Iterative regularization for low complexity regularizers

Iterative regularization for low complexity regularizers

Iterative regularization for low complexity regularizers

Cesare Molinari[1], Mathurin Massias[2], Lorenzo Rosasco[2,3,4], Silvia Villa[1]

[1]MaLGa, DIMA, Università di Genova
[2]MaLGa, DIBRIS, Università di Genova
[3] Center for Brains, Minds and Machines, MIT
[4] Istituto Italiano di Tecnologia

Abstract

Iterative regularization exploits the implicit bias of optimization algorithms to regularize ill-posed problems. Constructing algorithms with such built-in regularization mechanisms is a classic challenge in inverse problems but also in modern machine learning, where it provides both a new perspective on algorithms analysis, and significant speed-ups compared to explicit regularization. In this work, we propose and study the first iterative regularization procedure with explicit computational steps able to handle biases described by non smooth and non strongly convex functionals, prominent in low-complexity regularization. Our approach is based on a primal-dual algorithm of which we analyze convergence and stability properties, even in the case where the original problem is unfeasible. The general results are illustrated considering the special case of sparse recovery with the ℓ_1 penalty. Our theoretical results are complemented by experiments showing the computational benefits of our approach.

AMS Subject Classification: 65K10, 90C25, 90C46

L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR project 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. S. V. and L. R. acknowledge the support of the AFOSR project FA8655-22-1-7034 and of the H2020-MSCA-ITN Project Trade-OPT 2019. S. V. and C. M. are part of the INDAM research group "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro applicazioni".

1 Introduction

Many questions in applied sciences and engineering can be cast as an inverse problem corresponding to recovering a quantity of interest from finite and noisy data. In this context, ensuring a stable and reliable recovery requires selecting solutions according to some prior knowledge (bias) on the problem at hand (Engl et al., 1996). The latter takes the form of functionals, called regularizers, and Hilbert norms are classical examples. Allowing for a wide range of regularizers is crucial to encode the possible biases needed in different problems. Indeed, an effort has been made through the years to consider a large classes of regularizers called low complexity regularizers, see e.g. Iutzeler and Malick (2020), Mosci et al. (2010), Rosasco et al. (2010), Vaiter et al. (2015) and references therein. These regularizers promote solutions with a simple structure accordingly to some criteria. Relevant examples are total variation, sparsity or low rank inducing functionals. Given a bias and corresponding regularizer, the idea is that solutions should be selected to be faithful to the data while encoding the chosen bias. In practice, the best trade-off between data fit and bias is hardly ever known and needs be investigated developing suitable computational and algorithmic procedures.

A classical way to turn the above ideas into algorithms is Tikhonov regularization (Tikhonov and Arsenin, 1977). In this approach, an objective function is designed as the sum of a data fit term and the regularizer. The relative weight of the two terms is controlled by a parameter, called regularization parameter. Different criterions are known and can be used to determine the best regularization parameter value, hence the best data-fit/bias trade-off, see e.g. Deledalle et al. (2014), Goldenshluger and Pereverzev (2003) and references therein. However, independently from the specific criterion considered, multiple optimization problems must be solved, corresponding to different regularization parameter values, possibly resulting in numerical inefficiencies. An alternative algorithmic approach is given by iterative regularization. Here, the basic observation is that an optimization process *itself* can be used to explore different trade-offs between data-fit and bias, see Chapter 6 in Engl et al. (1996). Indeed, it is possible to design optimization procedures that iteratively control the amount of bias encoded in the solution at each step. Then, the number of iterations becomes the regularization parameter. Iterative regularization is a classic idea in inverse problems that has recently become popular in machine learning. In this context iterative regularization is also called implicit regularization, as opposed to Tikhonov regularization, which is called explicit regularization. A spur of interest in iterative (implicit) regularization followed the observation that it seems to play a key role in understanding the properties of deep neural networks, see e.g. Gunasekar et al. (2017). Another reason for recent interest is that it provides a substantially more efficient alternative to Tikhonov regularization, since only a single optimization process is needed, rather than multiple ones (Raskutti et al., 2014, Vaškevičius et al., 2019, Wei et al., 2017, Yao et al., 2007).

The above discussion motivates the question of developing iterative regularization approaches for wide classes of models (linear and nonlinear), data-fit terms and bias/regularizers. Indeed, for Tikhonov regularization, these combinations are quite straightforward, simply by designing suitable objective functions, and the issues largely reside in the solution of the corresponding optimization problems. In iterative regularization, the design and optimization steps are entangled, and so the above question is immediately challenging. In this work, we focus on dealing with general classes of bias/regularizers for linear models. The most classic example of iterative regularization is gradient descent, also called Landweber iteration in inverse problems, which corresponds to a bias given by a Hilbert norm (Engl et al., 1996). Extensions to regularizers that are Banach norms are also known (Brianzi et al., 2013, Kaltenbacher et al., 2008, Schöpfer et al., 2006). The case of regularizers given by strongly convex functionals has been investigated in two lines of work. The first one is based on mirror descent (Gunasekar et al., 2018, Vaškevičius et al., 2020), and can also be viewed as dual gradient descent Villa et al. (2022). The second one, arising from the imaging community, is called linearized Bregman iterations (Benning and Burger, 2018b, Cai et al., 2009a,b, Lorenz et al., 2014a,b, Yin, 2010). For a general convex functional, linearized Bregman iterations solve a relaxation of the problem where the bias is additionally penalized by a norm squared term.

We are interested in regularizers defined by functionals that are convex but neither smooth nor strongly convex. This setting allows to deal with total variation, sparsity/low rank inducing regularizers, and more generally low complexity regularizers. In this case, the explicit (variational) regularization procedures are well-understood, see Burger and Osher (2004). Iterative regularization schemes are known only for the ℓ_1-norm (Vaškevičius et al., 2019), require tuning additional parameters (Yin, 2010, Yin et al., 2008), or solving a non-trivial optimization problem at every step (see the literature about Bregman iteration, also called inverse scale space method in Benning and Burger (2018b), Burger et al. (2005, 2006, 2007, 2013)). Devising a generic and practical iterative regularization procedure for convex regularizers with associated theoretical guarantees is thus an open problem. In this work, we propose and analyze the first efficient iterative regularization procedure applicable to non smooth, non strongly convex regularizers. Our approach builds on ideas in optimization specifically primal dual approaches (Chambolle and Pock, 2011, Condat, 2013, Vũ, 2013). Further, beyond convergence we deal with stability with respect to noise, borrowing ideas from inexact optimization, see e.g. Salzo and Villa (2012) and references therein. Indeed, in the presence of noise, we derive a stopping time varying as the inverse of the noise level, in accordance with known results for strongly convex regularizers and matching known error bounds for Tikhonov regularization. Our study covers the case of ill-posed problems, where solutions might not exist, mentioned as an open problem in Benning and Burger (2018a) even for Tikhonov regularization. Further, we consider two relevant optimization ideas. The

first one is the use of preconditioning to speed up computations. The second one is to allow for approximate computations which further enlarge the class of bias that can be efficiently considered. Beyond the general analysis, we provide a specialized discussion when the regularizer is the ℓ_1 norm, and use ideas from Grasmair et al. (2011) to obtain recovery results. Finally, we validate our approach numerically and provide an open source Python package at `https://lcsl.github.io/iterreg/`. Our experiments are meant to be a proof of concept, and are by no means exhaustive. The practical performance of course depends on the application and on the stopping time selection criterion. To properly assess its effectiveness on real problems, an extensive analysis would be needed relying on classical parameter's selection rules and is deferred to future work. The structure of the paper is as follows: we first formalize in Section 2 the problem at hand and detail the notions of explicit and iterative regularization. In Section 3 we present the algorithm we use for iterative regularization and the setup under which we analyze it. In Section 4, we state our main result: stability in the presence of noise and a stopping time for iterative regularization. Section 5 contains a detailed comparison of our results to existing approaches. Section 6 is devoted to deeper results in the case of sparse recovery with the ℓ_1 norm. In Section 7 we study some cases of iterative regularization where the solution to the problem does not exist. Experiments in Section 8 demonstrate the validity of the approach.

Notation Let $\mathcal{X}$ be a real Hilbert space endowed with the scalar product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $f \colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. For $\varepsilon \geq 0$, the ε-subdifferential of f at the point $x \in \mathcal{X}$ is the set $\partial_\varepsilon f(x) = \{u \in \mathcal{X} : \forall y \in \mathcal{X}, f(x) - f(y) \leq \langle u, x - y \rangle + \varepsilon\}$; for $\varepsilon = 0$ we write $\partial f(x)$. For a symmetric positive definite $T \colon \mathcal{X} \to \mathcal{X}$, $\|x\|_T^2 := \langle T^{-1}x, x \rangle$. The T-preconditioned proximal operator of f at x is $\operatorname{prox}_f^T(x) = \operatorname{argmin}_{x' \in \mathcal{X}} f(x') + \frac{1}{2} \|x - x'\|_T^2$. The set of proper, convex and closed functions on the space $\mathcal{X}$ is denoted by $\Gamma_0(\mathcal{X})$. For a convex function R, $x' \in \mathcal{X}$ and $\theta \in \partial R(x')$, the Bregman divergence[1] induced by R with subgradient θ is defined as $D_R^\theta(x, x') = R(x) - R(x') - \langle \theta, x - x' \rangle$. When R is differentiable, its subdifferential at x' reduces to $\{\nabla R(x')\}$ and thus, for the Bregman divergence, we omit the θ superscript. The pointwise multiplication between vectors, or row-wise multiplication between a vector and a matrix, is denoted $\odot$. The vector of $\mathbb{R}^n$ with all entries equal to 1 is 1_n. For two vectors a and b, $a \succeq b$ is to be understood entrywise, meaning that $a_i \geq b_i$ for all i.

[1] This is a abuse of language since, when R is not strictly convex, D_R may not be a divergence; meaning that, in general, $D_R(x, x') = 0$ does not imply $x' = x$.

2 Background on explicit and iterative regularization

As one motivation for our setting, consider the problem of learning a mapping f from observations $(a_i, b_i)_{i\in[n]} \in \mathbb{R}^p \times \mathbb{R}$ such that $f(a_i) = b_i$. In the case where f is a linear function, this amounts to learn a vector x such that $\langle a_i, x\rangle = b_i$ for all $i \in [n]$, which, introducing the design matrix $A = (a_1^\top, \dots, a_n^\top)^\top \in \mathbb{R}^{n\times p}$, is equivalent to ~~solve~~

$$Ax = b \ . \tag{1}$$

Such inverse problems are ubiquitous in machine learning (Steinwart and Christmann, 2008), signal processing (Foucart and Rauhut, 2013) and image processing (Chambolle and Pock, 2016). Recent successful analyses of deep learning also consider linear approximations of this kind (Ghorbani et al., 2021). The solution to Eq. (1) is often not unique, for example in the overparametrized setting when $p > n$, common in machine learning, statistics and signal processing. In this situation, amongst all possible solutions, it is popular to favor a particular one, e.g. considering:

$$\min_{x\in\mathcal{X}} R(x) \quad \text{s.t.} \quad Ax = b \ , \tag{2}$$

where the regularizer R (also called penalty, or bias) selects the solutions of interest. In this work, we are interested in a special type of regularizers, low complexity ones, which force the solution x to lie on a reduced subset of the space, for instance to be sparse. Critically, these regularizers are neither smooth, nor strongly convex (see, for instance, Bach et al. (2012), Iutzeler and Malick (2020), Vaiter et al. (2015)). In Examples 1 to 3, we recall some well-known examples; other notable examples include the ℓ_∞ norm (Zhao et al., 2016), ordered ℓ_1 penalties (Figueiredo and Nowak, 2016) or block-sparse penalties (Obozinski et al., 2010, Simon et al., 2013).

Example 1 (Sparse regression and classification)**.** *When $\mathcal{X} = \mathbb{R}^p$, choosing $R(\cdot) = \|\cdot\|_1$ corresponds to finding the minimal ℓ_1-norm solution to a linear system, and in this case (2) is known as Basis Pursuit (Chen et al., 1998). Following the practical success of compressed sensing (Candès et al., 2006, Donoho, 2006), ℓ_1-based approaches have had a tremendous impact in imaging, signal processing and machine learning in the last decades (see Hastie et al. (2015) for a review). Problem (2) also encompasses classification with the following rewriting: if one searches for the minimal f-valued separator to a linearly separable dataset (a_i, b_i), the problem is:*

$$\min_{x\in\mathbb{R}^p} f(x) \quad s.t. \quad (b \odot A)x \succeq 1_n \ . \tag{3}$$

Introducing a slack variable $u = (b\odot A)x - 1_n$, (3) fits in the framework of Problem (2) using $\tilde{x} = \big(x, u\big)$, $R(\tilde{x}) = f(x) + \iota_{\{\cdot\succeq 0\}}(u)$, $\tilde{A} = \big(b \odot A, \quad -\mathrm{Id}\big)$ and $\tilde{b} = 1_n$.

Example 2 (Low rank matrix completion)**.** *In many practical applications, such as recommender systems, one seeks to recover a partially observed matrix B based on the assumption that its rank is low (Candès and Recht, 2009, Fazel, 2002). A convex approach to this problem is:*

$$\min_{X \in \mathbb{R}^{p_1 \times p_2}} \|X\|_* \quad s.t. \quad X_{ij} = B_{ij} \quad \forall (i,j) \in \mathcal{D} \ , \tag{4}$$

where $\|\cdot\|_*$ is the nuclear norm and $\mathcal{D} \subset [p_1] \times [p_2]$ is the set of observed entries of the matrix B. The problem in (4) is a special case of Problem (2), with A the linear operator from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^{p_1 \times p_2}$, such that $(AX)_{ij}$ has value X_{ij} if $(i,j) \in \mathcal{D}$ and 0 otherwise; and $b = AB$.

Example 3 (Total Variation)**.** *In imaging tasks such as deblurring and denoising, regularization via Total Variation allows to simultaneously preserve edges while removing noise in flat regions (Rudin et al., 1992). Given a blurring operator $A : \mathcal{X} \to \mathcal{X}$, the problem of Total Variation is:*

$$\min_{X \in \mathbb{R}^{p_1 \times p_2}} \|\nabla X\|_{2,1} \ \ s.t. \ AX = B \ , \tag{5}$$

where for $u = (u_1, u_2) \in \mathbb{R}^{2(p_1 \times p_2)}$, $\|u\|_{2,1} = \sum_{i,j} \sqrt{(u_1)_{ij}^2 + (u_2)_{ij}^2}$. The above problem can be re-written as:

$$\min_{\tilde{X} \in \mathbb{R}^{(p_1+p_2) \times p_2}} \Omega(\tilde{X}) \quad s.t. \quad \tilde{A}\tilde{X} = \tilde{B} \ , \tag{6}$$

with $\tilde{X} = \begin{pmatrix} X \\ U \end{pmatrix}$, $\Omega(\tilde{X}) = \|U\|_{2,1}$, $\tilde{A} = \begin{pmatrix} A & 0 \\ \nabla & -\mathrm{Id} \end{pmatrix}$ and $\tilde{B} = \begin{pmatrix} B \\ 0 \end{pmatrix}$. To avoid increasing the dimension of the problem, one can also consider directly problem (5) and compute the proximal operator of TV approximately, in which case it is necessary to handle errors in the prox (Villa et al., 2013), see Equation (13).

In practice, it is frequent that the observations in problem (2) are corrupted by noise: the true observations are only available through a noisy version b^δ. To avoid fitting the noise in the data, one should no longer impose the constraint $Ax = b^\delta$ and the approach (2) must be modified. Explicit regularization consists in relaxing the equality constraint into a penalization, and solving a composite optimization problem:

$$\min_{x \in \mathcal{X}} \frac{1}{2} \left\| Ax - b^\delta \right\|^2 + \lambda R(x) \ , \tag{7}$$

where the nonnegative scalar λ controls the trade-off between fitting the data and regularizing the solution. As mentioned in the introduction, selecting the correct value for λ is computationally costly.

Alternatively, it is possible to exploit the implicit bias of an optimization algorithm. As a classical example, it is well-known (Engl et al., 1996, Chap 6) that iterations of gradient descent on least-squares,

$$x_{k+1} = x_k - \gamma A^*(Ax_k - b) \ , \tag{8}$$

converge[2] to the solution of (2) with $R = \frac{1}{2}\|\cdot\|^2$. When applied to $b = b^\delta$, iterative regularization consists in stopping gradient descent iterates before convergence (Engl et al., 1996, Chap 6). What controls the regularization strength in this case is the number of iterations performed . Typically, when the noise level is of order of magnitude δ, one seeks an implicitly biased algorithm, a strictly positive scalar α and a stopping time $k(\delta)$ such that the algorithm, applied to b^δ, produces iterates (x_k) satisfying:

$$D(x_{k(\delta)}, x^\star) \leq \mathcal{O}(\delta^\alpha) \ , \tag{9}$$

where D is some discrepancy measure and $x^\star$ is a solution of (2) with exact data. The stopping time $k(\delta)$ usually depends on unknown constants. Indeed, our results are theoretical bounds with a-priori choice of the parameters. Thus, the best model across iterates must be selected according to some criterion, exactly as a model is chosen between the solutions of (7) in the explicit approach. Typical examples are, in inverse problems, the discrepancy principle; and, in machine learning, cross-validation.
As discussed next, it is the contribution of this paper to provide an algorithm, a stopping time and guarantees for a generic non-smooth convex regularizer R.

3 Algorithm and assumptions

In this section we present the algorithm we study and the mathematical assumptions we consider.

3.1 Algorithm

Let $\mathcal{X}$ and $\mathcal{Y}$ be real Hilbert spaces, $A\colon \mathcal{X} \to \mathcal{Y}$ a linear and bounded operator, and $b \in \mathcal{Y}$. Generalizing the above discussion, we consider the following minimization problem,

$$\min_{x \in \mathcal{X}} R(x) + F(x) \quad \text{s.t.} \quad Ax = b \ . \tag{10}$$

The functions $R\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $F\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ are both assumed to be convex, proper, and lower-semicontinuous. In addition, F is differentiable. Compared to (2), the

[2] If initialized at 0 and provided $\gamma < 2/\|A\|_{\text{op}}^2$.

splitting between a nonsmooth and a smooth term allows us to handle the smooth term F using only its gradient.

Let $b^\star \in \mathcal{Y}$ denote the exact observation, typically unavailable, and $b^\delta \in \mathcal{Y}$ denote the accessible noisy data. We study a worst-case situation; namely, for some $\delta \geq 0$, we assume that

$$\|b^\delta - b^\star\| \leq \delta \ . \tag{11}$$

The algorithm we consider for iterative regularization is a preconditioned and inexact version of a three steps primal-dual procedure (Chambolle and Pock, 2011, Condat, 2013, Vũ, 2013) applied to the noisy data b^δ. Given initializations $y_{-1},\ y_0 \in \mathcal{Y}$ and $x_0 \in \mathcal{X}$, consider

$$\begin{cases} \tilde{y}_k = 2y_k - y_{k-1} \ , \\ x_{k+1} = \operatorname{prox}_R^{T,\varepsilon_{k+1}}(x_k - T\nabla F(x_k) - TA^*\tilde{y}_k) \ , \\ y_{k+1} = y_k + \Sigma\left(Ax_{k+1} - b^\delta\right) \ . \end{cases} \tag{12}$$

The first step is an extrapolation on the dual variable; the second one is the update of the primal variable and involves the proximal-point operator of R and the gradient of F; finally, the third step is the update of the dual variable, which accumulates the residuals of the constraint $Ax = b^\delta$. The operators $T\colon \mathcal{X} \to \mathcal{X}$ and $\Sigma\colon \mathcal{Y} \to \mathcal{Y}$ are linear, positive and bounded and can be intepreted as preconditioners, or, if proportional to the identity, as step-sizes. The proximal-point operator of R is allowed to be computed inexactly with error $\varepsilon_{k+1} \geq 0$, recovering the exact case for $\varepsilon_{k+1} = 0$. The notation $\operatorname{prox}_R^{T,\varepsilon_{k+1}}$ is intended in terms of ε-subdifferential, namely

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_R^{T,\varepsilon_{k+1}}(x_k - T\nabla F(x_k) - TA^*\tilde{y}_k) \\ &\iff -T^{-1}(x_{k+1} - x_k) - \nabla F(x_k) - A^*\tilde{y}_k \in \partial_{\varepsilon_{k+1}} R(x_{k+1}) \ . \end{aligned} \tag{13}$$

To interpret algorithm (12) as an instance of the approach in Condat (2013), Vũ (2013), it is useful to cast the update of the dual variable y as a proximal step:

$$\begin{aligned} y_{k+1} &= \operatorname*{argmin}_{y\in\mathcal{Y}} \left\{ \langle b^\delta - Ax_{k+1}, y\rangle + \frac{1}{2}\|y - y_k\|_\Sigma^2 \right\} \\ &= \operatorname*{argmin}_{y\in\mathcal{Y}} \left\{ \langle b^\delta, y\rangle + \frac{1}{2}\|y - [y_k + \Sigma Ax_{k+1}]\|_\Sigma^2 \right\} \\ &= \operatorname{prox}^\Sigma_{\langle b^\delta,\cdot\rangle}(y_k + \Sigma Ax_{k+1}) \ . \end{aligned} \tag{14}$$

The above algorithm is cheap in terms of computations per iteration. Indeed, it only requires one (inexact) evaluation of the proximal operator of the non-smooth function R, one evaluation of the gradient of the smooth function F and one matrix-vector multiplication for A and

A^*. Its memory cost is also minimal, as only one primal and two dual variables need to be stored. The prox of R can be computed exactly for many penalties of interest (see Combettes and Pesquet (2011), Mosci et al. (2010)). Through ε_k, our framework also handles the case where the optimization problem defined by the proximal operator is numerically computed, in an approximate fashion, through an iterative inner-routine (see Bach et al. (2012), Barré et al. (2020), Salzo and Villa (2012)).

3.2 Assumptions

We first make the following general assumptions on functions and operators involved in the problem.

Assumption 4. *$\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces and $A : \mathcal{X} \to \mathcal{Y}$ is linear and bounded. The functions R and F belong to $\Gamma_0(\mathcal{X})$, meaning that they are proper, convex and lower-semicontinuous. Additionally, F is Fréchet-differentiable with L-Lipschitz continuous gradient on $\mathcal{X}$.*

In order to introduce the next assumptions on the problem and the existence of an exact solution, we first define the set of primal solutions, the set of dual solutions, and the Lagrangian functional with respect to the exact datum $b^\star$:

$$\mathcal{P}^\star := \operatorname*{argmin}_{x\in\mathcal{X}} \{R(x) + F(x) : \quad Ax = b^\star\} \quad , \tag{15}$$

$$\mathcal{D}^\star := \operatorname*{argmin}_{y\in\mathcal{Y}} \{[R + F]^* (-A^* y) + \langle b^\star, y\rangle\} \quad , \tag{16}$$

$$\mathcal{L}^\star(x, y) := R(x) + F(x) + \langle y, Ax - b^\star\rangle \quad . \tag{17}$$

We also denote by $\mathcal{S}^\star$ the set of saddle-points of $\mathcal{L}^\star$; namely, $(\bar{x}, \bar{y}) \in \mathcal{S}^\star$ if and only if $\mathcal{L}^\star(\bar{x}, y) - \mathcal{L}^\star(x, \bar{y}) \leq 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We write $\mathcal{P}^\delta$, $\mathcal{D}^\delta$, $\mathcal{L}^\delta$ and $\mathcal{S}^\delta$ for their respective counterparts when $b^\star$ is replaced by b^δ. We refer to the corresponding problems and quantities as the exact and noisy ones, respectively. In the rest of the paper we will assume that one only has access to the noisy quantities: hence our focus is on the iterative algorithm designed to solve the noisy problem $\mathcal{P}^\delta$, having in mind that the problem of interest is the exact one $\mathcal{P}^\star$. We make the following assumptions on the existence of solution *to the exact problem*. Notice that, on the other hand, we do not require the existence of solutions (or even feasibility) for the noisy one.

Assumption 5 (Existence of exact solution)**.** *There exists a saddle-point for the Lagrangian $\mathcal{L}^\star$ ($\mathcal{S}^\star \neq \emptyset$); namely, a pair $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ such that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$\mathcal{L}^\star(x^\star, y) - \mathcal{L}^\star(x, y^\star) \leq 0 \quad .$$

Remark 6. *Under Assumption 4, the following statements are equivalent:*

- $(x^\star, y^\star) \in \mathcal{S}^\star$*; namely, it is a saddle-point for the Lagrangian* $\mathcal{L}^\star$*;*
- $x^\star$ *and* $y^\star$ *satisfy the following optimality conditions:*

$$\begin{cases} -A^* y^\star - \nabla F(x^\star) \in \partial R(x^\star) \ , \\ Ax^\star = b^\star \ . \end{cases} \tag{18}$$

Moreover, either one of the previous conditions implies that $x^\star$ *is a primal solution and* $y^\star$ *is a dual solution; namely, it holds* $\mathcal{S}^\star \subseteq \mathcal{P}^\star \times \mathcal{D}^\star$*. Under usual qualification conditions (Bauschke and Combettes, 2011, Thm. 26.2) (also called source conditions, see Burger and Osher (2004)), the converse is also true: if* $x^\star \in \mathcal{P}^\star$ *and* $y^\star \in \mathcal{D}^\star$*, then* $(x^\star, y^\star)$ *is a saddle-point; namely,* $\mathcal{S}^\star = \mathcal{P}^\star \times \mathcal{D}^\star$*.*

As far as the parameters of the algorithm are concerned, we make the following assumptions on the preconditioners T and Σ.

Assumption 7. *The operator* $T : \mathcal{X} \to \mathcal{X}$ *is linear, bounded, self-adjoint and positive with spectrum lower and upper bounded by* $\tau_m > 0$ *and* τ_M *respectively. The same holds for* $\Sigma : \mathcal{Y} \to \mathcal{Y}$ *with lower and upper bounds* $\sigma_m > 0$ *and* σ_M*.*

Assumption 8. *Define the quantity* $\omega := 1 - \tau_M(L + \sigma_M \|A\|^2)$*. The operators* T *and* Σ *are chosen so that* $\omega \geq 0$*.*

Assumption 9. *For* $0 < \xi < 1$ *and* $\eta > 1$*, let* $\theta := \xi - \tau_M(\xi L + \sigma_M \|A\|^2)$ *and* $\rho := \sigma_m(\eta - 1) - \sigma_M \xi \eta$*. The operators* T *and* Σ *and the constants* ξ, η *are chosen so that* $\theta \geq 0$ *and* $\rho > 0$*.*

Note that Assumption 9 is stronger than Assumption 8, that is the classical one for primal-dual algorithm (see Chambolle and Pock (2011), Condat (2013), Vũ (2013)). We consider them separately because some of our results hold only for Assumption 9, while for others Assumption 8 is sufficient. For every value of $L \geq 0$ and $\|A\|$, it is always possible to choose T and Σ and constants ξ and η so that Assumption 9 (and so Assumption 8) is fulfilled. For instance, choosing $\xi = 1/4$ and $\eta = 3/2$ amounts to require $\sigma_M < (4/3)\sigma_m$ and $\tau_M \leq (L + 4\sigma_M \|A\|^2)^{-1}$. For simplicity, the two preconditioners can be taken diagonal or as $T = \tau \,\mathrm{Id}$ and $\Sigma = \sigma \,\mathrm{Id}$, where Id is the identity operator while τ and σ are positive parameters representing the primal and dual stepsizes of the algorithm. In this case $\tau_m = \tau_M = \tau$, $\sigma_m = \sigma_M = \sigma$ and Assumption 9 naturally simplifies to $\tau \leq \xi(\xi L + \sigma \|A\|^2)^{-1}$ for some $0 < \xi < 1$. For example, if $F = 0$ and thus $L = 0$, one recovers the classical step-size condition for the algorithm of Chambolle and Pock (2011), that is $\sigma\tau \|A\|^2 < 1$.

In the framework introduced above, we next show that Algorithm (12) is well-suited to iterative regularization, by studying its convergence and stability properties.

4 Convergence, stability and early-stopping bounds

In this section, we present the main results of the paper. We start with a generalization of a well-known result about convergence of primal-dual algorithms. We include it since it highlights the implicit bias of our algorithm in the case of exact data and exact computations ($b^\delta = b^\star$ and $\varepsilon_k = 0$). Indeed, we prove convergence to a solution of problem $\mathcal{P}^\star$, namely, amongst all solutions to $Ax = b^\star$, Algorithm (12) converges to one with minimal regularizer value.

Proposition 10. *Assume that Assumptions 4 and 5 hold. Let (x_k, y_k) be the sequence generated by iterations (12) applied to $b^\delta = b^\star$ under Assumptions 7 and 8. Let also $\varepsilon_k = 0$ for every $k \in \mathbb{N}$. Then (x_k, y_k) weakly converges to a pair in $\mathcal{S}^\star$. In particular, (x_k) weakly converges to a point in $\mathcal{P}^\star$.*

Proposition 10 is a first step towards an iterative regularization procedure: it shows that in the absence of noise, iterations (12) converge to a solution of interest. The proof is in Appendix B.1. It is a generalization of Pock and Chambolle (2011), which treats the preconditioning with $F = 0$; and of Condat (2013), Vũ (2013), considering $F \neq 0$, but without the preconditioning.

The next step is to show that when only b^δ is available, one can approximate the exact solution by early stopping the iterations (12) with noisy data. To this end, we prove stability results in terms of Lagrangian gap and feasibility, that allow to derive a stopping time depending on the noise level δ. Before stating our main result (Theorem 13), we first highlight why the Lagrangian gap and the feasibility are adequate quantities to measure convergence of the primal variable. In the next lemma we show that, if they are both zero, the primal variable is a solution of $\mathcal{P}^\star$.

Proposition 11. *Let $(x^\star, y^\star) \in \mathcal{S}^\star$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y) = 0$ and $Ax = b^\star$. Then $(x, y^\star) \in \mathcal{S}^\star$.*

We call the quantity $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y)$ *Lagrangian gap*, as it is always non negative since $(x^\star, y^\star)$ is a saddle point. More specifically, it is equal to the Bregman divergence $D_{R+F}^{-A^*y^\star}(x, x^\star)$, as we detail in the proof (Appendix B.2). The latter has been often used as an optimality measure in this context, see e.g. Burger et al. (2007). However, we emphasize that, contrarily to the α-strongly convex case (where $D_{R+F}^{-A^*y^*}(x, x^\star) \geq \frac{\alpha}{2}\|x - x^\star\|^2$), a vanishing Lagrangian gap is not enough for the primal variable to be a solution of the primal problem. For example, for $R(\cdot) = \|\cdot\|_1$ and $F(\cdot) = 0$, the quantity $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y)$ vanishes whenever x and $x^\star$ have the same support and sign (or simply when $x = 0$), while the primal variable x can still be arbitrarily far away from $x^\star$ (see Figure 1).

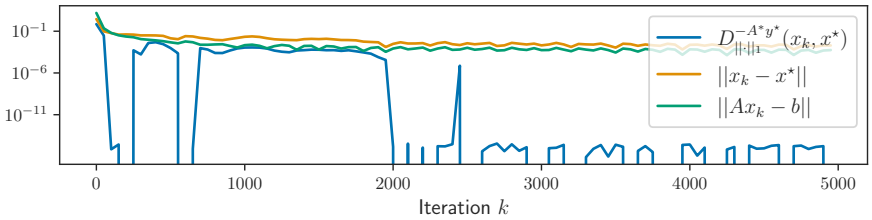

Figure 1: When $R + F$ is not strongly convex, the Bregman divergence alone is not enough to provide useful convergence rates as it may not be point separating. Here for $R = \|\cdot\|_1$, $F = 0$, $D_R^{-A^\star y^\star}(x_k, x^\star)$ vanishes quickly, while the iterates x_k of (12) are still far from their limit $x^\star$ (synthetic noiseless data, exact prox).

Remark 12 (Comparison with duality gap)**.** *Another quantity that is often considered as optimality measure for primal-dual algorithms is*

$$\sup_{v \in B_2} \mathcal{L}^\star(x, v) - \inf_{u \in B_1} \mathcal{L}^\star(u, y),$$

where $B_1 \subseteq \mathcal{X}$ and $B_2 \subseteq \mathcal{Y}$ are two bounded sets containing a primal-dual solution, see for example Chambolle and Pock (2011). We note that such a bound can be easily derived from our convergence bounds in the exact setting. As discussed in Chambolle and Pock (2011), the choice of B_1 and B_2 is tricky while our bound is more easily readable in our linearly constrained setting.

In the next result, we prove a stability bound for the iterates applied to the noisy problem, in terms of the optimality metric discussed above.

Theorem 13. *Let Assumptions 4 and 5 hold and $(x^\star, y^\star) \in \mathcal{S}^\star$ be a saddle-point of the exact problem. Let (x_k, y_k) be generated by (12) under Assumptions 7 and 8 with inexact data b^δ such that $\left\|b^\delta - b^\star\right\| \leq \delta$ and error $|\varepsilon_k| \leq C_0\delta$ in the proximal operator for all $k \in \mathbb{N}$. Denote by $(\hat{x}_k, \hat{y}_k)$ the averaged iterates $(\frac{1}{k}\sum_{j=1}^k x_j, \frac{1}{k}\sum_{j=1}^k y_j)$. Then there exist constants C_1, C_2, C_3 and C_4 such that, for every $k \in \mathbb{N}$,*

$$\mathcal{L}^\star(\hat{x}_k, y^\star) - \mathcal{L}^\star(x^\star, \hat{y}_k) \;\leq\; \frac{C_1}{k} + C_2\delta + C_3\delta^{3/2}k^{1/2} + C_4\delta^2 k \;. \tag{19}$$

Let also Assumption 9 hold. Then there exist constants C_5, C_6, C_7, C_8 and C_9 such that, for every $k \in \mathbb{N}$,

$$\|A\hat{x}_k - b^\star\|^2 \;\leq \frac{C_5}{k} + C_6\delta + C_7\delta^{3/2}k^{1/2} + C_8\delta^2 k + C_9\delta^2 \;. \tag{20}$$

The proof is given in Appendix B.3, where the reader can find also the explicit expression for all the constants involved in the bounds. Note that the bounds (19) and (20) are composed of two kinds of terms. The first kind, related to *optimization*, is of the form $\mathcal{O}(1/k)$ and vanishes with the iteration counter, as it is related to the convergence of the algorithm to the exact solution. The second kind, involving δ, is related to *stability* and is due to the unavailability of $b^\star$. In particular, when $\delta > 0$, the terms in k make the bound increase with the iteration counter.

The main consequence of Theorem 13 is an early stopping procedure that allows to obtain upper-bounds on both Lagrangian gap and feasibility.

Corollary 14. *Under the assumptions of Theorem 13, setting $k = \tilde{C}/\delta$ for some constant $\tilde{C} > 0$, there exist constants C and C' such that*

$$\begin{aligned} \mathcal{L}^\star(\hat{x}_k, y^\star) - \mathcal{L}^\star(x^\star, \hat{y}_k) \;&\leq\; C\delta \;, \\ \|A\hat{x}_k - b^\star\|^2 \;&\leq C'\delta + C_6\delta^2 \;. \end{aligned}$$

This result, combined with Proposition 11, shows that the exact solution can be approximated by the averaged iterates generated by algorithm (12) on the noisy data, even if the true data is unavailable, by stopping at an appropriate iteration. Assuming $\delta \leq 1$, the level of approximation between the early-stopped iterate and the exact solution is then proportional to the noise level δ, both for the Lagrangian gap and the feasibility. We provide further comments and comparisons with existing results in the next section. We add one remark first.

Remark 15 (Early stopping in absence of noisy solution)**.** *We have shown that Algorithm (12), with appropriate early-stopping strategies, provides a good approximation of the exact solution, even if the noiseless datum is unavailable. Ill-posedness of the problem may be due to instability or non existence of the noisy solution. Our bounds in Theorem 13 apply to both these situations. If the problem is ill-posed from the stability point of view, the noiseless and noisy solutions are far apart, and the bounds in Theorem 13 imply that early-stopping ensures a computationally efficient way to find a stable solution. If the noisy problem does not have a solution, the averaged primal iterates generated by Algorithm (12) may diverge (see the example in Appendix B.4). In this case, early-stopping is necessary to prevent unboundedness. This confirms that it is unavoidable to have a stability bound going to $+\infty$ with the number of iterations. For more results related to the unfeasible case, see also Section 7.*

5 Comparison with existing results

The idea of exploiting the implicit regularization properties of optimization algorithms has been studied, often under the name of iterative regularization, in the fields of inverse problems

(Engl et al., 1996), image restoration (Burger et al., 2007), and machine learning (Yao et al., 2007). Existing methods can be divided into two classes, depending on whether or not strong convexity of the regularizer is assumed. In the following we compare known results with ours.

5.1 Strongly convex regularizer

We begin noting that, to the best of our knowledge, our method is the only one to handle the smooth term F in the regularizer using only its gradient. We next provide an overview of the algorithms proposed for iterative regularization.

- *Gradient descent, stochastic or accelerated.* The study of implicit regularization properties of gradient descent, known in the inverse problem community as Landweber method, goes back to the 50's (Engl et al., 1996, Chap. 6). Accelerated versions of gradient descent, first proposed by Nesterov in Nesterov (1983), have been also studied in inverse problems (Neubauer, 2017). Approaches related to the heavy-ball method (Polyak, 1964) have also been considered in inverse problems under the name of ν-method, see Engl et al. (1996). Generalizations towards p norms with $p > 1$ have been considered (Brianzi et al., 2013, Schöpfer et al., 2006), while more general choices are not as studied. Interestingly, there is a rich literature in the non-convex setting for nonlinear inverse problems (Kaltenbacher et al., 2008). These ideas have been extended to machine learning considering regularizing properties of gradient descent (Yao et al., 2007), and its stochastic and accelerated versions (Moulines and Bach, 2011, Pagliana and Rosasco, 2019, Rosasco and Villa, 2015).

- *Linearized Bregman iterations and mirror descent.* Interest in regularizers beyond the Euclidean norm, in particular non strongly convex ones, has been mainly motivated by imaging applications and Total Variation regularization. Following the pioneering work of Lorenz et al. (2014b), Osher et al. (2005), a series of methods have been designed for iterative regularization with general convex regularizers (see Burger et al. (2007) and references therein). If R is α-strongly convex, the iterative algorithm to exploit is mirror descent (Nemirovski and Yudin, 1983, Teboulle and Beck, 2003), which has been popularized in the inverse/imaging problems community under the name of "Linearized Bregman iterations" (Yin, 2010, Yin et al., 2008):

$$\begin{cases} x_{k+1} = \operatorname{argmin}_{x \in \mathcal{X}} D^{p_k}_{R - \frac{\alpha}{2}\|\cdot\|^2}(x, x_k) + \langle x, A^*(Ax_k - b)\rangle + \frac{1}{2\alpha}\|x - x_k\|^2 \\ p_{k+1} = p_k - \frac{1}{\alpha}(x_{k+1} - x_k) - A^*(Ax_k - b) \end{cases}, \tag{21}$$

  It has been shown that this algorithm, in combination with a discrepancy type stopping rule, regularizes ill-posed problems (Cai et al., 2009a).

- *Accelerated dual gradient descent.* From a different perspective, the stability and regularization properties of the accelerated variant of Linearized Bregman iterations have been studied in Villa et al. (2022). In the latter, mirror descent is interpreted as gradient descent applied to the dual; this connection, without acceleration, can also be found in Yin (2010), and an accelerated version was previously studied under the optimization lens in Huang et al. (2013).

- *Diagonal approaches.* All the aforementioned techniques are tailored to the use of a quadratic datafitting term. They cannot be applied when the nature of the noise differs, calling for another loss. In that case, diagonal approaches offer an alternative, applying an optimization algorithm to successive approximations of the original problem (Bahraoui and Lemaire, 1994, Bredies and Chariy, 2013). Convergence rates and stability of diagonal approaches for inverse problems have been considered in Garrigos et al. (2018) and in Calatroni et al. (2021) for the accelerated case.

5.2 Non strongly convex regularizers

If the regularizer is only convex, as we consider, Linearized Bregman iterations cannot be applied and one must resort to one of the following.

- *Bregman iteration and ADMM.* The main algorithm in this case is ADMM (Boyd et al., 2011), which has been studied in the imaging community under the name of Bregman iterations. Starting from $x_0 = 0$ and $p_0 = 0$, its updates read

$$\begin{cases} x_{k+1} \in \operatorname{argmin}_{x \in \mathcal{X}} D_R^{p_k}(x, x_k) + \frac{1}{2} \|Ax - b\|^2 \\ p_{k+1} = p_k - A^*(Ax_{k+1} - b) \end{cases}, \tag{22}$$

  The algorithm converges to the solution of (2); its regularization properties can be found in Burger et al. (2007). It has been extended to nonlinear inverse problems in Bachmayr and Burger (2009). However, this method is impractical, since it has a high computational cost: at each iteration a Tikhonov regularized problem has to be solved.

- *Bregmanized Operator Splitting and linearized/preconditioned ADMM.*
  These variants of Bregman iterations and ADMM rely on preconditioning to avoid the resolution of a difficult optimization problem at each iteration. They have been used empirically as regularizing procedures in inverse and imaging problems (Zhang et al., 2010, 2011). While convergence results are known, we are not aware of any theoretical quantitative stability result.

- *Specific algorithms for sparse recovery and compressed sensing.* In the specific context of sparse recovery, Osher et al. (2016) and Vaškevičius et al. (2019) have devised specific iterative regularization procedures. These do not generalize to regularizers beyond ℓ_1, nor do they allow to handle F.

- *Exact regularization ($F = 0$).* Exact regularization (Friedlander and Tseng, 2008, Schopfer, 2012, Yin, 2010) refers to solving

$$\min_x \; R(x) + \frac{\alpha}{2}\|x\|^2 \quad \text{s.t.} \quad Ax = b \;, \tag{23}$$

  and to showing that there exists a value of α such that this new problem and (10) have the same minimizer as (10). Then, known iterative regularization algorithms for the strongly convex case can be applied. The main drawback is that the existence of such a value of α is not guaranteed in general, it is problem specific, and cannot be determined in advance; hence it becomes a value to be tuned, which in turn is costly. When the regularizer is given by the ℓ^1-norm, this approach is also related to the one of sparse Kaczmarz method proposed in Schöpfer and Lorenz (2019).

As clear from the above discussion, to the best of our knowledge, *there previously did not exist an implementable iterative regularization procedure able to handle any non strongly convex regularizer*. Our proposed method fills this gap, and can be applied to the many instances of non smooth non strongly convex regularizers.

6 The special case of sparse recovery with ℓ_1-norm

In this section, we strengthen the results of Section 4 in the case of sparse recovery. The choice $R(\cdot) = \|\cdot\|_1$ has had a tremendous impact on sparse model estimation (Foucart and Rauhut, 2013). Below, we specialize our results to this case, obtaining bounds not only in terms of Lagrangian gap and feasibility, but directly on the distance between the iterates and the true model. The main result of the section, Theorem 19 is a corollary of our results and a lemma in Grasmair et al. (2011) which allows to control the distance between a point and a solution in terms of feasibility and Lagrangian gap. Therefore, in the next subsection, we first recall some results in Grasmair et al. (2011), while the new result is in Section 6.2.

6.1 Sparse recovery and compressed sensing

We set $\mathcal{X} = \ell^2(\mathbb{N};\mathbb{R})$, $R(\cdot) = \|\cdot\|_1$ and $F(\cdot) = 0$. The support of $x \in \mathcal{X}$ is $\mathrm{supp}(x) := \{i \in \mathbb{N} : \; x_i \neq 0\}$ and $|\cdot|$ denotes the cardinality of a set. The Bregman divergence induced by $\|\cdot\|_1$ is simply denoted by D. We first recall some notions from Grasmair et al. (2011).

Proposition 16. *Fix a primal-dual solution $(x^\star, y^\star) \in \mathcal{S}^\star$. Let the extended support be $\Gamma := \{i \in \mathbb{N} : |\,(A^*y^\star)_i\,| = 1\}$ and the saturation gap be $m := \sup\{|\,(A^*y^\star)_i\,| : |\,(A^*y^\star)_i\,| < 1\}$. Then Γ is finite, and $m < 1$. Moreover, for every $x \in \mathcal{X}$, with $\Gamma_C := \mathbb{N} \setminus \Gamma$,*

$$D^{-A^*y^\star}(x, x^\star) \geq (1 - m) \sum_{i \in \Gamma_C} |x_i|. \tag{24}$$

For completeness, the proof is reported in Appendix C.1. As $D^{-A^*y^\star}(x^\star, x^\star) = 0$ and $m < 1$, the (finite) set Γ can be considered as an *extended support*, as Equation (24) shows that $x^\star$ is zero on the indices of Γ_C.

More generally, if for some $x \in \mathcal{X}$ we have $D^{-A^*y^\star}(x, x^\star) = 0$, then $x = 0$ (and so x coincides with $x^\star$) on Γ_C. On the other hand, as mentioned above, $D^{-A^*y^\star}(x, x^\star) = 0$ does not ensure any similarity between the two vectors on Γ, the finite subset of indices where the components of $x^\star$ may be non-zero (see Figure 1).

To obtain sparse recovery results, we rely on compressed sensing assumptions on the design operator A and on the exact primal solution $x^\star$. Based on Assumption 17, Lemma 18 will allow us to bound $\|x - x^\star\|$ by a combination of the feasibility and the Lagrangian gap.

Assumption 17 (Compressed sensing)**.** *For some $s \in \mathbb{N}$,*

1. *there exists a s-sparse solution $x^\star$ to Eq. (1); namely, $Ax^\star = b^\star$ with $|\operatorname{supp}(x^\star)| \leq s$;*
2. *there exist constants $\theta_s, \theta_{s,s}$ and $\theta_{s,2s}$ such that*
   a) *for every $x \in \mathcal{X}$ with $|\operatorname{supp}(x)| \leq s$,*
   $$(1 - \theta_s)\,\|x\|^2 \leq \|Ax\|^2 \leq (1 + \theta_s)\,\|x\|^2\,;$$
   b) *for every $x, x' \in \mathcal{X}$ with $|\operatorname{supp}(x)| \leq s$, $|\operatorname{supp}(x')| \leq s$ (resp. $|\operatorname{supp}(x')| \leq 2s$) and $\operatorname{supp}(x) \cap \operatorname{supp}(x') = \emptyset$,*
   $$|\langle Ax, Ax' \rangle| \leq \theta_{s,s}\,\|x\|\,\|x'\|\,.$$
   (resp. $|\langle Ax, Ax' \rangle| \leq \theta_{s,2s}\,\|x\|\,\|x'\|$).
   c) *$\theta_s + \theta_{s,s} + \theta_{s,2s} < 1$.*

Lemma 18 (Grasmair et al. (2011), Prop. 5.3)**.** *Suppose Assumption 17 holds. Then:*

- *The vector $x^\star$ is the unique primal solution of Problem (2) with $R(\cdot) = \|\cdot\|_1$; namely,*
$$\operatorname*{argmin}_{x \in \mathcal{X}} \{\|x\|_1 : \quad Ax = b^\star\} = \{x^\star\} \quad . \tag{25}$$

- *There exists a dual solution* $y^\star \in \mathcal{Y}$ *such that*

$$\|y^\star\| \leq W_s := \frac{\sqrt{s}}{\sqrt{1-\theta_s}} \frac{\theta_{s,s}}{1-\theta_s-\theta_{s,2s}} \quad \text{and} \quad m \leq M_s := \frac{\theta_{s,s}}{1-\theta_s-\theta_{s,2s}} < 1,$$

  where m *is the saturation gap (Proposition 16) related to* $y^\star$.

- *Let* $\mathcal{X}_\mathcal{S} := \operatorname{span}\{e_i : \quad i \in \operatorname{supp}(x^\star)\}$ *and by* $i_\mathcal{S} : \quad \mathcal{X}_\mathcal{S} \to \mathcal{X}$ *the identity embedding. Then* $A_\mathcal{S} := A \circ i_\mathcal{S}$ *is injective with*

$$\left\|A_\mathcal{S}^{-1}\right\| \leq Q_s := \frac{1}{\sqrt{1-\theta_s}}. \tag{26}$$

- *For every* $x \in \mathcal{X}$,

$$\|x - x^\star\| \leq Q_s \|Ax - b^\star\| + \frac{1 + Q_s \|A\|}{1 - M_s} D^{-A^\star y^\star}(x, x^\star). \tag{27}$$

The previous results applied to Tikhonov regularization with ℓ^1 norm (Lasso), allow to derive explicit regularization results, which we recall in Appendix C.2. In a similar fashion, we use these facts for our proposed iterative regularization method instead. Note that there exist many other algorithms and penalties for sparse recovery. In particular, non convex penalties such as MCP Zhang (2010), SCAD (Fan and Li, 2001) or CNC (Lanza et al., 2019) have proven to perform better than the ℓ_1 norm. In this paper, we focus on iterative regularization for sparse recovery in the convex case, which is simpler to analyze. Extending the iterative regularization framework to non convex regularizers to leverage the better properties of non convex penalties is a promising research direction.

6.2 Sparse recovery with iterative regularization

Combining Theorem 13 with Assumption 17 and the inequality in (27), we get the following theorem for sparse recovery with ℓ_1-norm. In this setting, we are able to get an upper bound for the distance between the iterates and the exact solution.

Theorem 19. *Suppose that Assumption 17 holds. Let* $x^\star \in \mathcal{X}$ *be the unique primal solution of the exact problem*

$$\min_{x \in \mathcal{X}} \{\|x\|_1 : \quad Ax = b^\star\} \quad ,$$

18

and $y^\star \in \mathcal{Y}$ the dual solution given by Lemma 18. Moreover, under Assumptions 7 to 9, let $(\hat{x}_k, \hat{y}_k)$ be the sequence of averaged iterates generated by the primal-dual algorithm 12 when applied to the inexact problem

$$\min_{x \in \mathcal{X}} \left\{ \|x\|_1 : \quad Ax = b^\delta \right\} \quad .$$

Then we have that, for every $k \in \mathbb{N}$,

$$\begin{aligned} \|\hat{x}_k - x^\star\| &\leq Q_s \|A\hat{x}_k - b^\star\| + \frac{1 + Q_s \|A\|}{1 - M_s} D^{-A^* y^\star}(\hat{x}_k, x^\star) \\ &\leq Q_s \sqrt{\frac{C_4}{k} + C_5\delta + C_6\delta^2 + C_7\delta^2 k} + \frac{1 + Q_s \|A\|}{1 - M_s} \left[\frac{C_1}{k} + C_2\delta + C_3\delta^2 k \right] \end{aligned} \tag{28}$$

Remark 20 (Dependence on initialization)**.** *Notice that the bound (28) depends on the initialization z_0, through $V(z^\star - z_0)$ in the C_i's, see Appendix B.3. Yet, using the initialization $z_0 = 0$, we can bound the term $V(z^\star - z_0)$ by quantities that do not involve the unknown solution $z^\star$:*

$$\begin{aligned} V(z_0 - z^\star) &= \frac{1}{2\tau} \|x^\star\|^2 + \frac{1}{2\sigma} \|y^\star\|^2 \\ (26) \quad &\leq \frac{\|A_{\mathcal{S}} x^\star\|^2}{2\tau Q_s^2} + \frac{W_s^2}{2\sigma} \\ &= \frac{\|b^\star\|^2}{2\tau Q_s^2} + \frac{W_s^2}{2\sigma} \\ &\leq \frac{2\left\|b^\star - b^\delta\right\|^2 + 2\left\|b^\delta\right\|^2}{2\tau Q_s^2} + \frac{W_s^2}{2\sigma} \\ &\leq \frac{\delta^2 + \left\|b^\delta\right\|^2}{\tau Q_s^2} + \frac{W_s^2}{2\sigma}. \end{aligned} \tag{29}$$

For comparison with Tikhonov explicit regularization, we recall a result from Grasmair et al. (2011) (see Corollary 34 in the Appendix for the precise statement). Under Assumption 17 and for $\alpha > 0$, let

$$x_\alpha \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \left\|Ax - b^\delta\right\|^2 + \alpha \|x\|_1 \right\}. \tag{30}$$

Then, defining $D := (1 + Q_s \|A\|) / (1 - M_s)$,

$$\|x_\alpha - x^\star\| \leq \left(Q_s W_s + D W_s^2/4\right) \alpha + (Q_s + D W_s)\, \delta + D \frac{\delta^2}{\alpha}.$$

In particular, in the case of Tikhonov regularization, the upper bound does not depend on the magnitude of the exact or noisy data. On the other hand, from (28) and (29) we do not get a bound independent from the magnitude of b^δ. However, in general, the exact solution of Tikhonov problem is not available in closed form and must be approximated numerically by some iterative algorithm; the main examples are forward-backward (also called ISTA in this context) or accelerated forward-backward (FISTA). When these methods are applied to Tikhonov problem, the distance between the iterate and the solution depends indeed on the initialization and on the magnitude of the data, as for the proposed primal-dual algorithm.

7 Unfeasible case: convergence and stability with respect to a normal solution

In this section, we consider the case where the ideal problem is not feasible, i.e. the linear equation $Ax = b^\star$ does not have a solution. We show that to provide convergence and stability results for Equation (12) it is enough to assume that the normal equation $A^*Ax = A^*b^\star$ has a solution. Indeed, this is the classical setting in ill-posed inverse problems (Engl et al., 1996), but rarely considered in the context of iterative regularization beyond Hilbertian norms. This generalization is especially relevant for infinite dimensional problems.

In the first part of this section we focus on convergence and we refer to a generic data $b \in \mathcal{Y}$ on which the algorithm is run, as the presented results can be applied both to the exact and the inexact data. We denote the set of primal solutions with data b simply as $\mathcal{P}$, the one of dual solutions as $\mathcal{D}$, the Lagrangian as $\mathcal{L}$ and the set of saddle-points as $\mathcal{S}$. Let (x_k, y_k) be the sequence generated by the primal-dual algorithm in Equation (12) with data b. First we have the following result, showing that every weak cluster point of the averaged iterates is a saddle-point. Therefore, if there are no saddle-points, the iterates must diverge.

Corollary 21. *Let Assumption 4 hold. Let (x_k, y_k) be the sequence generated by Equation (12) with data b under Assumption 7, Assumption 8 and summable error ($(\varepsilon_k) \in \ell^1$). Denote by $(\hat{x}_k, \hat{y}_k)$ the averaged iterates. Then, every weak cluster point of $(\hat{x}_k, \hat{y}_k)$ belongs to $\mathcal{S}$. In particular, if $\mathcal{S} = \emptyset$, then the primal-dual sequence $(\hat{x}_k, \hat{y}_k)$ diverges: $\|(\hat{x}_k, \hat{y}_k)\| \to +\infty$.*

The proof can be found in Appendix D.1. The above result ensures that every weak cluster point of the averaged sequence belongs to $\mathcal{S}$ (and so to $\mathcal{P} \times \mathcal{D}$ by Remark 6). Moreover, if there are no primal solutions ($\mathcal{P} = \emptyset$), then $\mathcal{P} \times \mathcal{D} = \emptyset$, $\mathcal{S} = \emptyset$ and so the joint sequence $(\hat{x}_k, \hat{y}_k)$ diverges. Yet we are mainly interested in the primal variable, which may still converge while $\|(\hat{x}_k, \hat{y}_k)\| \to +\infty$. In the sequel, we show sufficient conditions for the averaged primal iterates to converge even when $\mathcal{P} = \emptyset$. For this purpose, we introduce the feasible set and

the *normal* feasible set as

$$\mathcal{C} := \{x \in \mathcal{X} : \quad Ax = b\} \quad \text{and} \quad \tilde{\mathcal{C}} := \{x \in \mathcal{X} : \quad A^*Ax = A^*b\} \quad . \tag{31}$$

It is clear that $\mathcal{C} \subseteq \tilde{\mathcal{C}}$. Moreover, $\mathcal{C} \neq \emptyset$ implies that $\tilde{\mathcal{C}} = \mathcal{C}$. Indeed, let $u \in \tilde{\mathcal{C}}$ and pick any $x \in \mathcal{C}$. Then $A^*Au = A^*b$, $Ax = b$ and $u - x \in N(A^*A) = N(A)$. Thus $Au = Ax = b$; and so $u \in \mathcal{C}$.
In addition to the normal feasible set, define also the *normal* primal problem, its dual, and the *normal* Lagrangian as

$$\tilde{\mathcal{P}} := \operatorname*{argmin}_{x \in \mathcal{X}} \{R(x) + F(x) : \quad A^*Ax = A^*b\}, \tag{32}$$

$$\tilde{\mathcal{D}} := \operatorname*{argmin}_{v \in \mathcal{X}} \{[R + F]^* (-A^*Av) + \langle A^*b, v\rangle\}, \tag{33}$$

$$\tilde{\mathcal{L}}(x, v) := R(x) + F(x) + \langle v, A^*Ax - A^*b\rangle. \tag{34}$$

From $\mathcal{C} \neq \emptyset \implies \tilde{\mathcal{C}} = \mathcal{C}$, we have $\mathcal{C} \neq \emptyset \implies \tilde{\mathcal{P}} = \mathcal{P}$. But it may happen that $\mathcal{C} = \emptyset$ and $\tilde{\mathcal{C}} \neq \emptyset$; and, consequently, that there are no primal solutions ($\mathcal{P} = \emptyset$) but there are normal primal solutions ($\tilde{\mathcal{P}} \neq \emptyset$). Thus in the next results, considering the case $\mathcal{P} = \emptyset$ and $\tilde{\mathcal{P}} \neq \emptyset$, we show convergence and stability with respect to a normal solution. More precisely,

- in Theorem 23, we show sufficient conditions to get convergence of the averaged primal sequence to a point in $\tilde{\mathcal{P}}$ even though $\mathcal{P} = \emptyset$;
- in Theorem 24, we get stability and early-stopping results analogous to the ones in Theorem 13 but with respect to any normal solution.

For simplicity, in the remainder of this section we include neither the preconditioning nor the error in the proximal-operator, setting $T = \tau\,\mathrm{Id}$, $\Sigma = \sigma\,\mathrm{Id}$ and $\varepsilon_k = 0$ for every $k \in \mathbb{N}$. Then our algorithm can be written as: given x_0, y_{-1} and setting $y_0 = y_{-1} + \sigma(Ax_0 - b)$, for every $k \in \mathbb{N}$,

$$\begin{cases} \tilde{y}_k = 2y_k - y_{k-1} \ , \\ x_{k+1} = \operatorname{prox}_{\tau R}(x_k - \tau \nabla F(x_k) - \tau A^* \tilde{y}_k) \ , \\ y_{k+1} = y_k + \sigma\,(Ax_{k+1} - b) \ . \end{cases} \tag{35}$$

Assume that $\tilde{\mathcal{C}} \neq \emptyset$. Let $x^b \in \tilde{\mathcal{C}}$ (meaning that $A^*Ax^b = A^*b$) and let $S := (A^*A)^{\frac{1}{2}}$. The normal problem (32) then can be rewritten as:

$$\tilde{\mathcal{P}} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{R(x) + F(x) : \quad Sx = Sx^b\right\} \ . \tag{36}$$

Indeed, $N(S) = N(S^*S) = N(A^*A)$ and $A^*Ax = A^*b = A^*Ax^b \Leftrightarrow x - x^b \in N(A^*A) = N(S)$.

In Lemma 22, we show that, under mild conditions, the primal variable generated by the algorithm, when applied to problem $\mathcal{P}$, is an instance of the same procedure but applied to the normal problem $\tilde{\mathcal{P}}$ in the form (36).

Lemma 22. *Let Assumption 4 hold. Assume that $\tilde{\mathcal{C}} \neq \emptyset$. Let (x_k) be the primal sequence generated by algorithm (35); namely, with $T = \tau \,\mathrm{Id}$, $\Sigma = \sigma \,\mathrm{Id}$, $\varepsilon_k = 0$ for every $k \in \mathbb{N}$ and $y_0 = y_{-1} + \sigma(Ax_0 - b)$. Then, there exists a primal sequence (u_k) generated by the same procedure but applied to problem $\tilde{\mathcal{P}}$ (as stated in (36)) such that $x_k = u_k$ for every $k \in \mathbb{N}$.*

The proof can be found in Appendix D.2. We are now ready to state the two main results of this section. The first one shows weak convergence of the averaged primal iterate of the algorithm, when applied to $\mathcal{P}$, to a solution of the normal problem $\tilde{\mathcal{P}}$.

Theorem 23. *Let Assumption 4 hold. Assume that $\tilde{\mathcal{P}}$ (as stated in 32) admits a saddle-point; namely, that there exists a pair $(\tilde{x}, \tilde{v}) \in \mathcal{X} \times \mathcal{X}$ such that*

$$\begin{cases} -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x}) \ , \\ A^*A\tilde{x} = A^*b \ . \end{cases} \tag{37}$$

Let (x_k, y_k) be the sequence generated by Equation (35), namely with initialization $y_0 = y_{-1} + \sigma(Ax_0 - b)$, and under Assumption 8. Denote by $(\hat{x}_k)$ the averaged primal iterates. Then there exists $\tilde{x}_\infty \in \tilde{\mathcal{P}}$ such that $\hat{x}_k \rightharpoonup \tilde{x}_\infty$. Moreover, if $\mathcal{P} = \emptyset$, then $\hat{y}_k$ diverges.

The proof can be found in Appendix D.3. Since we assume that the normal problem has a saddle point, a priori we could apply the primal-dual algorithm directly to the normal problem $\tilde{\mathcal{P}}$ and therefore with A^*A in place of A. To fix the ideas, consider the final dimensional setting, in which $A \in \mathbb{R}^{n \times d}$. If $d > n$, as is usual in compressed sensing, working with the matrix A^*A can be disadvantageous.

Two questions remain open from our previous analysis, that we leave as future work. Consider for simplicity the case $F = 0$. From the definition of the primal iterates in the proposed algorithm and the properties of the *prox* operator, we know that, if the domain of R is bounded, then the primal iterates remain bounded. Suppose that the normal equation has solutions, namely $\tilde{\mathcal{C}} \neq \emptyset$. If the domain of R does not intersect $\tilde{\mathcal{C}}$, we expect - but we could not prove - that the primal iterates of the algorithm converge to an element in

$$\operatorname*{argmin}_{x \in \mathrm{dom}(F)} \ \inf_{y \in \tilde{\mathcal{C}}} \|x - y\| \ .$$

On the other hand, now suppose - for instance - that the function R has full domain. We have seen that if the normal problem admits a saddle-point, then the averaged primal sequence converges to an element in $\tilde{\mathcal{P}}$ (see Theorem 23). We expect that, on the contrary, the

absence of solution for the primal normal problem (for instance, if $\tilde{\mathcal{C}} = \emptyset$) implies divergence of the primal iterates. This is the case of the example discussed in Remark 15, but we could not prove it in general.

To conclude this section, we show a stability result for the iterates generated by the algorithm on the noisy data with respect to any saddle-point of the exact *normal* problem. For this theorem we come back to the separated notation $b^\star$ for the exact data and b^δ for the noisy one, while we keep the symbol tilde for normal problems and solutions; for instance, $\tilde{\mathcal{P}}^\star$ will denote the exact normal primal problem, as stated for instance in Equation (32) but with data $b^\star$.

Theorem 24. *Let Assumption 4 hold and suppose that there exists a pair $(\tilde{x}, \tilde{v}) \in \mathcal{X} \times \mathcal{X}$ such that*

$$\begin{cases} -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x}) \ , \\ A^*A\tilde{x} = A^*b^\star \end{cases} \tag{38}$$

(namely, a saddle-point for the normal exact problem $\tilde{\mathcal{P}}^\star$). Let $b^\delta \in \mathcal{Y}$ be a noisy data such that $\left\|b^\delta - b^\star\right\| \leq \delta$ for some $\delta \geq 0$. Moreover, suppose that $\tilde{\mathcal{C}}^\delta \neq \emptyset$; namely, that there exists $x^\delta \in \mathcal{X}$ such that $A^*Ax^\delta = A^*b^\delta$. Let Assumption 8 and Assumption 9 hold and (x_k, y_k) be the sequence generated by the algorithm Equation (35) on the noisy data b^δ; namely, for the initialization $y_0 = y_{-1} + \sigma(Ax_0 - b^\delta)$,

$$\begin{cases} \tilde{y}_k = 2y_k - y_{k-1} \ , \\ x_{k+1} = \operatorname{prox}_{\tau R}(x_k - \tau\nabla F(x_k) - \tau A^*\tilde{y}_k) \ , \\ y_{k+1} = y_k + \sigma\left(Ax_{k+1} - b^\delta\right) . \end{cases}$$

Denote by $(\hat{x}_k)$ the averaged primal iterates. Then,

$$D^{-A^*A\tilde{v}}(\hat{x}_k, \tilde{x}) \leq \frac{C_1}{k} + C_2\delta + C_4\delta^2 k$$

and

$$\|A^*A\hat{x}_k - A^*b^\star\|^2 \leq \|S\|\left[\frac{C_5}{k} + C_6\delta + C_8\delta^2 k + C_9\delta^2\right],$$

where the constants involved in the bounds are specified in the proof.

The proof can be found in Appendix D.4.

Remark 25. *We think that the assumption $\tilde{\mathcal{C}}^\delta \neq \emptyset$ is a technical byproduct of our analysis (we need to assume it to use Lemma 22), but not necessary in order to get the results in Theorem 24.*

Example 26. *It is easy to find an example explaining the meaning and the importance of the previous result. Consider the following setting in* $\mathcal{X} = \mathbb{R}^2$. *Let the inexact linear system* $Ax = b^\delta$ *identify a line on the plane and let* $R : \mathbb{R}^2 \to \mathbb{R}$ *be a convex and lower-semicontinuous function that is an exponential when restricted to the inexact constraint* $\mathcal{C}^\delta$. *Then,* $\tilde{\mathcal{C}}^\delta = \mathcal{C}^\delta \neq \emptyset$ *but* $\tilde{\mathcal{P}}^\delta = \mathcal{P}^\delta = \emptyset$. *In particular we are in a case of severe instability: the averaged primal iterates* $(\hat{x}_k)$, *generated by the algorithm when applied to problem* $\mathcal{P}^\delta = \emptyset$, *may diverge. Now consider the two following scenarios.*

- *Let the exact linear system* $Ax = b^\star$ *identify a line in* $\mathbb{R}^2$ *(parallel to* $\mathcal{C}^\delta$*) and let* $R : \mathbb{R}^2 \to \mathbb{R}$ *be coercive on the exact constraint* $\mathcal{C}^\star$. *Then the primal exact problem admits minimizers (*$\mathcal{P}^\star \neq \emptyset$*), while the noisy one does not have solutions even if it is feasible. In this setting, the assumptions of Theorem 13 hold and thus our early-stopping bounds guarantee an efficient way to find a stable solution.*

- *Now suppose that the exact linear system* $Ax = b^\star$ *does not admit solutions (*$b^\star \notin R(A)$*) and let the exact normal system* $A^*Ax = A^*b^\star$ *identify a line in* $\mathbb{R}^2$. *Moreover, similarly to the previous example, let* $R : \mathbb{R}^2 \to \mathbb{R}$ *be coercive on the exact normal constraint* $\tilde{\mathcal{C}}^\star$. *The primal exact problem does not admit feasible points and so neither minimizers (*$\mathcal{P}^\star = \emptyset$*). Then, in this case, the assumptions in Theorem 13 are not verified. On the other hand, the exact normal problem has solutions (*$\tilde{\mathcal{P}}^\star = \emptyset$*) and* $\tilde{\mathcal{C}}^\delta \neq \emptyset$, *so we still can apply Theorem 24 to get an a similar early-stopping result, but with respect to any exact normal solution.*

8 Experiments

The results of this paper apply to any convex regularizer, with stronger results in the case of the ℓ_1 norm. As a proof of concept, we illustrate them on two seminal problems, the Lasso and low-rank matrix completion. A high quality Python package implementing our iterative regularization approach, with reproducible experiments, is available at `https://lcsl.github.io/iterreg`.

8.1 Sparse recovery with the ℓ_1 norm

First we illustrate numerically the results of Section 6 ($R(\cdot) = \|\cdot\|_1$, $F = 0$) on both real data and simulations. The simulated data is generated as $b^\delta = b^\star + \epsilon = A\bar{x} + \epsilon$. The design matrix A has Gaussian entries with a Toeplitz correlation structure (correlation between columns i and j is $\rho^{|i-j|}$ for $\rho \in [0, 1[$; as ρ approaches 1, the problem becomes more and more difficult). The noise vector ϵ has i.i.d. Gaussian entries, with standard deviation scaled

to control the signal-to-noise ratio (SNR), defined as $\|A\bar{x}\| / \|\epsilon\|$. The true parameter vector $\bar{x}$ has 10 % non zero entries set to 1 ; note that the noiseless solution $x^\star$ is not necessarily $\bar{x}$ – in particular the ℓ_0 and ℓ_1 solutions tend to differ if the feature correlation parameter ρ is too high or if the sparsity of $\bar{x}$ is not low enough. In Algorithm (12), unless specified otherwise, we use exact prox ($\varepsilon_k = 0$), as well as scalar preconditioners $T = \tau\,\mathrm{Id}$ and $\Sigma = \sigma\,\mathrm{Id}$.

The explicit, Tikhonov regularization competitor in this case is the Lasso.

Datadriven choice of stepsize σ. A key distinction between iterative and Tikhonov regularization is that our iterative approach produces discrete iterates, while the Tikhonov path can be discretized with arbitrary precision. Hence, our algorithm could converge too fast to the noisy solution, preventing us from finding a good early stopped iterate. Fortunately, it is possible to act on the dual stepsize σ so that the iterates remain sparse in the beginning (in the same way as, for the Lasso, the solutions are sparse for large regularization strength λ). On Figure 2 we illustrate multiple choices for σ, keeping $\sigma\tau$ equal to $0.99/\|A\|^2$: $\sigma \in \{\tau, \tau/100, 1/\|A^* b^\delta\|_\infty, \tau/10000\}$. The order of magnitude $\sigma = 1/\|A^* b^\delta\|_\infty$ is reversed engineered from the first iterations of (12) with $x_0 = 0$, $y_{-1} = y_0 = 0$, yielding $y_1 = -\sigma b^\delta$ and ensuring that $x_2 = \mathrm{prox}_{\tau\|\cdot\|_1}(2\tau\sigma A^* b^\delta)$ remains sparse enough.
The performance of iterative regularization is measured by the F1 score between the support of the iterates and the support of the true parameters, $\bar{x}$. As visible on Figure 2, the higher σ, the faster the primal iterates x_k become dense, thus overestimating the support of $\bar{x}$. From the figure, one can see that the datadriven choice of σ provides a good balance between quality of the regularization (it reaches the highest F1 score) and convergence speed (optimal score reached after 15 iterations only).

Comparison with the Lasso on simulations. In this experiment, we compare the support recovery performance to that of the Lasso. In order to have a ground truth available, we use a simulated setup. The data for this experiment has 1000 samples and 2000 features. The performance of iterative and Tikhonov regularization is evaluated with the F1 score for support estimation, and normalized mean squared error on left out data (250 additional samples) for prediction, $\|b^{\delta,\,\mathrm{test}} - A^{\mathrm{test}}\|^2 / \|b^{\delta,\,\mathrm{test}}\|^2$. We study two scenarios: an "easy" one (SNR $= 5$, low feature correlation factor $\rho = 0.2$) and a more challenging one (SNR $= 3$, $\rho = 0.8$). On Figure 3, one can see that the estimation and prediction performances are comparable between iterative regularization and explicit regularization, illustrating the numerical guarantees of Section 6.

Timing comparison with the Lasso on real data. Finally, we benchmark our approach on real data, where the true support is unknown and the best model must be selected by cross validation

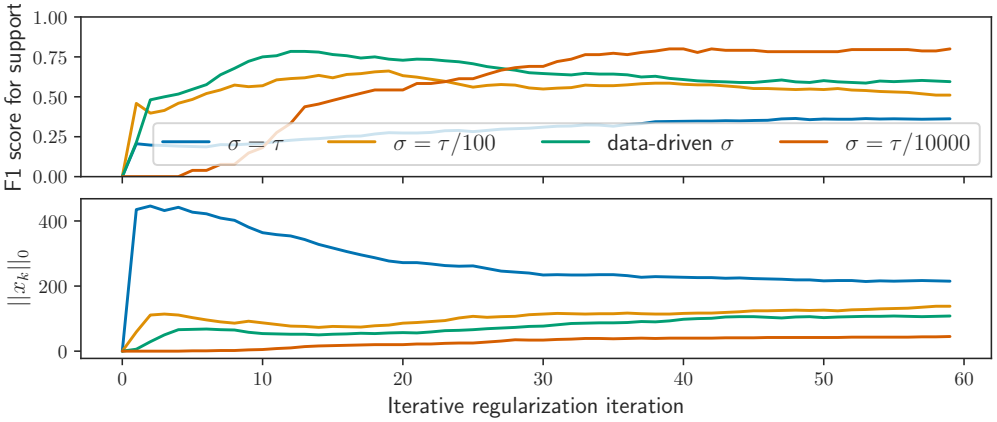

Figure 2: To maintain sparsity in the early iterates, it is important to set σ correctly: if it is too big, the iterates are dense too quickly (blue curve); if it is too low, convergence is too slow (red). Our datadriven choice behaves well: the iterates sparsity increases steadily, and they reach the highest F1 score. $(n, d, \rho) = (200, 500, 0.2)$, $\|A\bar{x}\| / \|\epsilon\| = 10$.

In Figure 4, we compare the quality of solutions obtained by iterative regularization and explicit regularization. The dataset for this experiment is *rcv1* from the LIBSVM package[3], for which $(n, d) = (20{,}242, 19{,}959)$. In order to select the best regularization strength for each approach (iteration or value of λ), we use the prediction mean squared error with 4-fold cross validation: the data (A, b^δ) is split in 4 folds and each method is run 4 times on 3 folds, while the MSE is computed on the remaining, unseen fold (dashed colored lines). The MSE is then averaged across folds (thick black line), and the best iteration/λ is determined by its minimum. Note that this approach does not rely on the knowledge of the true parameters $\bar{x}$ and is thus the one we advocate to use to determine the optimal stopping time in practice.

To solve the Lasso problem, many algorithms are available, the most popular ones being ISTA Daubechies et al. (2004). Important improvements of the previous method were given by coordinate descent (see Wu and Lange (2008)), active set strategies (see Friedman et al. (2010)) and acceleration (see FISTA, Beck and Teboulle (2009)). In the numerical experiments, we use the state-of-the-art solver `celer` (Massias et al., 2020), based on coordinate descent, an active set strategy and Anderson acceleration. Extensive validation in Massias et al. (2020) showed that this algorithm was currently the fastest one available to solve the Lasso (see also Moreau et al. (2022, Section 4). Warm-start is used along the path: the solution for the previous λ is used as initialization for the next one. With all these improvements over a basic forward-backward solver, the time to compute the best solution (the path up to

[3] https://github.com/mathurinm/libsvmdata

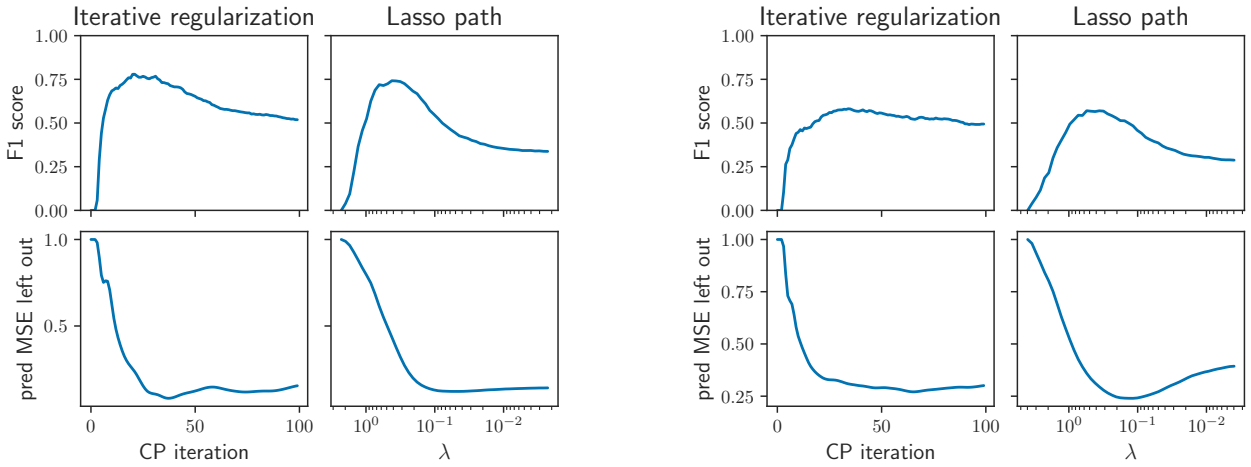

Figure 3: Comparison of estimation and prediction performances of iterative and Tikhonov regularization for sparse recovery. Left: feature correlation factor $\rho = 0.2$, SNR $= 5$. Right: correlation factor $\rho = 0.8, \text{SNR} = 3$. In both scenarios, iterative regularization attains performances similar to explicit regularization, but in a few iterations.

the best λ, if it were known in advance) is 125 seconds. This is because 69 Lasso problems must be solved (the optimal λ is the 69-th on the grid), each one being increasingly difficult as λ decreases.

On the contrary, iterative regularization finds its optimal solution along the optimization path in 2.5 s. The cost of each iteration is $\mathcal{O}(nd)$, making the algorithm very fast. One can see that in terms of prediction error on left-out data (4-fold cross validation being used to determine both the best λ for the Lasso and the best early stopping for our approach), both methods reach a similar performance, with a best average MSE around 0.2. In addition, using our proposed datadriven stepsize, we obtain a sparser solution than the Lasso: ours has 1,583 non zeros entries, while the optimal Lasso one has 2,820.

As mentioned above, non convex penalties such as MCP Zhang (2010), SCAD (Fan and Li, 2001) or CNC (Lanza et al., 2019) have a better perfomance in practice in terms of support recovery than the ℓ_1 norm. Our goal here is to compare explicit regularization to iterative regularization, in the convex case, which is the setting covered by our theoretical results. An exhaustive comparison of sparse recovery methods would be worthy of a separate effort.

27

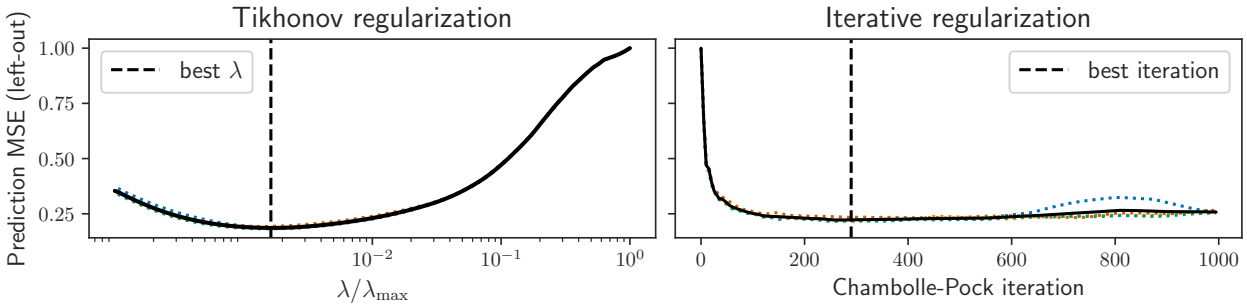

Figure 4: Comparison of Tikhonov regularization and iterative regularization for sparse recovery. The figure of merit is 4-fold cross validation prediction error. Both methods reach similar lowest prediction errors (left: 0.195, right: 0.21) while the iterative approach is much faster (2.5 s vs. 125 s).

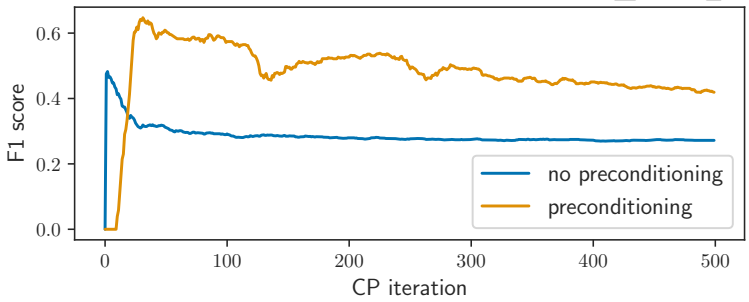

Figure 5: Benefit of preconditioning for sparse recovery on an unnormalized simulated dataset. $(nd) = (500, 1000)$.

8.2 Preconditioning

In this experiment we highlight the usefulness of a preconditioning. We consider two diagonal preconditioners, following Pock and Chambolle (2011): $T = \theta \operatorname{diag}(||A_{:1}||^2, \ldots, ||A_{:d}||^2)$ and $\Sigma = \frac{1}{\theta} \operatorname{diag}(\|A_{1:}\|_0, \ldots, \|A_{n:}\|_0) = \frac{d}{\theta} \operatorname{Id}$. The scaling factor θ is set to get σ as in the datadriven choice detailed above. This choice of T and Σ satisfies $\tau_M \sigma_M \leq 1/\|A\|^2$ (Pock and Chambolle, 2011, Lemma 2). The design matrix A is generated as in Section 8.1, but each column is then scaled by a uniform random number between 1 and 5, resulting in different column norms and thus in T being different from a scalar matrix. On Figure 5, one an see that using coordinate-wise stepsizes through the use of T in the update of the primal variable, is beneficial for iterative regularization as a higher F1 score is reached.

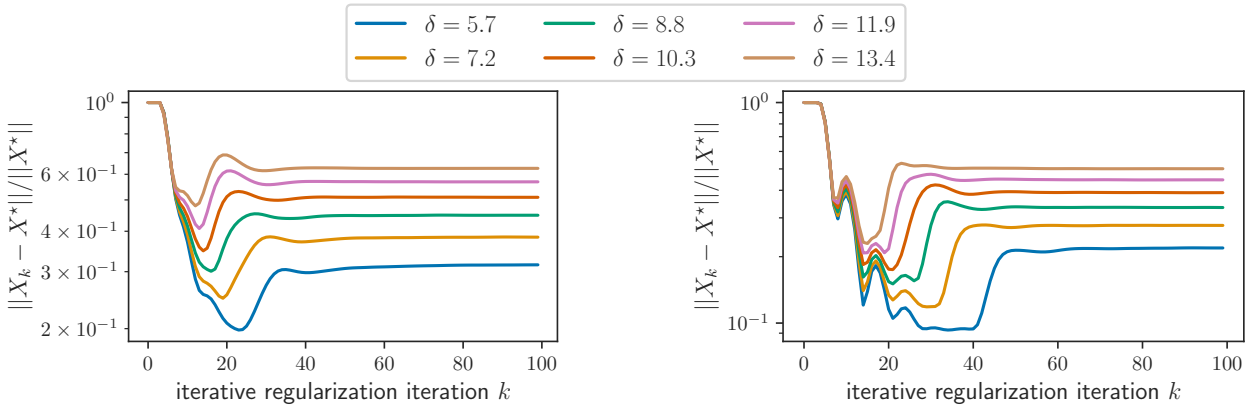

Figure 6: Semiconvergence of iterates for the low rank matrix completion problem, in dimension 200×200 (left) and 500×500 (right). The iterates first get close to the noiseless solution, before converging to the noisy solution.

8.3 Low rank matrix completion

In this experiment we highlight the versatility of our approach, considering the matrix completion setting of Example 2. The goal is to recover a low-rank matrix from the noisy observation of a subset of its entries. Both Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ are taken equal to $\mathbb{R}^{d\times d}$, and we use upper case letters X and B to denote the primal variable and the observations. The true matrix to recover is chosen as $B^\star = UV^\top$ where $U, V \in \mathbb{R}^{d\times 5}$ have i.i.d. normal entries. In order to get meaningful values for δ, we scale $B^\star$ so that it has a norm equal to 20. Finally, for a range of values of δ, various B^δ are obtained by adding scaled random Gaussian noise to the observed entries of $B^\star$. We choose to hide 80 % of entries of B^δ, uniformly sampled. The matrix A corresponds to the masking operator; we have $\|A\|_2 = 1$ and thus use $\sigma = \tau = 0.99$. ~~We tune the parameter σ similarly to the ℓ_1 case, taking $\sigma = 1/\|A^*B^\delta\|_2$.~~ Figure 6 highlights the semiconvergence behavior exploited by iterative regularization: the iterates produced by (12) first get closer to the noiseless solution, before converging to the noisy solution. Early-stopping the iterate at a correct iteration is thus beneficial.

Finally, Figure 7 is the equivalent of Figure 4 for low rank matrix recovery. The true matrix is 50×50. It is the sum of a rank 5 matrix and random Gaussian noise, with variance such that the SNR is 3. 40 % of the indices are randomly masked. On the remaining 60 %, a train-test split is performed (using 3/4th of the available observations for training and the remaining 1/4th for validation). Both explicit and iterative approaches are run on the train set, and their performance evaluated on the test set. Note that, as discussed above, this approach is implementable in real life: some part of the available data (the validation set) is kept apart for hyperparameter tuning. The competitor algorithm for explicit regularization

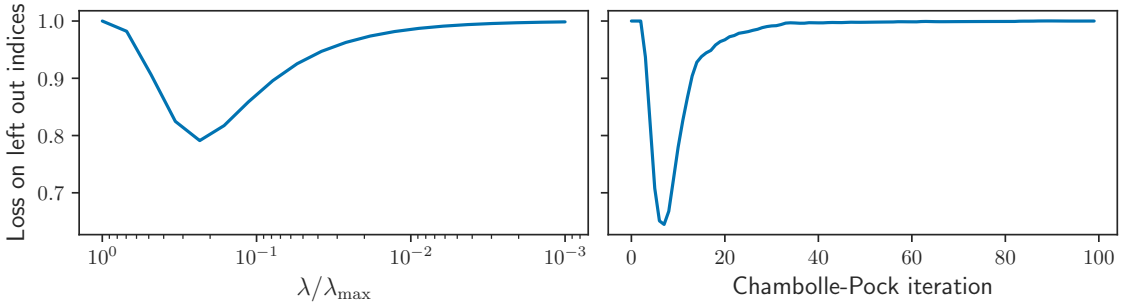

Figure 7: Comparison of Tikhonov regularization and iterative regularization for low rank matrix recovery. The figure of merit is error on left-out data. Iterative regularization outperforms explicit regularization, while requiring significantly less time (0.2 s vs 100 s).

is proximal gradient descent, with warm-start along the path. As visible on the figure, iterative regularization is able to achieve better results than explicit regularization in that case, while solving a single problem, thus being faster.

9 Conclusion

In this work, we have considered the problem of designing iterative regularization algorithms for bias described by a wide class of convex functionals. We proposed and study an iterative regularization method based on a primal-dual approach of which we characterize convergence and especially stability in the presence of noisy data. This latter results allow to derive and early stopping procedure and corresponding error bounds, comparable with those obtainable with variational regularization techniques. Empirical results complement and confirm our theoretical findings, showing that iterative regularization can be at the same time accurate and efficient.

A number of research directions remains unexplored. For example it would be interesting to consider stochastic gradient approaches, that often results in further efficiency improvement. It would also be interesting to extend the considered model to account for other form of noise/errors, including data models in machine learning, but also considering other, possibly non convex, penalties. Finally, it would be interesting to consider nonlinear models, and in particular compositional models such as those defining neural networks.

30

A Preliminary lemmas

Lemma 27 ((Schmidt et al., 2011, Lemma 2)). *Assume that (u_j) is a non-negative sequence, (S_j) is a non-decreasing sequence with $S_0 \geq u_0^2$ and $\lambda \geq 0$ such that, for every $j \in \mathbb{N}$,*

$$u_j^2 \leq S_j + \lambda \sum_{i=1}^{j} u_i \ . \tag{39}$$

Then, for every $j \in \mathbb{N}$,

$$u_j \leq \frac{\lambda j}{2} + \sqrt{S_j + \left(\frac{\lambda j}{2}\right)^2} \ . \tag{40}$$

Lemma 28 (Descent lemma, (Bauschke and Combettes, 2011, Thm 18.15 (iii))). *Let $f : \mathcal{X} \to \mathbb{R}$ be Fréchet differentiable with L-Lipschitz continuous gradient. Then, for every x and $y \in \mathcal{X}$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \ . \tag{41}$$

Lemma 29. *Let $\mathcal{Z}$ denote $\mathcal{X}$ or $\mathcal{Y}$ and U denote T or Σ accordingly. Let $f \in \Gamma_0(\mathcal{Z})$ and $\varepsilon \geq 0$. It follows easily from the definition of the ε-subdifferential that if $a, b \in \mathcal{Z}$ satisfy*

$$U^{-1}(a - b) \in \partial_\varepsilon f(b) \ , \tag{42}$$

then, for every $c \in \mathcal{Z}$,

$$f(b) - f(c) + \frac{1}{2}\|b - c\|_U^2 - \frac{1}{2}\|a - b\|_U^2 + \frac{1}{2}\|b - a\|_U^2 \leq \varepsilon \ . \tag{43}$$

A.1 Primal-dual estimates

Lemma 30 (One step estimate). *Let Assumption 4 hold. Let (x_k, y_k) be the sequence generated by iterations (12) under Assumption 7. Then, for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and for any $k \in \mathbb{N}$, with $V(z) := \frac{1}{2}\|x\|_T^2 + \frac{1}{2}\|y\|_\Sigma^2$,*

$$\begin{aligned} V(z_{k+1} - z) - V(z_k - z) + \frac{1 - \tau_M L}{2\tau_M}\|x_{k+1} - x_k\|^2 + \frac{1}{2}\|y_{k+1} - y_k\|_\Sigma^2 \\ + \left[\mathcal{L}^\delta(x_{k+1}, y) - \mathcal{L}^\delta(x, y_{k+1})\right] + \langle y_{k+1} - \tilde{y}_k, A(x - x_{k+1})\rangle \leq \varepsilon_{k+1} \ . \end{aligned} \tag{44}$$

Proof. Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Applying Lemma 29 to the definition of x_{k+1} yields

$$\begin{aligned} \frac{1}{2}\|x_{k+1} - x\|_T^2 - \frac{1}{2}\|x_k - x\|_T^2 + \frac{1}{2}\|x_{k+1} - x_k\|_T^2 + [R(x_{k+1}) - R(x)] \\ + \langle \tilde{y}_k, A(x_{k+1} - x)\rangle + \langle \nabla F(x_k), x_{k+1} - x\rangle \ \leq \ \varepsilon_{k+1} \ . \end{aligned} \tag{45}$$

For the dual update, similarly,

$$\frac{1}{2}\left\|y_{k+1}-y\right\|_{\Sigma}^{2}-\frac{1}{2}\left\|y_{k}-y\right\|_{\Sigma}^{2}+\frac{1}{2}\left\|y_{k+1}-y_{k}\right\|_{\Sigma}^{2}+\left\langle y_{k+1}-y,b^{\delta}-Ax_{k+1}\right\rangle\leq 0\ . \tag{46}$$

Recall that $z:=(x,y)$ and the definition of V. Sum Equations (45) and (46):

$$\begin{aligned}&V(z_{k+1}-z)-V(z_{k}-z)+V(z_{k+1}-z_{k})+\left[R\left(x_{k+1}\right)-R(x)\right]\\&+\left\langle\tilde{y}_{k},A\left(x_{k+1}-x\right)\right\rangle+\left\langle y_{k+1}-y,b^{\delta}-Ax_{k+1}\right\rangle+\left\langle\nabla F(x_{k}),x_{k+1}-x\right\rangle\leq\varepsilon_{k+1}\ .\end{aligned} \tag{47}$$

From the Lemma 28,

$$F(x_{k+1})\leq F(x_{k})+\left\langle\nabla F(x_{k}),x_{k+1}-x_{k}\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2}\ ,$$

while from the convexity of F,

$$F(x_{k})+\left\langle\nabla F(x_{k}),x-x_{k}\right\rangle\leq F(x)\ .$$

Summing the last two equations, one obtains the 3 points descent lemma:

$$F(x_{k+1})\leq F(x)+\left\langle\nabla F(x_{k}),x_{k+1}-x\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2}\ . \tag{48}$$

Summing Equations (47) and (48),

$$\begin{aligned}&V(z_{k+1}-z)-V(z_{k}-z)+V(z_{k+1}-z_{k})\\&+\left[R+F\right]\left(x_{k+1}\right)-\left[R+F\right](x)+\left\langle\tilde{y}_{k},A\left(x_{k+1}-x\right)\right\rangle+\left\langle y_{k+1}-y,b^{\delta}-Ax_{k+1}\right\rangle\\\leq\ &\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2}+\varepsilon_{k+1}\ .\end{aligned}$$

Now compute

$$\begin{aligned}&\left[R+F\right]\left(x_{k+1}\right)-\left[R+F\right](x)+\left\langle\tilde{y}_{k},A\left(x_{k+1}-x\right)\right\rangle+\left\langle y_{k+1}-y,b^{\delta}-Ax_{k+1}\right\rangle\\=\ &\left[\mathcal{L}^{\delta}(x_{k+1},y)-\mathcal{L}^{\delta}(x,y_{k+1})\right]-\left\langle y,Ax_{k+1}-b^{\delta}\right\rangle+\left\langle y_{k+1},Ax-b^{\delta}\right\rangle\\&+\left\langle\tilde{y}_{k},A\left(x_{k+1}-x\right)\right\rangle+\left\langle y_{k+1}-y,b^{\delta}-Ax_{k+1}\right\rangle\\=\ &\left[\mathcal{L}^{\delta}(x_{k+1},y)-\mathcal{L}^{\delta}(x,y_{k+1})\right]-\left\langle y_{k+1}-y,b^{\delta}\right\rangle-\left\langle y,Ax_{k+1}\right\rangle+\left\langle y_{k+1},Ax\right\rangle\\&+\left\langle\tilde{y}_{k},Ax_{k+1}\right\rangle-\left\langle\tilde{y}_{k},Ax\right\rangle+\left\langle y_{k+1}-y,b^{\delta}\right\rangle-\left\langle y_{k+1}-y,Ax_{k+1}\right\rangle\\=&\left[\mathcal{L}^{\delta}(x_{k+1},y)-\mathcal{L}^{\delta}(x,y_{k+1})\right]\\&-\left\langle y,Ax_{k+1}\right\rangle+\left\langle y_{k+1},Ax\right\rangle+\left\langle\tilde{y}_{k},Ax_{k+1}\right\rangle-\left\langle\tilde{y}_{k},Ax\right\rangle-\left\langle y_{k+1},Ax_{k+1}\right\rangle+\left\langle y,Ax_{k+1}\right\rangle\\=\ &\left[\mathcal{L}^{\delta}(x_{k+1},y)-\mathcal{L}^{\delta}(x,y_{k+1})\right]+\left\langle y_{k+1}-\tilde{y}_{k},A\left(x-x_{k+1}\right)\right\rangle\ .\end{aligned}$$

Notice that

$$\frac{1}{2\tau_M}\left\|x_{k+1}-x_k\right\|^2 \leq \frac{1}{2}\left\|x_{k+1}-x_k\right\|_T^2 \quad . \tag{49}$$

Finally,

$$\begin{aligned} V(z_{k+1}-z)-V(z_k-z)+\frac{1-\tau_M L}{2\tau_M}\left\|x_{k+1}-x_k\right\|^2+\frac{1}{2}\left\|y_{k+1}-y_k\right\|_\Sigma^2 \\ +\left[\mathcal{L}^\delta(x_{k+1},y)-\mathcal{L}^\delta(x,y_{k+1})\right]+\left\langle y_{k+1}-\tilde{y}_k, A\left(x-x_{k+1}\right)\right\rangle \leq \varepsilon_{k+1} \quad . \end{aligned}$$

□

Lemma 31 (First cumulating estimate)**.** *Let Assumption 4 hold. Let (x_k, y_k) be the sequence generated by iterations (12) under Assumption 7. Define $\omega := 1-\tau_M(L+\sigma_M\|A\|^2)$. Then, for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and for any $k\in\mathbb{N}$,*

$$\begin{aligned} &\tfrac{1-\tau_M\sigma_M\|A\|^2}{2\tau_M}\left\|x_k-x\right\|^2+\frac{1}{2}\left\|y_k-y\right\|_\Sigma^2+\sum_{j=1}^{k}\left[\mathcal{L}^\delta(x_j,y)-\mathcal{L}^\delta(x,y_j)\right]+\frac{\omega}{2\tau_M}\sum_{j=1}^{k}\left\|x_j-x_{j-1}\right\|^2 \\ &\leq V(z_0-z)+\sum_{j=1}^{k}\varepsilon_j. \end{aligned} \tag{50}$$

Proof. We start from the inequality in Lemma 30, switching the index from k to j. Recall that $\tilde{y}_j := 2y_j - y_{j-1}$, to get

$$\begin{aligned} &V(z_{j+1}-z)-V(z_j-z)+\frac{1-\tau_M L}{2\tau_M}\left\|x_{j+1}-x_j\right\|^2+\frac{1}{2}\left\|y_{j+1}-y_j\right\|_\Sigma^2 \\ &\qquad +\left[\mathcal{L}^\delta(x_{j+1},y)-\mathcal{L}^\delta(x,y_{j+1})\right] \\ \leq\ & \varepsilon_{j+1}-\left\langle y_{j+1}-(2y_j-y_{j-1}), A\left(x-x_{j+1}\right)\right\rangle \\ =\ & \varepsilon_{j+1}-\left\langle y_{j+1}-y_j, A\left(x-x_{j+1}\right)\right\rangle+\left\langle y_j-y_{j-1}, A\left(x-x_{j+1}\right)\right\rangle \\ =\ & \varepsilon_{j+1}-\left\langle y_{j+1}-y_j, A\left(x-x_{j+1}\right)\right\rangle+\left\langle y_j-y_{j-1}, A\left(x-x_j\right)\right\rangle+\left\langle y_j-y_{j-1}, A\left(x_j-x_{j+1}\right)\right\rangle \quad . \end{aligned}$$

Now focus on the term

$$\begin{aligned} \left\langle y_j-y_{j-1}, A\left(x_j-x_{j+1}\right)\right\rangle &= \left\langle \Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(y_j-y_{j-1}\right), A\left(x_j-x_{j+1}\right)\right\rangle \\ &= \left\langle \Sigma^{-\frac{1}{2}}\left(y_j-y_{j-1}\right), \Sigma^{\frac{1}{2}}A\left(x_j-x_{j+1}\right)\right\rangle \\ &\leq \left\|\Sigma^{-\frac{1}{2}}\left(y_j-y_{j-1}\right)\right\|\left\|\Sigma^{\frac{1}{2}}A\left(x_j-x_{j+1}\right)\right\| \\ &\leq \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\left(y_j-y_{j-1}\right)\right\|^2+\frac{1}{2}\left\|\Sigma^{\frac{1}{2}}A\left(x_j-x_{j+1}\right)\right\|^2 \\ &\leq \frac{1}{2}\left\|y_j-y_{j-1}\right\|_\Sigma^2+\frac{\sigma_M\|A\|^2}{2}\left\|x_{j+1}-x_j\right\|^2 \quad , \end{aligned} \tag{51}$$

where we used Cauchy-Schwarz and Young inequalities. Then, using the definition of $\omega := 1 - \tau_M(L + \sigma_M \|A\|^2)$, we have

$$\begin{aligned} & V(z_{j+1} - z) - V(z_j - z) + [\mathcal{L}(x_{j+1}, y) - \mathcal{L}(x, y_{j+1})] \\ & + \frac{\omega}{2\tau_M} \|x_{j+1} - x_j\|^2 + \frac{1}{2} \|y_{j+1} - y_j\|_\Sigma^2 - \frac{1}{2} \|y_j - y_{j-1}\|_\Sigma^2 \\ \leq \; & \varepsilon_{j+1} - \langle y_{j+1} - y_j, A(x - x_{j+1}) \rangle + \langle y_j - y_{j-1}, A(x - x_j) \rangle \; . \end{aligned} \tag{52}$$

Imposing $y_{-1} = y_0$, summing-up Equation (52) from $j = 0$ to $j = k - 1$:

$$\begin{aligned} & V(z_k - z) - V(z_0 - z) + \sum_{j=0}^{k-1} \left[\mathcal{L}^\delta(x_{j+1}, y) - \mathcal{L}^\delta(x, y_{j+1})\right] + \frac{\omega}{2\tau_M} \sum_{j=0}^{k-1} \|x_{j+1} - x_j\|^2 \\ & \qquad + \frac{1}{2} \|y_k - y_{k-1}\|_\Sigma^2 \\ & \leq \sum_{j=0}^{k-1} \varepsilon_{j+1} - \langle y_k - y_{k-1}, A(x - x_k) \rangle \\ & \leq \frac{1}{2} \|y_k - y_{k-1}\|_\Sigma^2 + \frac{\sigma_M \|A\|^2}{2} \|x_k - x\|^2 + \sum_{j=1}^{k} \varepsilon_j \; , \end{aligned}$$

where in the last inequality we used again Cauchy-Schwarz and Young inequalities as before. Reordering, we obtain the claim. □

Lemma 32 (Second cumulative estimate)**.** *Let Assumption 4 hold. Let (x_k, y_k) be the sequence generated by iterations (12) under Assumption 7. Given $\xi > 0$ and $\eta > 0$, define $\theta := \xi - \tau_M(\xi L + \sigma_M \|A\|^2)$ and $\rho := \sigma_m(\eta - 1) - \sigma_M \xi \eta$. Then, for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and for any $k \in \mathbb{N}$,*

$$\begin{aligned} & V(z_k - z) + \frac{\theta}{2\tau_M \xi} \sum_{j=1}^{k} \|x_j - x_{j-1}\|^2 + \frac{\rho}{2\eta} \sum_{j=1}^{k} \|Ax_j - Ax\|^2 + \sum_{j=1}^{k} \left[\mathcal{L}^\delta(x_j, y) - \mathcal{L}^\delta(x, y_j)\right] \\ & \leq V(z_0 - z) + \sum_{j=1}^{k} \varepsilon_j + \frac{\sigma_m (\eta - 1) k}{2} \left\|Ax - b^\delta\right\|^2 \; . \end{aligned} \tag{53}$$

Proof. In a similar fashion as in the previous proof, we start again from the main inequality in Lemma 30, switching the index from k to j. Since $\tilde{y}_j = y_j + (y_j - y_{j-1}) = y_j + \Sigma(Ax_j - b^\delta)$

and $y_{j+1} - y_j = \Sigma(Ax_{j+1} - b^\delta)$, we get

$$\begin{aligned} &V(z_{j+1} - z) - V(z_j - z) + \frac{1 - \tau_M L}{2\tau_M} \|x_{j+1} - x_j\|^2 + \frac{1}{2} \left\|\Sigma \left(Ax_{j+1} - b^\delta\right)\right\|_\Sigma^2 \\ &\qquad + \left[\mathcal{L}^\delta(x_{j+1}, x) - \mathcal{L}^\delta(x, y_{j+1})\right] \\ &\leq \varepsilon_{j+1} + \langle y_{j+1} - y_j - \Sigma \left(Ax_j - b^\delta\right), Ax_{j+1} - Ax\rangle \\ &= \varepsilon_{j+1} + \langle \Sigma A \left(x_{j+1} - x_j\right), Ax_{j+1} - Ax\rangle \ . \end{aligned}$$

Now estimate

$$\begin{aligned} \frac{1}{2} \left\|\Sigma \left(Ax_{j+1} - b^\delta\right)\right\|_\Sigma^2 &= \frac{1}{2} \langle \Sigma \left(Ax_{j+1} - b^\delta\right), Ax_{j+1} - b^\delta\rangle \\ &\geq \frac{\sigma_m}{2} \left\|Ax_{j+1} - b^\delta\right\|^2 \\ &= \frac{\sigma_m}{2} \|Ax_{j+1} - Ax\|^2 + \frac{\sigma_m}{2} \left\|Ax - b^\delta\right\|^2 + \sigma_m \langle Ax_{j+1} - Ax, Ax - b^\delta\rangle \ . \end{aligned}$$

So,

$$\begin{aligned} &V(z_{j+1} - z) - V(z_j - z) + \frac{1 - \tau_M L}{2\tau_M} \|x_{j+1} - x_j\|^2 + \frac{\sigma_m}{2} \|Ax_{j+1} - Ax\|^2 \\ &\qquad + \left[\mathcal{L}^\delta(x_{j+1}, y) - \mathcal{L}^\delta(x, y_{j+1})\right] \\ &\leq \varepsilon_{j+1} + \langle \Sigma A \left(x_{j+1} - x_j\right), Ax_{j+1} - Ax\rangle + \sigma_m \langle Ax_{j+1} - Ax, b^\delta - Ax\rangle - \frac{\sigma_m}{2} \left\|Ax - b^\delta\right\|^2 \\ &\leq \varepsilon_{j+1} + \frac{\sigma_M \|A\|^2}{2\xi} \|x_{j+1} - x_j\|^2 + \frac{\xi \sigma_M}{2} \|Ax_{j+1} - Ax\|^2 - \frac{\sigma_m}{2} \left\|Ax - b^\delta\right\|^2 \\ &\quad + \frac{\sigma_m}{2\eta} \|Ax_{j+1} - Ax\|^2 + \frac{\sigma_m \eta}{2} \left\|Ax - b^\delta\right\|^2 \ . \end{aligned}$$

In the last inequality we used three times Cauchy-Schwarz inequality and twice Young inequality with parameters $\xi > 0$ and $\eta > 0$. Then, reordering and recalling the definitions of $\theta := \xi - \tau_M(\xi L + \sigma_M \|A\|^2)$, we obtain

$$\begin{aligned} &V(z_{j+1} - z) - V(z_j - z) + \frac{\theta}{2\tau_M \xi} \|x_{j+1} - x_j\|^2 + \frac{\sigma_m(\eta - 1) - \sigma_M \xi \eta}{2\eta} \|Ax_{j+1} - Ax\|^2 \\ &+ \left[\mathcal{L}^\delta(x_{j+1}, y) - \mathcal{L}^\delta(x, y_{j+1})\right] \quad \leq \quad \varepsilon_{j+1} + \frac{\sigma_m(\eta - 1)}{2} \left\|Ax - b^\delta\right\|^2 \ . \end{aligned}$$

Summing-up the latter from $j = 0$ to $j = k-1$, we get

$$V(z_k - z) - V(z_0 - z) + \frac{\theta}{2\tau_M \xi} \sum_{j=0}^{k-1} \|x_{j+1} - x_j\|^2 + \frac{\sigma_m(\eta-1) - \sigma_M \xi \eta}{2\eta} \sum_{j=0}^{k-1} \|Ax_{j+1} - Ax\|^2$$
$$+ \sum_{j=0}^{k-1} \left[\mathcal{L}^\delta(x_{j+1}, y) - \mathcal{L}^\delta(x, y_{j+1})\right] \quad \leq \quad \sum_{j=0}^{k-1} \varepsilon_{j+1} + \frac{\sigma_m\,(\eta-1)\,k}{2} \left\|Ax - b^\delta\right\|^2 \quad .$$

By trivial manipulations, we get the claim. □

B Proofs of main results

B.1 Proof of Proposition 10

Proposition 10. *Assume that Assumptions 4 and 5 hold. Let (x_k, y_k) be the sequence generated by iterations (12) applied to $b^\delta = b^\star$ under Assumptions 7 and 8. Let also $\varepsilon_k = 0$ for every $k \in \mathbb{N}$. Then (x_k, y_k) weakly converges to a pair in $\mathcal{S}^\star$. In particular, (x_k) weakly converges to a point in $\mathcal{P}^\star$.*

Proof. Up to a change of initialization and offset of index, the steps of algorithm (12) when $\varepsilon_k = 0$ correspond to

$$\begin{cases} y_{k+1} = y_k + \Sigma\,(Ax_k - b^\star) \\ x_{k+1} = \mathrm{prox}_R^T(x_k - T\nabla F(x_k) - TA^*(2y_{k+1} - y_k)) \quad . \end{cases} \tag{54}$$

We now show that the previous iterations correspond to Algorithm 3.2 in Condat (2013), setting $\sigma = \tau = 1$ and applying it in the metrics defined by the preconditioning operators; namely, in the primal and dual spaces $(\mathcal{X},\ \langle T^{-1}\cdot,\cdot\rangle)$ and $(\mathcal{Y},\ \langle\Sigma\cdot,\cdot\rangle)$ - respectively. Comparing problem (15) with (1) in Condat (2013), their notation in our setting reads as $F = F,\ G = R,\ H = \iota_{\{b^\star\}}$ and $K = A$. The Fenchel conjugate of H in $(\mathcal{Y},\ \langle\Sigma\cdot,\cdot\rangle)$ is

$$H^\star(y) = \sup_{z\in\mathcal{Y}} \left\{\langle\Sigma z, y\rangle - \iota_{\{b^\star\}}(z)\right\} = \langle\Sigma b^\star, y\rangle \tag{55}$$

and its proximal-point operator, again in $(\mathcal{Y},\ \langle\Sigma\cdot,\cdot\rangle)$, is

$$\mathrm{prox}_{H^\star}(y) = \operatorname*{argmin}_{z\in\mathcal{Y}} \left\{\langle\Sigma b^\star, z\rangle + \frac{1}{2}\langle\Sigma(z-y), z-y\rangle\right\} = y - b^\star \quad . \tag{56}$$

The gradient of F in $(\mathcal{X},\ \langle T^{-1}\cdot,\cdot\rangle)$ is denoted by $\nabla_T F(x)$ and satisfies, for x and v in $\mathcal{X}$,

$$\langle T^{-1}\nabla_T F(x), v\rangle = \langle\nabla F(x), v\rangle \quad .$$

It is easy to see that one has $\nabla_T F(x) = T\nabla F(x)$.
The adjoint operator of $K : (\mathcal{X}, \langle T^{-1}\cdot,\cdot\rangle) \to (\mathcal{Y}, \langle \Sigma\cdot,\cdot\rangle)$ satisfies, for every $(x,y) \in \mathcal{X}\times\mathcal{Y}$,

$$\langle T^{-1}K^*y, x\rangle = \langle \Sigma Kx, y\rangle = \langle \Sigma Ax, y\rangle = \langle x, A^*\Sigma y\rangle \ , \tag{57}$$

implying that $T^{-1}K^* = A^*\Sigma$ and so that $K^* = TA^*\Sigma$. Then Algorithm 3.2 in Condat (2013) (with $\sigma = \tau = 1$, $\rho_k = 1$ for every $k \in \mathbb{N}$ and no errors involved) is:

$$\begin{cases} \bar{y}_{k+1} = \mathrm{prox}_{H^\star}(\bar{y}_k + K\bar{x}_k) \\ \bar{x}_{k+1} = \mathrm{prox}_R(\bar{x}_k - \nabla_T F(\bar{x}_k) - K^*(2\bar{y}_{k+1} - \bar{y}_k)) \ , \end{cases}$$

and becomes, applied to our setting in the spaces $(\mathcal{X}, \langle T^{-1}\cdot,\cdot\rangle)$ and $(\mathcal{Y}, \langle\Sigma\cdot,\cdot\rangle)$,

$$\begin{cases} \bar{y}_{k+1} = \bar{y}_k + A\bar{x}_k - b^\star \\ \bar{x}_{k+1} = \mathrm{argmin}_{x\in\mathcal{X}} \left\{ R(x) + \frac{1}{2}\left\| x - [\bar{x}_k - T\nabla F(\bar{x}_k) - TA^*\Sigma(2\bar{y}_{k+1} - \bar{y}_k)]\right\|^2_{T^{-1}} \right\} \ . \end{cases}$$

Define the variable $\bar{z}_k = \Sigma\bar{y}_k$ and multiply the first line by Σ. Then,

$$\begin{cases} \bar{z}_{k+1} = \bar{z}_k + \Sigma\left(A\bar{x}_k - b^\star\right) \\ \bar{x}_{k+1} = \mathrm{prox}^T_R\left(\bar{x}_k - T\nabla F(\bar{x}_k) - TA^*(2\bar{z}_{k+1} - \bar{z}_k)\right) \ . \end{cases}$$

Comparing the previous with (54), we get that they are indeed the same algorithm. To conclude, we want to use Theorem 3.1 in Condat (2013), that ensures the weak convergence of the sequence generated by the algorithm to a saddle-point. It remains to check that, under our assumptions, the hypothesis of the above result are indeed satisfied; namely, that

$$1 - \|K\|^2 - \frac{L_T}{2} \geq 0 \ , \tag{58}$$

where $\|K\|$ represents the operator norm of $K : (\mathcal{X}, \langle T^{-1}\cdot,\cdot\rangle) \to (\mathcal{Y}, \langle\Sigma\cdot,\cdot\rangle)$ and L_T is the Lipschitz constant of $\nabla_T F$. Notice that

$$\|K\|^2 = \sup_{x\in\mathcal{X}} \frac{\langle \Sigma Ax, Ax\rangle}{\langle T^{-1}x, x\rangle} \leq \sigma_M\tau_M \|A\|^2 \ .$$

Moreover, $L_T \leq \tau_M L$. Indeed, for every x and $x' \in \mathcal{X}$,

$$\|\nabla_T F(x') - \nabla_T F(x)\| = \|T\nabla F(x') - T\nabla F(x)\| \leq \tau_M \|\nabla F(x') - \nabla F(x)\| \ .$$

Then, by Assumption 8 and the previous considerations,

$$0 \leq 1 - \tau_M(L + \sigma_M\|A\|^2) \leq 1 - L_T - \|K\|^2 \leq 1 - \frac{L_T}{2} - \|K\|^2 \ .$$

In particular, (58) is satisfied and we the claim is proved. □

B.2 Proof of Proposition 11

Proposition 11. *Let $(x^\star, y^\star) \in \mathcal{S}^\star$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y) = 0$ and $Ax = b^\star$. Then $(x, y^\star) \in \mathcal{S}^\star$.*

Proof. For simplicity, denote $J := R + F$. First notice that, for our problem, the Lagrangian gap is equal to the Bregman divergence. Indeed, using $-A^*y^\star \in \partial J(x^\star)$ and $Ax^\star = b^\star$:

$$\begin{aligned}\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y) &= J(x) - J(x^\star) + \langle y^\star, Ax - b^\star\rangle - \langle y, Ax^\star - b^\star\rangle \\ &= J(x) - J(x^\star) + \langle A^*y^\star, x - x^\star\rangle = D_J^{-A^*y^\star}(x, x^\star) \ , \end{aligned} \tag{59}$$

We then show that if $v \in \partial J(x^\star)$ and $D_J^v(x, x^\star) = 0$, then $v \in \partial J(x)$. Indeed, $J(x) - J(x^\star) - \langle v, x - x^\star\rangle = 0$ and so, for all $x' \in \mathcal{X}$,

$$J(x') \geq J(x^\star) + \langle v, x' - x^\star\rangle = J(x) - \langle v, x - x^\star\rangle + \langle v, x' - x^\star\rangle = J(x) + \langle v, x' - x\rangle \ . \tag{60}$$

Proposition 11 follows by taking $v = -A^*b^\star$. □

B.3 Proof of Theorem 13

Theorem 13. *Let Assumptions 4 and 5 hold and $(x^\star, y^\star) \in \mathcal{S}^\star$ be a saddle-point of the exact problem. Let (x_k, y_k) be generated by (12) under Assumptions 7 and 8 with inexact data b^δ such that $\left\|b^\delta - b^\star\right\| \leq \delta$ and error $|\varepsilon_k| \leq C_0\delta$ in the proximal operator for all $k \in \mathbb{N}$. Denote by $(\hat{x}_k, \hat{y}_k)$ the averaged iterates $(\frac{1}{k}\sum_{j=1}^k x_j, \frac{1}{k}\sum_{j=1}^k y_j)$. Then there exist constants C_1, C_2, C_3 and C_4 such that, for every $k \in \mathbb{N}$,*

$$\mathcal{L}^\star(\hat{x}_k, y^\star) - \mathcal{L}^\star(x^\star, \hat{y}_k) \ \leq \ \frac{C_1}{k} + C_2\delta + C_3\delta^{3/2}k^{1/2} + C_4\delta^2 k \ . \tag{19}$$

Let also Assumption 9 hold. Then there exist constants C_5, C_6, C_7, C_8 and C_9 such that, for every $k \in \mathbb{N}$,

$$\|A\hat{x}_k - b^\star\|^2 \ \leq \frac{C_5}{k} + C_6\delta + C_7\delta^{3/2}k^{1/2} + C_8\delta^2 k + C_9\delta^2 \ . \tag{20}$$

Proof. Recall that we denote $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ a primal-dual pair, and define

$$V(z) := \frac{1}{2}\|x\|_T^2 + \frac{1}{2}\|y\|_\Sigma^2 \ . \tag{61}$$

38

Use Lemma 31 at $x = x^\star$ and $y = y^\star$, to get

$$\begin{aligned}&\tfrac{1-\tau_M\sigma_M\|A\|^2}{2\tau_M}\|x_k - x^\star\|^2 + \tfrac{1}{2}\|y_k - y^\star\|_\Sigma^2 + \sum_{j=1}^{k}[\mathcal{L}^\delta(x_j, y^\star) - \mathcal{L}^\delta(x^\star, y_j)] + \tfrac{\omega}{2\tau_M}\sum_{j=1}^{k}\|x_j - x_{j-1}\|^2 \\ &\leq V(z_0 - z^\star) + \sum_{j=1}^{k}\varepsilon_j \ .\end{aligned} \tag{62}$$

Notice that

$$\mathcal{L}^\delta(x_j, y^\star) - \mathcal{L}^\delta(x^\star, y_j) = \mathcal{L}^\star(x_j, y^\star) - \mathcal{L}^\star(x^\star, y_j) + \langle y_j - y^\star, b^\delta - b^\star\rangle \ . \tag{63}$$

Then,

$$\begin{aligned}&\tfrac{1-\tau_M\sigma_M\|A\|^2}{2\tau_M}\|x_k - x^\star\|^2 + \frac{1}{2}\|y_k - y^\star\|_\Sigma^2 + \sum_{j=1}^{k}[\mathcal{L}^\star(x_j, y^\star) - \mathcal{L}^\star(x^\star, y_j)] + \tfrac{\omega}{2\tau_M}\sum_{j=1}^{k}\|x_j - x_{j-1}\|^2 \\ &\leq V(z_0 - z^\star) + \sum_{j=1}^{k}\varepsilon_j + \delta\sum_{j=1}^{k}\|y_j - y^\star\| \, .\end{aligned} \tag{64}$$

Recall that $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y) \geq 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Moreover, $\omega \geq 0$ by Assumption 8 and so $1 - \tau_M\sigma_M\|A\|^2 \geq 0$. Then, for every $j \in \mathbb{N}$, we have that

$$\|y_j - y^\star\|_\Sigma^2 \ \leq \ 2V(z_0 - z^\star) + 2\sum_{i=1}^{j}\varepsilon_i + 2\delta\sum_{i=1}^{j}\|y_i - y^\star\| \tag{65}$$

and so

$$\|y_j - y^\star\|^2 \ \leq \ 2\sigma_M\left[V(z_0 - z^\star) + \sum_{i=1}^{j}\varepsilon_i\right] + 2\delta\sigma_M\sum_{i=1}^{j}\|y_i - y^\star\| \ . \tag{66}$$

Apply Lemma 27 to Equation (66) with $u_j = \|y_j - y^\star\|$, $S_j = 2\sigma_M\left[V(z_0 - z^\star) + \sum_{i=1}^{j}\varepsilon_i\right]$ and $\lambda = 2\delta\sigma_M$. We get, for $1 \leq j \leq k$,

$$\begin{aligned}\|y_j - y^\star\| &\leq \delta\sigma_M j + \sqrt{2\sigma_M\left[V(z_0 - z^\star) + \sum_{i=1}^{j}\varepsilon_i\right] + (\delta\sigma_M j)^2} \\ &\leq 2\delta\sigma_M k + \sqrt{2\sigma_M\left[V(z_0 - z^\star) + \sum_{i=1}^{k}\varepsilon_i\right]} \ .\end{aligned} \tag{67}$$

Insert the latter in Equation (64), to obtain

$$\begin{aligned}
&\sum_{j=1}^{k}\left[\mathcal{L}^{\star}(x_j,y^{\star})-\mathcal{L}^{\star}(x^{\star},y_j)\right]\\
&\qquad\leq V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+\delta\sum_{j=1}^{k}\left(2\delta\sigma_M k+\sqrt{2\sigma_M\left[V(z_0-z^{\star})+\sum_{i=1}^{k}\varepsilon_i\right]}\right)\\
&\qquad= V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+\delta k\sqrt{2\sigma_M\left[V(z_0-z^{\star})+\sum_{i=1}^{k}\varepsilon_i\right]}+2\delta^2\sigma_M k^2\\
&\qquad\leq V(z_0-z^{\star})+C_0k\delta+\delta k\left(\sqrt{2\sigma_M V(z_0-z^{\star})}+\sqrt{2\sigma_M C_0 k\delta}\right)+2\delta^2\sigma_M k^2\ ,
\end{aligned}$$

where the last line uses $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$. By Jensen's inequality, we get the first claim. For the second result, apply Lemma 32 at $x=x^{\star}$ and $y=y^{\star}$:

$$\begin{aligned}
&V(z_k-z^{\star})+\frac{\theta}{2\tau_M\xi}\sum_{j=1}^{k}\|x_j-x_{j-1}\|^2+\frac{\rho}{2\eta}\sum_{j=1}^{k}\|Ax_j-Ax^{\star}\|^2+\sum_{j=1}^{k}\left[\mathcal{L}^{\delta}(x_j,y^{\star})-\mathcal{L}^{\delta}(x^{\star},y_j)\right]\\
&\leq V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+\frac{\sigma_m(\eta-1)k}{2}\left\|Ax^{\star}-b^{\delta}\right\|^2\ .
\end{aligned}\tag{68}$$

Using Equations (63) and (67), we have

$$\begin{aligned}
&V(z_k-z^{\star})+\frac{\theta}{2\tau_M\xi}\sum_{j=1}^{k}\|x_j-x_{j-1}\|^2+\frac{\rho}{2\eta}\sum_{j=1}^{k}\|Ax_j-b^{\star}\|^2+\sum_{j=1}^{k}\left[\mathcal{L}^{\star}(x_j,y^{\star})-\mathcal{L}^{\star}(x^{\star},y_j)\right]\\
&\leq V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+\sum_{j=1}^{k}\langle y_j-y^{\star},b^{\star}-b^{\delta}\rangle+\frac{\sigma_m(\eta-1)k}{2}\left\|b^{\star}-b^{\delta}\right\|^2\\
&\leq V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+\delta\sum_{j=1}^{k}\|y_j-y^{\star}\|+\frac{\sigma_m(\eta-1)k}{2}\delta^2\\
&\leq V(z_0-z^{\star})+\sum_{j=1}^{k}\varepsilon_j+2\sigma_M\delta^2k^2+\delta k\sqrt{2\sigma_M\left[V(z_0-z^{\star})+\sum_{i=1}^{k}\varepsilon_i\right]}+\frac{\sigma_m(\eta-1)k}{2}\delta^2\ .
\end{aligned}\tag{69}$$

Recall that $\theta \geq 0$ and that $\mathcal{L}^\star(x, y^\star) - \mathcal{L}^\star(x^\star, y) \geq 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. By Jensen's inequality, rearranging the terms and using $\sum_{i=1}^{k} \varepsilon_i \leq C_0 k\delta$, we get the claim. The exact values of the constants of Theorem 13 are therefore:

$$\begin{aligned}
C_1 &= V(z_0 - z^\star) \ , \\
C_2 &= C_0 + \sqrt{2\sigma_M V(z_0 - z^\star)} \ , \\
C_3 &= \sqrt{2\sigma_M C_0} \ , \\
C_4 &= 2\sigma_M \ , \\
C_5 &= \frac{2\eta}{\rho} C_1 \ , \\
C_6 &= \frac{2\eta}{\rho} C_2 \ , \\
C_7 &= \frac{2\eta}{\rho} C_3 \ , \\
C_8 &= \frac{2\eta}{\rho} C_4 \ , \\
C_9 &= \frac{\eta\sigma_m(\eta - 1)}{\rho} \ ,
\end{aligned} \tag{70}$$

□

B.4 Example of divergence in absence of noisy solution (see Remark 15)

We present an example in which the primal exact problem has solution, but the noisy one does not and the averaged primal iterates generated by Algorithm (12) indeed diverge. First note that, if the function R has bounded domain, the primal iterates remain bounded. So, to exhibit a case of divergence of the primal iterates, we consider a function R with full domain: set $R(\cdot) = \frac{1}{2}\|\cdot\|^2$ (and $F = 0$). The exact problem is then

$$\min_{x \in \mathcal{X}} \ \frac{1}{2}\|x\|^2 \quad \text{s.t.} \quad Ax = b^\star \ . \tag{71}$$

Now consider a noisy datum b^δ such that $Ax = b^\delta$ does not have a solution. If the associated normal equation, namely $A^*Ax = A^*b^\delta$ is feasible, in Section 7 we prove not only boundedness of the iterates but also convergence to a normal solution. On the contrary, to get divergence of the iterates, here we consider a classic scenario in which the perturbation of the exact data generates an unfeasible constraint, even for the associated normal equation. We recall that

this may happen only in the infinite dimensional setting, as when $R(A)$ is finite dimensional, it is also closed and a solution to the normal equation always exists. As a prototype of ill-posed problem, let $\mathcal{X} = \mathcal{Y} = \ell^2$ and A be defined by, for every $x \in \ell^2$ and for every $i \in \mathbb{N}$,

$$(Ax)^i = a^i x^i \ ,$$

where, for every $i \in \mathbb{N}$, $a^i \in (0, M)$ for a fixed constant $M > 0$ and $\inf_{i\in\mathbb{N}} a_i = 0$. Note that $A\colon \ell^2 \to \ell^2$ is well-defined, linear, continuous, self-adjoint and compact. Let $b^\star$ in the range of A and denote by $x^\star$ the unique solution to $\mathcal{P}^\star$ defined in Equation (71); namely, $(x^\star)^i := (b^\star)^i/a^i$ for every $i \in \mathbb{N}$. In particular, the $(b^\star)^i$ are such that $x^\star$ belongs to ℓ^2. Let also $b^\delta \in \ell^2$ with $\|b^\delta - b^\star\| \leq \delta$, but such that the noisy equation does not have a normal solution. Defining, for every $i \in \mathbb{N}$,

$$(x^\delta)^i := (b^\delta)^i/a^i, \tag{72}$$

the previous means that x^δ does not belong to ℓ^2. For an explicit example, consider $a^i = 1/i, (b^\star)^i = 1/i^2$ and $(b^\delta)^i = (b^\star)^i + C/i$, with $C = \delta/\sqrt{\sum_{j=1}^{+\infty} 1/j^2}$.

Apply the algorithm with step-sizes $\sigma > 0$ and $\tau > 0$ such that $\sigma\tau < 1/\|A\|^2$ and notice that it implies, for every $i \in \mathbb{N}$, $\sigma\tau < 1/(a^i)^2$. As $a^i > 0$ for every $i \in \mathbb{N}$, the coordinates of the averaged sequence $(\hat{x}_k^i)$ are convergent to a solution of the following (one-dimensional) optimization problem:

$$\mathcal{P}^i := \operatorname*{argmin}_{x^i\in\mathbb{R}} \left\{ \frac{1}{2}(x^i)^2 : a^i x^i = (b^\delta)^i \right\} = \left\{ \frac{(b^\delta)^i}{a^i} \right\}.$$

Hence, for the primal-dual algorithm, if $x^\delta \notin \ell^2$, then $(\hat{x}_k)$ diverges. Indeed, by contradiction, suppose that $(\hat{x}_k)$ is bounded. As it is bounded and converges coordinate-wise to x^δ, then it weakly converges to x^δ. But this is not possible since x^δ is not in ℓ^2.
Note that the problem considered in this example can be treated by Landweber method and it is well-known that also the iterates generated by this method, while being different from the ones of primal-dual algorithm, diverge.

C Sparse recovery

C.1 Proof of Proposition 16

Proposition 16. *Fix a primal-dual solution $(x^\star, y^\star) \in \mathcal{S}^\star$. Let the extended support be $\Gamma := \{i \in \mathbb{N} : |(A^*y^\star)_i| = 1\}$ and the saturation gap be $m := \sup\{|(A^*y^\star)_i| : |(A^*y^\star)_i| < 1\}$. Then Γ is finite, and $m < 1$. Moreover, for every $x \in \mathcal{X}$, with $\Gamma_C := \mathbb{N} \setminus \Gamma$,*

$$D^{-A^*y^\star}(x, x^\star) \geq (1 - m) \sum_{i\in\Gamma_C} |x_i|. \tag{24}$$

Proof. Recall that $x^\star, y^\star$ is a primal-dual solution, hence $-A^*y^\star \in \partial \|x^\star\|_1$. For every $i \in \mathbb{N}$ we have that $[\partial \|\cdot\|_1]_i (x^\star) \subseteq [-1, 1]$ and so $|(A^*y^\star)_i| \leq 1$. Recall that $\Gamma_C := \mathbb{N} \setminus \Gamma$. As $A^*y^\star$ belongs to $\mathcal{X} = \ell^2(\mathbb{N}; \mathbb{R})$, we have

$$\sum_{i \in \mathbb{N}} |(A^*y^\star)_i|^2 < +\infty. \tag{73}$$

Indeed, $m \leq 1$ by definition and from Equation (73) the coefficients $|(A^*y^\star)_i|$ converge to 0 (and so they can not accumulate at 1). We have also that

$$\begin{aligned}
D^{-A^*y^\star}(x, x^\star) &= \sum_{i \in \mathbb{N}} [|x_i| - |x_i^\star| + (A^*y^\star)_i (x_i - x_i^\star)] \\
&= \sum_{i \in \mathbb{N}} [|x_i| + (A^*y^\star)_i x_i] \\
&\geq \sum_{i \in \Gamma} \left[|x_i| - \underbrace{|(A^*y^\star)_i|}_{=1} |x_i| \right] + \sum_{i \in \Gamma_C} \left[|x_i| - \underbrace{|(A^*y^\star)_i|}_{\leq m} |x_i| \right] \\
&\geq (1 - m) \sum_{i \in \Gamma_C} |x_i|.
\end{aligned}$$

□

C.2 Tikhonov regularization: Lasso

For Tikhonov regularisation, the results in terms of Bregman divergence and feasibility are the following.

Lemma 33 (Grasmair et al. (2011), Lemma 3.5)**.** *Let $Ax^\star = b^\star$, $-A^*y^\star \in \partial \|\cdot\|_1 (x^\star)$ and, for $\alpha > 0$,*

$$x_\alpha \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \left\| Ax - b^\delta \right\|^2 + \alpha \|x\|_1 \right\}. \tag{74}$$

Then it holds that

$$\|Ax_\alpha - b^\star\| \leq \delta + \alpha \|y^\star\| \qquad \text{and} \qquad D^{-A^*y^\star}(x_\alpha, x^\star) \leq \frac{(\delta + \alpha \|y^\star\| / 2)^2}{\alpha}.$$

The previous bounds, combined with Assumption 17 and the last inequality in Lemma 18, lead naturally to the following corollary.

Corollary 34 (Grasmair et al. (2011), Theorem 5.6)**.** *Suppose Assumption 17 holds. Then, for x_α defined as in Lemma 33 and $C := \alpha/\delta$,*

$$\|Ax_\alpha - b^\star\| \leq (1 + CW_s)\,\delta \qquad \text{and}$$
$$\|x_\alpha - x^\star\| \leq Q_s\,(1 + CW_s)\,\delta + \frac{1 + Q_s\,\|A\|}{1 - M_s}\,\frac{(1 + CW_s/2)^2}{C}\delta.$$

D Proofs of Section 7

D.1 Proof of Corollary 21

Corollary 21. *Let Assumption 4 hold. Let (x_k, y_k) be the sequence generated by Equation (12) with data b under Assumption 7, Assumption 8 and summable error ($(\varepsilon_k) \in \ell^1$). Denote by $(\hat{x}_k, \hat{y}_k)$ the averaged iterates. Then, every weak cluster point of $(\hat{x}_k, \hat{y}_k)$ belongs to $\mathcal{S}$. In particular, if $\mathcal{S} = \emptyset$, then the primal-dual sequence $(\hat{x}_k, \hat{y}_k)$ diverges: $\|(\hat{x}_k, \hat{y}_k)\| \to +\infty$.*

Proof. From Lemma 31, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and for any $k \in \mathbb{N}$, we have

$$\begin{aligned}&\frac{1 - \tau_M\sigma_M\,\|A\|^2}{2\tau_M}\,\|x_k - x\|^2 + \frac{1}{2\sigma}\,\|y_k - y\|_\Sigma^2 + \sum_{j=1}^{k}\left[\mathcal{L}(x_j, y) - \mathcal{L}(x, y_j)\right] + \frac{\omega}{2\tau_M}\sum_{j=1}^{k}\|x_j - x_{j-1}\|^2 \\ &\leq V(z_0 - z) + \sum_{j=1}^{k}\varepsilon_j,\end{aligned} \tag{75}$$

where $\omega := 1 - \tau_M(L + \sigma_M\,\|A\|^2) \geq 0$ by Assumption 8. Using Jensen's inequality, we get

$$\mathcal{L}(\hat{x}_k, y) - \mathcal{L}(x, \hat{y}_k) \leq \frac{1}{k}\left[V(z_0 - z) + \sum_{j=1}^{+\infty}\varepsilon_j\right]. \tag{76}$$

Let (x_∞, y_∞) be a weak cluster point of $(\hat{x}_k, \hat{y}_k)$; namely, there exists a subsequence $(\hat{x}_{k_j}, \hat{y}_{k_j}) \subseteq (\hat{x}_k, \hat{y}_k)$ such that $(\hat{x}_{k_j}, \hat{y}_{k_j}) \rightharpoonup (x_\infty, y_\infty)$. By weak lower-semicontinuity of R and F, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathcal{L}(x_\infty, y) - \mathcal{L}(x, y_\infty) \leq \liminf_j \mathcal{L}(\hat{x}_{k_j}, y) - \mathcal{L}(x, \hat{y}_{k_j}) \leq \liminf_j \frac{1}{k_j}\left[V(z_0 - z) + \sum_{j=1}^{+\infty}\varepsilon_j\right] = 0. \tag{77}$$

Thus (x_∞, y_∞) is a saddle-point for the Lagrangian.
Now suppose that the set of saddle-points of $\mathcal{L}$ is empty. Assume also, for contradiction, that $(\hat{x}_k, \hat{y}_k)$ does not diverge. Then we can extract a bounded subsequence, that consequently admits a weakly converging subsequence. But then, the limit is a saddle-point, which contradicts the assumption. □

D.2 Proof of Lemma 22

Lemma 22. *Let Assumption 4 hold. Assume that $\tilde{\mathcal{C}} \neq \emptyset$. Let (x_k) be the primal sequence generated by algorithm (35); namely, with $T = \tau\,\mathrm{Id}$, $\Sigma = \sigma\,\mathrm{Id}$, $\varepsilon_k = 0$ for every $k \in \mathbb{N}$ and $y_0 = y_{-1} + \sigma(Ax_0 - b)$. Then, there exists a primal sequence (u_k) generated by the same procedure but applied to problem $\tilde{\mathcal{P}}$ (as stated in (36)) such that $x_k = u_k$ for every $k \in \mathbb{N}$.*

Proof. As $\tilde{\mathcal{C}} \neq \emptyset$, there exists $x^b \in \mathcal{X}$ such that $A^*Ax^b = A^*b$. First consider the algorithm in (35). Note that, for every $k \in \mathbb{N}$, $\tilde{y}_k = y_k + \sigma\,(Ax_k - b)$ and multiply the last step by A^*. We get, for every $k \in \mathbb{N}$,

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\tau R}(x_k - \tau\nabla F(x_k) - \tau A^*y_k - \sigma\tau A^*A(x_k - x^b)) \\ A^*y_{k+1} &= A^*y_k + \sigma A^*A(x_{k+1} - x^b). \end{aligned}$$

Recall that $S := (A^*A)^{\frac{1}{2}}$ and introduce $p_k := A^*y_k$. Then the primal sequence (x_k) is equivalently defined by the following recursion: given x_0 and $p_0 = A^*y_0$, for every $k \in \mathbb{N}$,

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\tau R}(x_k - \tau\nabla F(x_k) - \tau p_k - \sigma\tau S^2(x_k - x^b)) \\ p_{k+1} &= p_k + \sigma S^2(x_{k+1} - x^b). \end{aligned} \tag{78}$$

As A^*y_{-1} belongs to $R(A^*)$ and $R(A^*) = R(S)$ (Engl et al., 1996, Prop 2.18), there exists v_{-1} such that $Sv_{-1} = A^*y_{-1}$. Now consider the primal-dual algorithm applied to problem (36) starting at $u_0 = x_0$, v_{-1} and $v_0 = v_{-1} + \sigma(Su_0 - Sx^b)$. It reads as: for every $k \in \mathbb{N}$,

$$\begin{aligned} \tilde{v}_k &= 2v_k - v_{k-1} \\ u_{k+1} &= \mathrm{prox}_{\tau R}(u_k - \tau\nabla F(u_k) - \tau S\tilde{v}_k) \\ v_{k+1} &= v_k + \sigma(Su_{k+1} - Sx^b). \end{aligned}$$

Then, noticing that $\tilde{v}_k = v_k + \sigma\left(Su_k - Sx^b\right)$ and multiplying the last step by S,

$$\begin{aligned} u_{k+1} &= \mathrm{prox}_{\tau R}(u_k - \tau\nabla F(u_k) - \tau Sv_k - \sigma\tau S^2(u_k - x^b)) \\ Sv_{k+1} &= Sv_k + \sigma S^2(u_{k+1} - x^b)\ . \end{aligned}$$

Define the change of variable $q_k := Sv_k$, so that $q_{-1} = Sv_{-1} = A^*y_{-1}$ and

$$q_0 = Sv_0 = S\left(v_{-1} + \sigma(Su_0 - Sx^b)\right) = A^*(y_{-1} + \sigma(Ax_0 - b)) = A^*y_0 = p_0.$$

Then the primal sequence (u_k) is alternatively defined by the following recursion: for every $k \in \mathbb{N}$,

$$\begin{aligned} u_{k+1} &= \mathrm{prox}_{\tau R}(u_k - \tau\nabla F(u_k) - \tau q_k - \sigma\tau S^2(u_k - x^b)) \\ q_{k+1} &= q_k + \sigma S^2(u_{k+1} - x^b) \ . \end{aligned} \tag{79}$$

Comparing Equation (78) with Equation (79), with $(u_0, q_0) = (x_0, p_0)$, we get the claim. □

D.3 Proof of Theorem 23

Theorem 23. *Let Assumption 4 hold. Assume that $\tilde{\mathcal{P}}$ (as stated in 32) admits a saddle-point; namely, that there exists a pair $(\tilde{x}, \tilde{v}) \in \mathcal{X} \times \mathcal{X}$ such that*

$$\begin{cases} -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x}) \ , \\ A^*A\tilde{x} = A^*b \ . \end{cases} \tag{37}$$

Let (x_k, y_k) be the sequence generated by Equation (35), namely with initialization $y_0 = y_{-1} + \sigma(Ax_0 - b)$, and under Assumption 8. Denote by $(\hat{x}_k)$ the averaged primal iterates. Then there exists $\tilde{x}_\infty \in \tilde{\mathcal{P}}$ such that $\hat{x}_k \rightharpoonup \tilde{x}_\infty$. Moreover, if $\mathcal{P} = \emptyset$, then $\hat{y}_k$ diverges.

Proof. From Lemma 22, we know that the sequence $(\hat{x}_k)$ generated by Equation (35) coincides with the primal iterate of a sequence $(\hat{u}_k, \hat{v}_k)$ generated by the same algorithm on problem (36). Notice that $\|S\| = \|(A^*A)^{\frac{1}{2}}\| = \|A\|$ and so, if Assumption 8 holds, the analogue also holds for problem (36): namely, $1 - \tau(L + \sigma\|S\|^2) \geq 0$. The same is true for Assumption 5. Indeed, defining $\bar{v} = S\tilde{v}$, $-S\bar{v} = -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x})$. Moreover, we have seen already that $A^*Ax = A^*b$ if and only if $Sx = Sx^b$, where x^b is any vector in $\mathcal{X}$ such that $A^*Ax^b = A^*b$. Then, $S\tilde{x} = Sx^b$ and $(\tilde{x}, \bar{v})$ is a saddle-point for (36). So, by Proposition 10, we know that the averaged primal-dual sequence $(\hat{u}_k, \hat{v}_k)$ weakly converges to a saddle-point for (36). In particular, there exists $\tilde{x}_\infty \in \tilde{\mathcal{P}}$ such that $\hat{u}_k \rightharpoonup \tilde{x}_\infty$ and so the same holds for $(\hat{x}_k)$. For the second claim, by assumption we have that $\mathcal{P} = \emptyset$, which implies that $\mathcal{S} = \emptyset$. All the assumptions of Corollary 21 are verified, so $(\hat{x}_k, \hat{y}_k)$ diverges. As $(\hat{x}_k)$ is weakly convergent and so bounded, we conclude that $(\hat{y}_k)$ has to diverge. □

D.4 Proof of Theorem 24

Theorem 24. *Let Assumption 4 hold and suppose that there exists a pair* $(\tilde{x}, \tilde{v}) \in \mathcal{X} \times \mathcal{X}$ *such that*

$$\begin{cases} -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x}) \ , \\ A^*A\tilde{x} = A^*b^\star \end{cases} \tag{38}$$

(namely, a saddle-point for the normal exact problem $\tilde{\mathcal{P}}^\star$*). Let* $b^\delta \in \mathcal{Y}$ *be a noisy data such that* $\left\|b^\delta - b^\star\right\| \leq \delta$ *for some* $\delta \geq 0$*. Moreover, suppose that* $\tilde{\mathcal{C}}^\delta \neq \emptyset$*; namely, that there exists* $x^\delta \in \mathcal{X}$ *such that* $A^*Ax^\delta = A^*b^\delta$*. Let Assumption 8 and Assumption 9 hold and* (x_k, y_k) *be the sequence generated by the algorithm Equation (35) on the noisy data* b^δ*; namely, for the initialization* $y_0 = y_{-1} + \sigma(Ax_0 - b^\delta)$*,*

$$\begin{cases} \tilde{y}_k = 2y_k - y_{k-1} \ , \\ x_{k+1} = \operatorname{prox}_{\tau R}(x_k - \tau\nabla F(x_k) - \tau A^*\tilde{y}_k) \ , \\ y_{k+1} = y_k + \sigma\left(Ax_{k+1} - b^\delta\right) . \end{cases}$$

Denote by $(\hat{x}_k)$ *the averaged primal iterates. Then,*

$$D^{-A^*A\tilde{v}}(\hat{x}_k, \tilde{x}) \leq \frac{C_1}{k} + C_2\delta + C_4\delta^2 k$$

and

$$\|A^*A\hat{x}_k - A^*b^\star\|^2 \leq \|S\|\left[\frac{C_5}{k} + C_6\delta + C_8\delta^2 k + C_9\delta^2\right],$$

where the constants involved in the bounds are specified in the proof.

Proof. From the assumption $\tilde{\mathcal{C}}^\delta \neq \emptyset$ and Lemma 22, we know that the sequence $(\hat{x}_k)$ coincides with the primal iterate of a sequence $(\hat{u}_k, \hat{v}_k)$ generated by the same algorithm on problem

$$\tilde{\mathcal{P}}^\delta = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{R(x) + F(x) : \quad Sx = Sx^\delta\right\}, \tag{80}$$

where x^δ is any vector in $\mathcal{X}$ such that $A^*Ax^\delta = A^*b^\delta$. As in the proof of the previous theorem, notice that $\|S\| = \|A\|$ and so, as Assumption 8 and Assumption 9 hold by hypothesis, the analogue also holds for problem (80): namely, $1 - \tau(L + \sigma\|S\|^2) \geq 0$, $\xi - \tau(\xi L + \sigma\|S\|^2) \geq 0$ and $\sigma(\eta - 1) - \sigma\xi\eta > 0$. The same is true for Assumption 5. Indeed, define $\bar{v} = S\tilde{v}$. Then, from Equation (38), $-S\bar{v} = -A^*A\tilde{v} \in \partial R(\tilde{x}) + \nabla F(\tilde{x})$ and $(\tilde{x}, \bar{v})$ is a saddle-point for

$$\tilde{\mathcal{P}}^\star = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{R(x) + F(x) : \quad Sx = S\tilde{x}\right\}. \tag{81}$$

In particular, we can apply Theorem 13 for $(\hat{u}_k, \hat{v}_k)$ - averaged primal-dual sequence generated on the noisy problem in (80) - with respect to $(\tilde{x}, \bar{v})$ - saddle-point for the exact problem in (81) - to get that

$$D^{-S\bar{v}}(\hat{u}_k, \tilde{x}) \leq \frac{C_1}{k} + C_2\tilde{\delta} + C_4(\tilde{\delta})^2 k$$

and

$$\|S\hat{u}_k - S\tilde{x}\|^2 \leq \frac{C_5}{k} + C_6\tilde{\delta} + C_8(\tilde{\delta})^2 k + C_9(\tilde{\delta})^2.$$

The constants in the previous bounds are the same as in (70) with $z_0 = (u_0, v_0)$, $z^\star = (\tilde{x}, \bar{v})$, $C_3 = C_7 = 0$ (because $C_0 = 0$ as we suppose $\varepsilon_k = 0$ for every $k \in \mathbb{N}$), $\sigma_m = \sigma_M = \sigma$ and

$$\tilde{\delta} := \|Sx^\delta - S\tilde{x}\|.$$

From Lemma 22, we recall also that $u_0 = x_0$ and $v_0 = v_{-1} + \sigma(Su_0 - Sx^\delta)$, where v_{-1} is any element in $\mathcal{X}$ such that $Sv_{-1} = A^*y_{-1}$ (v_{-1} exists due to $R(A^*) = R(S)$). Now it remains to show that $\tilde{\delta} \leq \delta$. Denote by $(\mu_i, f_i, g_i)_{i\in\mathbb{N}} \subseteq \mathbb{R}_+ \times \mathcal{X} \times \mathcal{Y}$ the singular value decomposition of the operator A. First, notice that $S^2(x^\delta - \tilde{x}) = A^*(b^\delta - b^\star)$ and so that, for every $i \in \mathbb{N}$,

$$\mu_i^2\langle x^\delta - \tilde{x}, f_i\rangle = \mu_i\langle b^\delta - b^\star, g_i\rangle.$$

Then, for every $i \in \mathbb{N}$ such that $\mu_i \neq 0$, $\mu_i\langle x^\delta - \tilde{x}, f_i\rangle = \langle b^\delta - b^\star, g_i\rangle$ and so

$$\begin{aligned}\tilde{\delta}^2 = \|Sx^\delta - S\tilde{x}\|^2 &= \sum_{i\in\mathbb{N}} \left(\mu_i\langle x^\delta - \tilde{x}, f_i\rangle\right)^2 = \sum_{\mu_i\neq 0} \left(\mu_i\langle x^\delta - \tilde{x}, f_i\rangle\right)^2 \\ &= \sum_{\mu_i\neq 0} \left(\langle b^\delta - b^\star, g_i\rangle\right)^2 \leq \sum_{i\in\mathbb{N}} \left(\langle b^\delta - b^\star, g_i\rangle\right)^2 = \|b^\delta - b^\star\|^2 \leq \delta^2.\end{aligned}$$

We conclude the claim simply by noticing that

$$D^{-A^*A\tilde{v}}(\hat{x}_k, \tilde{x}) = D^{-S\bar{v}}(\hat{u}_k, \tilde{x})$$

and

$$\|A^*A\hat{x}_k - A^*b^\star\| = \left\|S^2\hat{u}_k - S^2\tilde{x}\right\| \leq \|S\| \, \|S\hat{u}_k - S\tilde{x}\| \, .$$

□

E A dual view on the implicit bias of gradient descent on least squares

Here we provide an interesting view on why the "implicit" bias of gradient descent on least squares is not so implicit. Recall that these iterations,

$$x_{k+1} = x_k - \gamma A^*(Ax_k - b) \ , \tag{82}$$

converge, for $\gamma < 2/\|A\|_{\text{op}}^2$, to the minimal Euclidean norm solution of $Ax = b$:

$$\min_{x \in \mathcal{X}} \frac{1}{2} \|x\|^2 \quad \text{s.t.} \quad Ax = b \ , \tag{83}$$

provided that Problem (83) is feasible and $x_0 = 0$.

It turns out that the iterations (82) correspond, up to multiplication by $-A^*$, to the iterates of gradient descent to the dual of (83), namely:

$$\min_{y \in \mathcal{Y}} \frac{1}{2} \|A^* y\|^2 + \langle b, y \rangle \ , \quad \text{and} \quad y_{k+1} = y_k - \gamma(AA^* y_k + b) \ . \tag{84}$$

By setting $x_{k+1} = -A^* y_{k+1}$ one recovers the iterates of gradient descent on least squares (82). Therefore the "implicit bias" of gradient descent on least squares is not so implicit: its iterates x_k are dual to iterates y_k on Problem (84), which is itself the dual of Problem (83) in which the bias appears explicitly.

References

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statist. Sci.*, 27(4):450–468, 2012.

M. Bachmayr and M. Burger. Iterative total variation schemes for nonlinear inverse problems. *Inverse Problems*, 25(10):105004, 2009.

M. Bahraoui and B. Lemaire. Convergence of diagonally stationary sequences in convex optimization. *Set-Valued Anal.*, 2:49–61, 1994.

M. Barré, A. Taylor, and F. Bach. Principled analyses and design of first-order methods with inexact proximal operators. *arXiv preprint arXiv:2006.06041*, 2020.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

M. Benning and M Burger. Modern regularization methods for inverse problems. *Acta Numer.*, 27, 2018a.

M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numer.*, 27:1–111, 2018b.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends in Mach. Learn.*, 3(1):1–122, 2011.

K. Bredies and M. Zhariy. A discrepancy-based parameter adaptation and stopping rule for minimization algorithms aiming at Tikhonov-type regularization. *Inverse problems*, 29(2): 025008, 2013.

P. Brianzi, F. Di Benedetto, and C. Estatico. Preconditioned iterative regularization in banach spaces. *Comput. Optim. Appl.*, 54(2):263–282, 2013.

M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse problems*, 20(5):1411, 2004.

M. Burger, S. Osher, J. Xu, and G Gilboa. Nonlinear inverse scale space methods for image restoration. In *International Workshop on Variational, Geometric, and Level Set Methods in Computer Vision*, pages 25–36. Springer, 2005.

M. Burger, G. Gilboa, S. Osher, and J. Xu. Nonlinear inverse scale space methods. *Commun. Math. Sci.*, 4(1):179–212, 2006.

M. Burger, E. Resmerita, and L. He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2-3):109–135, 2007.

M. Burger, M. Möller, M. Benning, and S. Osher. An adaptive inverse scale space method for compressed sensing. *Math. Comp.*, 82(281):269–299, 2013.

J.-F. Cai, S. Osher, and Z. Shen. Convergence of the linearized Bregman iteration for ℓ_1-norm minimization. *Math. Comp.*, 78(268):2127–2136, 2009a.

J-F. Cai, S. Osher, and Z. Shen. Linearized bregman iterations for compressed sensing. *Math. Comp.*, 78(267):1515–1536, 2009b.

L. Calatroni, G. Garrigos, L. Rosasco, and S. Villa. Accelerated iterative regularization via dual diagonal descent. *SIAM J. Optim.*, 31(1):754–784, 2021.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2): 489–509, 2006.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J.Optim. Theory Appl.*, 158(2):460–479, 2013.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11): 1413–1457, 2004.

C. Deledalle, S. Vaiter, G. Peyré, and J. Fadili. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection . *SIAM J. Imaging Sci.*, 7(4):2448–2487, 2014.

D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

H. W. Engl, W. Heinz, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

M. Figueiredo and R. Nowak. Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects. In *AISTATS*, pages 930–938. PMLR, 2016.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, New York, 2013.

M. P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM J. Optim.*, 18(4):1326–1350, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

G. Garrigos, L. Rosasco, and S. Villa. Iterative regularization via dual diagonal descent. *J.Math. Imaging Vision*, 60(2):189–215, 2018.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *Ann. Statist.*, 49(2):1029–1054, 2021.

A. Goldenshluger and S. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli*, 9(5):783 – 807, 2003.

M. Grasmair, O. Scherzer, and M. Haltmeier. Necessary and sufficient conditions for linear convergence of l1-regularization. *Communications on Pure and Applied Mathematics*, 64 (2):161–182, 2011.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *NeurIPS*, pages 6151–6159, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. *Proceedings of the 35th International Conference on Machine Learning*, 80:1832–1841, 2018.

T. J. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

Bo Huang, Shiqian Ma, and Donald Goldfarb. Accelerated linearized bregman method. *Journal of Scientific Computing*, 54(2-3):428–453, 2013.

F. Iutzeler and J. Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued Var. Anal.*, 28(4):661–678, 2020.

B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative regularization methods for nonlinear ill-posed problems*, volume 6. Walter de Gruyter, 2008.

Alessandro Lanza, Serena Morigi, Ivan W Selesnick, and Fiorella Sgallari. Sparsity-inducing nonconvex nonseparable regularization for convex image processing. *SIAM Journal on Imaging Sciences*, 12(2):1099–1134, 2019.

D. Lorenz, S. Wenger, F. Schöpfer, and M. Magnor. A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing. In *2014 IEEE international conference on image processing (ICIP)*, pages 1347–1351. IEEE, 2014a.

D. A Lorenz, F. Schopfer, and S. Wenger. The linearized bregman method via split feasibility problems: Analysis and generalizations. *SIAM J. Imaging Sci.*, 7(2):1237–1262, 2014b.

M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. Dual extrapolation for sparse generalized linear models. *J. Mach. Learn. Res.*, 2020.

Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupre la Tour, Ghislain Durif, Cassio F Dantas, et al. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. *NeuRIPS*, 35:25404–25421, 2022.

S. Mosci, L. Rosasco, M. Santoro, A Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 418–433. Springer, 2010.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NeurIPS*, pages 451–459, 2011.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983.

Y. E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ2). In *Dokl. Akad. Nauk. Sssr.*, volume 269, pages 543–547, 1983.

A. Neubauer. On nesterov acceleration for landweber iteration of linear ill-posed problems. *J. Inverse Ill-posed Probl.*, 25(3):381–390, 2017.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.*, 20(2):231–252, 2010.

S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *SIAM Multiscale Model. Simul.*, 4:460–489, 2005.

S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin. Sparse recovery via differential inclusions. *Appl. Comput. Harmon. Anal.*, 41(2):436–469, 2016.

N. Pagliana and L. Rosasco. Implicit regularization of accelerated methods in Hilbert spaces. In *NeurIPRS*, pages 14454–14464, 2019.

T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 International Conference on Computer Vision*, pages 1762–1769, 2011.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Comput. Math. Math. Phys.*, 4(5):1–17, 1964.

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(1):335–366, 2014.

L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *NeurIPS*, pages 1630–1638, 2015.

L. Rosasco, M. Santoro, S. Mosci, A. Verri, and S. Villa. A regularization approach to nonlinear variable selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 653–660, 2010.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *J. Convex Anal.*, 19(4):1167–1192, 2012.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NeurIPS*, pages 1458–1466, 2011.

F. Schopfer. Exact regularization of polyhedral norms. *SIAM J. Optim.*, 22(4):1206–1223, 2012.

F. Schöpfer and D. Lorenz. Linear convergence of the randomized sparse kaczmarz method. *Math. Program.*, 173(1):509–536, 2019.

F. Schöpfer, A. Louis, and T. Schuster. Nonlinear iterative methods for linear ill-posed problems in Banach spaces. *Inverse problems*, 22(1):311, 2006.

N. Simon, J. Friedman, T. J. Hastie, and R. Tibshirani. A sparse-group lasso. *J. Comput. Graph. Statist.*, 22(2):231–245, 2013. ISSN 1061-8600.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN 9780387772424.

M. Teboulle and A. Beck. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Letters*, 31:167–175, 2003.

A. N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

Samuel Vaiter, Gabriel Peyré, and Jalal Fadili. Low complexity regularization of linear inverse problems. *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 103–153, 2015.

T. Vaškevičius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. In *NeurIPS*, pages 2968–2979, 2019.

T. Vaškevičius, V. Kanade, and P. Rebeschini. The statistical complexity of early stopped mirror descent. *NeurIPS*, pages 253–264, 2020.

S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.

S. Villa, S. Matet, B.C. Vu, and L. Rosasco. Implicit regularization with strongly convex bias: stability and acceleration. *Anal. Appl.*, 2022. doi: 10.1142/S0219530522400139.

B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.*, 38(3):667–681, 2013.

Y. Wei, F. Yang, and M. J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a081cab429ff7a3b96e0a07319f1049e-Paper.pdf.

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. 2008.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM J. Imaging Sci.*, 3(4):856–877, 2010.

W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l1- minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168, 2008.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. 2010.

X. Zhang, M. Burger, X. Bresson, and S. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM J. Imaging Sci.*, 3:253–276, 2010.

X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *J. Sci. Comput.*, 46:20–46, 2011.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497, 2016.

56