Analysis and Applications

Implicit regularization with strongly convex bias: stability and acceleration

Silvia Villa

MaLGa, DIMA, Università degli Studi di Genova,
Via Dodecaneso 35, 16146, Genova, Italy
silvia.villa@unige.it

Simon Matet

Squarepoint Capital, 16 Avenue Matignon
75008 Paris France
simonmatet@hotmail.com

Bằng Công Vũ

3NLab, Huawei Belgium Research Center (BeRC),
Gaston Geenslaan 10, 3001, Leuven, Belgium
bangcvvn@gmail.com

Lorenzo Rosasco [*]

MaLGa, DIBRIS, Università degli Studi di Genova,
Via Dodecaneso 35, 16146, Genova, Italy
lrosasco@unige.it

Received (Day Month Year)
Revised (Day Month Year)

Implicit regularization refers to the property of optimization algorithms to be biased towards a certain class of solutions. This property is relevant to understand the behavior of modern machine learning algorithms as well as to design efficient computational methods. While the case where the bias is given by a Euclidean norm is well understood, implicit regularization schemes for more general classes of biases are much less studied. In this work, we consider the case where the bias is given by a strongly convex functional, in the context of linear models, and data possibly corrupted by noise. In particular, we propose and analyze accelerated optimization methods and highlight a trade-off between convergence speed and stability. Theoretical findings are complemented by an empirical analysis on high-dimensional inverse problems in machine learning and signal processing, showing excellent results compared to the state of the art.

Keywords: Implicit/iterative regularization, machine learning, inverse problems, high-

[*] Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

dimensional and overparameterized problems, convex optimization, duality.

Mathematics Subject Classification 2010: 47H05, 49M29, 49M27, 90C25

1. Introduction

Many data driven problems require estimating a quantity of interest based on finitely many, and often noisy, measurements. When considering linear models this question can naturally be phrased as a linear inverse problem. The most classical approach in this setting is Tikhonov regularization, corresponding to the minimization of an empirical objective [?] where a data fit term is penalized using a regularizing functional, encoding a bias for the problem. From an optimization perspective, this approach can be seen as the relaxation of a problem where the regularizing functional is minimized under linear equality constraints defined by the data. The latter would be the natural formulation in absence of noise or perturbations. From a numerical point of view, minimization is often performed using first order methods because of their simplicity and small memory requirement [?]. In practice, a regularization parameter balancing data fit and regularization needs to be chosen, and the solution of multiple optimization problems is typically required. Algorithm design is then typically split in two steps studied separately: first the design of an objective and then its minimization.

In this paper, we investigate a different approach classically called iterative regularization [?] and more recently implicit regularization in machine learning, see e.g.[?,?,?,?,?]. We will use the term iterative and implicit regularization interchangeably. This approach naturally combines modeling and computational aspects, potentially improving efficiency. Further, it was shown to be crucial to understand the learning properties of modern machine learning approaches such as deep learning, see [?]. A basic idea to derive implicit regularization algorithms is to solve directly the minimization of the regularizing functional under the linear constraints for the noisy data, rather than a relaxation. This ensures the sequence of obtained solutions to be biased to the regularization of interest. In this approach, the robustness of the considered optimization procedure in the presence of noise has to be considered. Indeed, depending on the noise level, the iterations might need to be stopped before convergence (early stopping [?]) to obtain a robust solution. In iterative regularization, the number of iterations is the regularization parameter. In this sense, this approach has a built-in warm restart property, that allows to easily compute a whole set of solutions corresponding to different regularization levels, while running the iteration a single time.

Robustness of optimization in the presence of disturbances is a classic topic of study [?]. However, in the optimization literature, the disturbances are typically assumed to disappear as the number of iterations increases so to preserve convergence, see e.g. [?]. The situation where disturbances might persist is typically studied in inverse problems, where they are seen as noise in the data. In this context there is a large literature on implicit/iterative regularization and [?] provides an exhaustive overview. Of interest to our discussion are the results considering regularizations which are not Euclidean norms and in particular the case of possibly non-smooth regularization functionals. One relevant research direction is the one considering regularization in Banach spaces, see e.g. [?] and references therein. For general nonsmooth functionals a number of works have considered the so called Bregman iteration approach [?]. As we discuss in the following, this is very much related and often equivalent with classic and older ideas in optimization methods, such ad proximal methods [?] and mirror descent methods [?], see also [?].

In this paper, we consider non-smooth strongly convex regularization and analyze the regularization properties of a dual gradient descent approach. The latter can be shown to be equivalent to mirror descent and the linearized Bregman iteration, but has a more direct derivation, see [?]. Notably, the approach reduces to Landweber iteration when the bias is given by a Euclidean norm. Our approach allows to consider all biases for which a corresponding proximity operator can be computed in closed form [?], or possibly up-to a given precision, see e.g. [?]. Compared to previous results we focus on two main novel aspects. First, we analyze the interplay between convergence and stability for deterministic noise. In particular, we derive explicit stability estimates for the noisy iterates and convergence results for the noiseless iterates. This contrasts in particular with previous results where, in the presence of noise, error estimates are studied only for suitable stopping rules, e.g. given by the discrepancy principle [?]. The interplay between convergence and stability is not clear in this setting. Second, we consider accelerated approaches using ideas from [?]. In this context, explicit stability bounds are crucial since a a trade-off between convergence speed and stability arises. Faster convergence is at the expense of stability and the advantage of accelerated methods is that more aggressive stopping rules can be considered. These results generalize similar observations made for Euclidean norms [?] and complement recent results in optimization [?,?,?]. From a technical point of view, we largely draw from the optimization literature, in particular results considering non-smooth convex optimization [?,?] and related robustness results [?].
Our theoretical findings are complemented by empirical results on three different applications from machine learning and signal processing: variable selection, matrix completion, and image deblurring. The experiments confirm the theoretical results and show that the recovery properties of iterative regularization are comparable to penalization approaches with much lower computational costs.

The rest of the paper is organized as follows: in Section ?? we describe the setting and the main assumptions, in Section ?? we introduce the iterations we study, and in Section ?? we state the main results, discuss them, and provide the main elements of the proof. The complete proof of the results is given in Section ??. In Section ?? we present several experimental results on matrix completion, variable selection, and deblurring problems.

2. Problem setting

We consider a problem of the form

$$y = Xw, \tag{2.1}$$

for a given matrix[a] $X\colon \mathbb{R}^p \to \mathbb{R}^n$, an observation $y \in \mathbb{R}^n$, and a vector $w \in \mathbb{R}^p$. Such a formulation include for instance regression, feature selection, as well as many image/signal processing problems. In general, the solution of the above linear equation is not unique, and a selection principle is needed to choose an appropriate solution (e.g. in the high dimensional scenario, where $p > n$). In this paper, we assume that the solution of interest $w^\dagger$ minimizes a function $R\colon \mathbb{R}^p \to \left]-\infty, +\infty\right]$ encoding some bias known a priori on the problem at hand. We assume R to be proper, lower semicontinuous, strongly convex, and we let $w^\dagger$ to be the unique solution of the optimization problem

$$\underset{y=Xw}{\text{minimize}}\; R(w). \tag{2.2}$$

Note that existence of a solution to problem (**??**) is assumed. We will see later that a weaker formulation of the problem can also be analyzed. Further, in practice, one does not have access to y, but only to a noisy version $\widehat{y}$. In particular, we consider a worst case scenario, where the noise is deterministic, i.e. $\|y - \widehat{y}\| \leq \delta$, for some $\delta > 0$. The goal is then to find a stable solution observing only X and $\widehat{y}$.

The classical way to achieve this goal is to relax the equality constraints, and use a Tikhonov regularization scheme,

$$\min_{w \in \mathbb{R}^p} \|\widehat{y} - Xw\|^2 + \lambda R(w).$$

A data fidelity term is added to the function R, multiplied by a regularization parameter λ. Such an approach usually requires two steps: first, the solution of a regularized problem for several values of the regularizing parameter, and second the best regularized solution is selected among the computed ones (model selection).

In this paper we follow a different route, which avoids relaxation, and is based on iterative regularization. This latter idea dates back at least to [**?**], it is classical in inverse problems [**?**,**?**], but it is also a common trick in machine learning, where it is referred to as early stopping [**?**] or implicit regularization [**?**,**?**]. Within the setting of the paper, this idea translates into defining a sequence $(\widehat{w}_t)_{t\in\mathbb{N}}$ derived by applying an appropriate minimization algorithm to the noisy problem

$$\underset{\widehat{y}=Xw}{\text{minimize}}\; R(w). \tag{2.3}$$

This is somewhat odd from an optimization perspective, since the sequence converges to a minimizer of the noisy problem (**??**), which is not the solution we are

[a] For simplicity, the results are stated in finite dimensional euclidean spaces, but all the conclusions hold if $\mathbb{R}^p$ and $\mathbb{R}^n$ are replaced by Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$.

looking for. The key insight is that the optimization itself iteratively (and implicitly) enforces regularization. The stopping time is the regularization parameter, and model selection coincides with defining a suitable stopping criterion, hence the name early stopping.

We next provide some discussion illustrating why such a procedure is sensible. The goal is to see why depending on the noise level, we can select an element $\widehat{w}_{t_\delta}$, of the sequence $(\widehat{w}_t)_{t\in\mathbb{N}}$, that converges to $w^\dagger$ when the noise goes to zero. An intuition of why this is possible can be derived from the proof's strategy. To analyze the behavior of the sequence $(\widehat{w}_t)_{t\in\mathbb{N}}$ we define an auxiliary (regularizing) sequence $(w_t)_{t\in\mathbb{N}}$, that is the sequence obtained applying the same minimization algorithm devised for problem (**??**), to the ideal problem (**??**), which therefore converges to $w^\dagger$. The choice of the stopping time is derived from the following error decomposition

$$\|\widehat{w}_t - w^\dagger\| \leq \|\widehat{w}_t - w_t\| + \|w_t - w^\dagger\|.$$

The term $\|w_t - w^\dagger\|$ is an optimization error, but can be seen as the regularization or approximation error [**?**]. We will show that it vanishes for increasing t and in fact will prove non asymptotic bounds. The term $\|\widehat{w}_t - w_t\|$ measures stability to noise and we will see to increase with t and δ. Given data and knowledge of the noise level, our actual regularization procedure is specified by a suitable choice t_δ and this results in the explicit bound $\|\widehat{w}_{t_\delta} - w^\dagger\| \leq c\delta^{1/2}$. Note that the dependence on the noise level δ is the same as in Tikhonov regularization [**?**]. In the rest of the paper, we develop the above idea providing all the details.

3. Iterative regularization for general penalty

In this section we begin presenting the iterative regularization procedures we study based on dual gradient descent (DGD) and accelerated dual gradient descent (ADGD). The first one is a basic algorithm, while the second is its accelerated version, requiring some additional steps. First, recall that the regularizing function R in (**??**) is assumed to be strongly convex. This implies that there exists $\alpha \in]0,+\infty[$ and a proper, lower semicontinuous, and convex function $F\colon \mathbb{R}^p \to [0,+\infty]$ such that

$$R = F + \frac{\alpha}{2}\|\cdot\|^2. \tag{3.1}$$

Both DGD and ADGD belong to the class of first order methods, requiring only matrix and vector multiplications, and the computation of the proximity operator of $\alpha^{-1}F$, which is defined as

$$(\forall w \in \mathbb{R}^p) \qquad \operatorname{prox}_{\alpha^{-1}F}(w) = \operatorname{argmin}_{u\in\mathbb{R}^p} \Big\{F(u) + \frac{\alpha}{2}\|u-w\|^2\Big\}. \tag{3.2}$$

The computation of the proximity operator involves a minimization problem, which can be solved explicitly in many relevant cases, see e.g. [**?**]. In particular, it reduces to the well-known soft-thresholding operator when F is equal to the ℓ^1 norm, and to a projection, when F is the indicator function of a convex and closed set [?]. We will show in the next section that DGD reduces to a gradient descent on the dual of problem of (??). Its convergence properties for Problem (??), which is not the one we want to solve, have been studied in [?]. By considering a Nesterov acceleration [?] of gradient descent, we derive ADGD, that is the FISTA variant on the dual problem, which has been considered in [?,?]. The algorithms DGD and ADGD can also be seen from an inexact optimization perspective, see e.g. [?] and reference therein. In this view, they solve the dual of the original noise free problem in (??), in the presence of a nonvanishing error on the gradient.

3.1. *Derivation of the algorithms*

We start showing that the proposed procedures DGD and ADGD are indeed a gradient and an accelerated gradient descent algorithm applied to the dual problem of the noisy minimization problem

$$\min_{Xw=\widehat{y}} R(w), \quad \text{with } R = F + \frac{\alpha}{2}\|\cdot\|^2. \tag{3.3}$$

Let C be a convex and closed subset of $\mathbb{R}^n$. With δ_C we denote the indicator function of C, which takes value 0 on C and $+\infty$ otherwise. The optimization problem in (??) can be equivalently written as

$$\min_{w\in\mathbb{R}^p} R(w) + \delta_{\widehat{y}}(Xw). \tag{3.4}$$

The above optimization problem is given by the sum of two convex, proper, and lower semicontinuous functions, where one of the two is composed with a linear operator. This is the suitable form to apply Fenchel-Rockafellar duality (see Appendix ?? for the definition of Fenchel conjugate of a convex function). The dual of the problem in (??) is then (see Appendix ??)

$$\min_{v\in\mathbb{R}^n} R^*(-X^T v) + \langle \widehat{y}, v\rangle. \tag{3.5}$$

As recalled in Appendix ??, its conjugate is differentiable with Lipschitz continuous gradient and

$$\nabla R^*(v) = \operatorname{prox}_{\alpha^{-1}F}(\alpha^{-1}v).$$

We derive that one step of gradient descent applied to the problem in (??) can be written as

$$v_{t+1} = v_t + \gamma\big(X \operatorname{prox}_{\alpha^{-1}F}(-\alpha^{-1}X^T v_t) - \widehat{y}\big),$$

and this is the main iteration in DGD. The derivation of ADGD is analogous, simply the gradient descent method is replaced by FISTA acceleration [?], see also [?]. Here not one, but two previous iterates are used at every step, resulting in faster convergence speed to the objective.

Dual Gradient Descent (DGD)

```
Let $\widehat{v}_0 = 0 \in \mathbb{R}^p$ and $\gamma = \alpha\|X\|^{-2}$
For $t = 0, 1, \ldots$ iterate
    $\widehat{w}_t = \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big)$
    $\widehat{v}_{t+1} = \widehat{v}_t + \gamma(X\widehat{w}_t - \widehat{y})$
If $t > 0$
    $\widehat{u}_t = (1/t)\sum_{k=1}^{t}\widehat{w}_k$
```

Accelerated Dual Gradient Descent (ADGD)

```
Let $\widehat{v}_0 = \widehat{z}_{-1} = \widehat{z}_0 = 0 \in \mathbb{R}^p$, $\gamma = \alpha\|X\|^{-2}$, and $\theta_0 = 1$
For $t = 0, 1, \ldots$ iterate
    $\widehat{r}_t = \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big)$
    $\widehat{z}_t = \widehat{v}_t + \gamma(X\widehat{r}_t - \widehat{y})$
    $\theta_{t+1} = (1 + \sqrt{1 + 4\theta_t^2})/2$
    $\widehat{v}_{t+1} = \widehat{z}_t + \frac{\theta_t - 1}{\theta_{t+1}}(\widehat{z}_t - \widehat{z}_{t-1})$
    $\widehat{w}_t = \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{z}_t\big)$
```

3.2. *DGD and ADGD with inexact proximity operators*

In this section, we present and discuss the situation in which the proximity operator of F is not available and can be computed only up to a certain precision. There are various ways to appropriately define computational errors in the approximation of the proximity operator [**?**,**?**]. Here we adopt the more general one, which corresponds to an inexact solution of the minimization problem defining the proximal point (**??**). To this aim, given $w \in \mathbb{R}^p$, define

$$(\forall u \in \mathbb{R}^p)(\forall \sigma > 0) \quad \Phi_\sigma(u) = F(u) + \frac{1}{2\sigma}\|u - w\|^2.$$

Definition 3.1. Let $w \in \mathbb{R}^p$ and $\sigma > 0$. We say that $\bar{p} \in \mathbb{R}^p$ is an approximation of $\mathrm{prox}_{\sigma F}(w)$ with ε-precision and we write $\bar{p} \approx_\varepsilon \mathrm{prox}_{\sigma F}(w)$ if

$$\Phi_\sigma(\bar{p}) - \min \Phi_\sigma \leq \frac{\varepsilon^2}{2\sigma}.$$

Since Φ_σ is σ^{-1} strongly convex, if $\bar{p} \approx_\varepsilon \mathrm{prox}_{\sigma F}(w)$, then $\bar{p} \in \mathrm{dom}F$ and

$$\|\bar{p} - \mathrm{prox}_{\sigma F}(x)\| \leq \varepsilon. \tag{3.6}$$

This observation will be relevant for the subsequent analysis of the inexact version of the DGD and ADGD algorithms given below.

Inexact Dual Gradient Descent (IDGD)

```
Let $\widehat{v}_0 = 0 \in \mathbb{R}^p$ and $\gamma = \alpha\|X\|^{-2}$
For $t = 0, 1, \ldots$ iterate
    $\widehat{w}_t \approx_\delta \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big)$
    $\widehat{v}_{t+1} = \widehat{v}_t + \gamma(X\widehat{w}_t - \widehat{y})$
If $t > 0$
    $\widehat{u}_t = (1/t)\sum_{k=1}^{t}\widehat{w}_k$
```

Inexact Accelerated Dual Gradient Descent (

```
Let $\widehat{v}_0 = \widehat{z}_{-1} = \widehat{z}_0 = 0 \in \mathbb{R}^p$, $\gamma = \alpha\|X\|^{-2}$, and $\theta_0 =$
For $t = 0, 1, \ldots$ iterate
    $\widehat{r}_t \approx_\delta \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big)$
    $\widehat{z}_t = \widehat{v}_t + \gamma(X\widehat{r}_t - \widehat{y})$
    $\theta_{t+1} = (1 + \sqrt{1 + 4\theta_t^2})/2$
    $\widehat{v}_{t+1} = \widehat{z}_t + \frac{\theta_t - 1}{\theta_{t+1}}(\widehat{z}_t - \widehat{z}_{t-1})$
    $\widehat{w}_t = \mathrm{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{z}_t\big)$
```

3.3. *Connections with other approaches*

Before studying the regularizing properties of the proposed procedures, we add a few remarks discussing our approach in the context of related studies. First, we show that DGD is a generalization of the well-known Landweber iteration (see [**?**]).

Remark 3.1 (Connection to Landweber iteration). Consider Algorithm DGD in the special case $F = 0$. Noting that, for every $w \in \mathbb{R}^p$, $\operatorname{prox}_{\alpha^{-1}F}(w) = w$, we derive

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma\alpha^{-1}X^T(X\widehat{w}_t - \widehat{y}), \tag{3.7}$$

which coincides with the Landweber iteration for solving Problem **??**, studied in the context of regression in [**?**]. ADGD provides a FISTA variant of Landweber iteration, for which we prove here regularization properties, see also [**?**].

The previous remark shows that the proposed algorithms are generalizations of the Landweber iteration for a more general penalty term of the form in (**??**). The next remark shows that the choice of $F \neq 0$ is in a sense equivalent to a gradient descent algorithm with respect to a different geometry.

Remark 3.2 (Connection to Mirror Descent and Linearized Bregman algorithms). Define $G\colon \mathbb{R}^p \to \mathbb{R}^n$, $G(w) = \|Xw - \widehat{y}\|^2/2$, for every $w \in \mathbb{R}^p$. Each step of the Landweber iteration (**??**) coincides with a gradient descent step on G, e.g. with the solution of the minimization problem

$$\widehat{w}_{t+1} = \operatorname{argmin}\left\{G(\widehat{w}_t) + \langle \nabla G(\widehat{w}_t), w - \widehat{w}_t\rangle + \frac{\gamma^{-1}\alpha}{2}\|w - \widehat{w}_t\|^2\right\}, \tag{3.8}$$

where the objective function is linearized and regularized with the Euclidean distance. The idea behind mirror descent [**?**] is to replace the Euclidean distance with the Bregman divergence induced by R and, starting from $\widetilde{w}_0 \in \mathbb{R}^p$ and $\widetilde{v}_0 \in \partial R(\widetilde{w}_0)$, to consider the iteration

$$\begin{cases}\widetilde{v}_t \in \partial R(\widetilde{w}_t)\\ \widetilde{w}_{t+1} = \operatorname{argmin}\left\{G(\widetilde{w}_t) + \langle \nabla G(\widetilde{w}_t), w - \widetilde{w}_t\rangle + \gamma^{-1}\alpha\left(R(w) - R(\widetilde{w}_t) - \langle \widetilde{v}_t, w - \widetilde{w}_t\rangle\right)\right\}\end{cases} \tag{3.9}$$

which yields

$$0 \in \gamma\alpha^{-1}\nabla G(\widetilde{w}_t) + \partial R(\widetilde{w}_{t+1}) - \widetilde{v}_t. \tag{3.10}$$

If we define $\widetilde{v}_{t+1} = \widetilde{v}_t - \gamma\alpha^{-1}\nabla G(\widetilde{w}_t)$ we obtain

$$\widetilde{v}_{t+1} \in \partial R(\widetilde{w}_{t+1})$$

The Young-Fenchel equality yields

$$\widetilde{w}_{t+1} = \nabla R^*(\widetilde{v}_{t+1}).$$

and therefore

$$\begin{cases}\widetilde{w}_{t+1} = \operatorname{prox}_{\alpha^{-1}F}(\alpha^{-1}\widetilde{v}_{t+1})\\ \widetilde{v}_{t+1} = \widetilde{v}_t - \gamma\alpha^{-1}X^T(X\widetilde{w}_t - \widehat{y})\end{cases}$$

Such an iteration corresponds to a change of variables in the DGD algorithm. Indeed, if we start from DGD, it is easy to see that $\widetilde{v}_t = -X^T\widehat{w}_t$ for every $t \in \mathbb{N}$ satisfies the update rule in (**??**). The point of view taken above is discussed and studied in a series of paper [**?**,**?**,**?**,**?**,**?**,**?**], where however no acceleration is considered.

Remark 3.3 (Connection to Bregman iterations). Another algorithm used to solve constrained optimization problems of the form (**??**) is the Bregman iteration introduced in [**?**] (closely related to the augmented lagrangian method [**?**]). As before, define the Bregman divergence induced by R and, starting from $\widetilde{w}_0 \in \mathbb{R}^p$ and $\widetilde{v}_0 \in \partial R(\widetilde{w}_0)$, and consider the iteration

$$\begin{cases} \widetilde{v}_t \in \partial R(\widetilde{w}_t) \\ \widetilde{w}_{t+1} = \operatorname{argmin} \left\{ G(w) + \gamma^{-1}\alpha \left(R(w) - R(\widetilde{w}_t) - \langle \widetilde{v}_t, w - \widetilde{w}_t \rangle \right) \right\}. \end{cases} \tag{3.11}$$

Regularization properties of (**??**) have been studied in [**?**] (see also references therein), but the main drawback is the fact that the solution of the subproblems in (**??**) are rarely available in closed form and require the solution of a regularized problem at each iteration. To tackle this issue, the linearized Bregman iteration described in the previous remark is introduced

While it is well known that early stopping of the Landweber iteration leads to stable approximations of the minimal norm solution of an inverse problem, here we generalize such a result to obtain stable approximations of the solution defined by a wide class of regularizing functionals. The presence of the additional term F in the regularization function introduces in the algorithm a (nonlinear) proximal operation.

4. Theoretical analysis

In this section, we present and discuss the main results of the paper.

4.1. *Properties of DGD with and without averaging*

We start with DGD and state an early stopping result for two variants of the method, namely with and without iterates averaging.

Theorem 4.1 (Dual gradient descent). *Let $\delta \in]0,1]$ and consider the (DGD) algorithm. Assume that there exists $\bar{v} \in \mathbb{R}^p$ such that $-X^T\bar{v} \in \partial R(w^\dagger)$. Set $a = 2\|X\|^{-1}$ and $b = \|X\|\|v^\dagger\|\alpha^{-1}$, where $v^\dagger$ is a solution of the dual problem of* (**??**). *Then the following hold:*

(i) For every $t \in \mathbb{N}$,

$$\|\widehat{u}_t - w^\dagger\| \le a t^{1/2}\delta + b t^{-1/2}. \tag{4.1}$$

In particular, choosing $t_\delta = \lceil c\delta^{-1} \rceil$ for some $c > 0$, we derive

$$\|\widehat{u}_{t_\delta} - w^\dagger\| \le \left[a(c^{1/2} + 1) + b c^{-1/2} \right] \delta^{1/2}. \tag{4.2}$$

(ii) There exists $t_\delta \in \{\lfloor c\delta^{-1}\rfloor, \dots, 2\lfloor c\delta^{-1}\rfloor\}$ *such that*

$$\|\widehat{w}_{t_\delta} - w^\dagger\| \le \big[a(c^{1/2}+1) + bc^{-1/2}\big]\delta^{1/2}. \tag{4.3}$$

The proof of this result can be found in Section **??**; here we briefly discuss it. Regarding the assumptions, the condition $-X^T\overline{v} \in \partial R(w^\dagger)$ can be interpreted as an abstract regularity condition on the subdifferential of R at the solution of interest [**?**]. When $R = \|\cdot\|^2/2$, and more generally when R is real-valued, it is automatically satisfied under our assumptions, and it corresponds to what is called a source condition [**?**,**?**].

As anticipated in Section **??**, the bounds in (**??**) and (**??**) are derived by optimizing a stability plus regularization/optimization bound. Note in particular that the constants appearing in the regularization error are determined by the strong convexity constant of R, the norm of a dual solution, and the norm of the operator X. The above result shows that, given a noise level δ, regularization is achieved by computing a suitable number t_δ of iterations of DGD, both for the averaged and the original sequence. The number of required iterations is a decreasing function of the noise level, guaranteeing efficiency of the proposed approach, and tends to infinity as the noise goes to zero. The definition of t_δ in Theorem **??** is an early stopping rule. The dependence of the noise that we get in Theorem **??** is optimal [**?**], and coincides with the Tikhonov regularization one.

The stability result that we obtain for the original sequence $\widehat{w}_t$ does not hold for every $t \in \mathbb{N}$, see equation (**??**) later in the proof. Nevertheless, we are still able to derive an early stopping rule for the non averaged sequence. This may be useful when structural properties of the solution are of interest. The main example is sparsity: if the solution to be recovered is sparse, averaging of the iterates is not appropriate, since the averaged iterates would not be sparse, invalidating for instance the advantages of a soft-thresholding step.

4.2. *Properties of ADGD*

Next, we derive an analogous of Theorem **??** for the accelerated variant of the algorithm.

Theorem 4.2 (Accelerated dual gradient descent). *Let* $\delta \in \,]0,1]$ *and let* $(\widehat{w}_t)_{t\in\mathbb{N}}$ *be the sequence generated by ADGD. Assume that there exists* $\overline{v} \in \mathbb{R}^p$ *such that* $-X^T\overline{v} \in \partial R(w^\dagger)$. *Set* $a = 4\|X\|^{-1}$ *and* $b = 2\|X\|\|v^\dagger\|\alpha^{-1}$, *where* $v^\dagger$ *is a solution of the dual problem of* (**??**). *Then, for every* $t \ge 3$,

$$\|\widehat{w}_t - w^\dagger\| \le at\delta + bt^{-1}. \tag{4.4}$$

In particular, choosing $t_\delta = \lceil c\delta^{-1/2}\rceil$ *for some* $c > 0$,

$$\|\widehat{w}_t - w^\dagger\| \le \big[a(c+1) + bc^{-1}\big]\delta^{1/2}. \tag{4.5}$$

As can be directly observed, the results proved in Theorem **??** are analogous to the ones in Theorem **??**. In particular, the constants involved in the bounds are the ones appearing in the DGD algorithm. As for DGD, the early stopping rule gives optimal dependence with respect to the noise, and again coincides with the Tikhonov one. The difference between DGD and ADGD is in the computational aspects: indeed, to achieve the same recovery accuracy, a number of iterations of the order of δ^{-1} are needed for the basic scheme, and only $\delta^{-1/2}$ iterations are needed for the accelerated method. This kind of result resembles the behaviour of the ν-method for the minimal norm solution [**?**].

4.3. *Results for inexact proximity operators*

In some interesting cases, the proximity operator is not available in closed form, but can be still computed at reasonable cost, see [**?**,**?**] for a throughout discussion. The results in Theorem **??** and **??** hold also if the proximity operator is computed inexactly. The proof is a straightforward generalization of the one for the exact case. We include it for the sake of completeness.

Theorem 4.3 (Dual gradient descent with inexact prox). *Let $\delta \in \,]0,1]$. Let $(\widehat{u}_t)_{t\in\mathbb{N}}$ be the averaged sequence generated by IDGD. Assume that there exists $\bar{v} \in \mathbb{R}^p$ such that $-X^T\bar{v} \in \partial R(w^\dagger)$. Set $a = 12(1+\|X\|)\|X\|^{-1}$ and $b = \|X\|\|v^\dagger\|\alpha^{-1}$, where $v^\dagger$ is a solution of the dual problem of* (**??**). *Then, for every $t \in \mathbb{N}$,*

$$\|\widehat{u}_t - w^\dagger\| \leq at^{1/2}\delta + bt^{-1/2}. \tag{4.6}$$

In particular, choosing $t_\delta = \lceil c\delta^{-1}\rceil$ for some $c > 0$, we derive

$$\|\widehat{u}_{t_\delta} - w^\dagger\| \leq \left[a(c^{1/2}+1) + bc^{-1/2}\right]\delta^{1/2}. \tag{4.7}$$

Before discussing the above result, we state an analogous result for the accelerated variant.

Theorem 4.4 (Accelerated dual gradient descent with inexact prox). *Let $\delta \in \,]0,1]$ and let $(\widehat{w}_t)_{t\in\mathbb{N}}$ be the sequence generated by IADGD. Assume that there exists $\bar{v} \in \mathbb{R}^p$ such that $-X^T\bar{v} \in \partial R(w^\dagger)$. Set $a = 4(1+\|X\|)\|X\|^{-1}$ and $b = 2\|X\|\|v^\dagger\|/\alpha$, where $v^\dagger$ is a solution of the dual problem of* (**??**). *Then, for every $t \geq 2$,*

$$\|\widehat{w}_t - w^\dagger\| \leq at\delta + bt^{-1}. \tag{4.8}$$

In particular, choosing $t_\delta = \lceil c\delta^{-1/2}\rceil$ for some $c > 0$,

$$\|\widehat{w}_t - w^\dagger\| \leq \left[a(c+1) + bc^{-1}\right]\delta^{1/2}. \tag{4.9}$$

Remark 4.1 (Beyond worst case). While we considered a general regularization R and obtained worst-case results, an interesting question is if these results can be improved under additional assumptions on R, e.g. assuming it is sparsity inducing. We refer to [**?**] and [**?**,**?**] for some results in this direction.

4.4. *Comparison with related work*

We next compare our iterative regularization methods with related work. The case $R = \|\cdot\|^2$ is classic, see [**?**]. In [**?**] an iterative regularization procedure based on the so called Bregman iteration is considered. An early stopping rule based on a discrepancy principle in the case of noisy data is also presented. There is one main difference with respect to our contribution. Each DGD or ADGD step does not require inner algorithms if the proximity operator is available in closed form, while Bregman iteration requires the solution of a nontrivial minimization problem at each step. Such step is computationally as costly as solving a Tikhonov regularized problem. A stability analysis for the Bregman iteration is presented in Theorem 4.2 in [**?**], while weak convergence without the strong convexity assumption is proved in [**?**]. A qualitative early stopping rule for the DGD algorithm has been considered in [**?**] for the total variation case. Finally, a related algorithm to the DGD is the one considered in [**?**]. The setting of [**?**] is more general than ours, but the obtained results are weaker: the stopping rule is of the form $O(\delta^{-2})$ and no quantitative bounds of $\|\widehat{w}_{t_\delta} - w^\dagger\|$ are given. In this paper we focused on general strongly convex regularization functions. For results related to general data fit terms see [**?**,**?**].

5. Proofs: convergence and stability

In this section we develop the proofs of our main results. The proof of Theorem **??** is based on a decomposition of the error to be estimated in two terms. The idea is to build an auxiliary sequence and to majorize the error with the sum of two quantities that can be interpreted as a stability and an optimization (regularization) error, respectively. Bounds on these two terms are then provided. We first introduce the corresponding algorithm to solve the target problem in (**??**). This algorithm is not used in practice, but is needed only for the theoretical analysis, and is the noise free version of DGD, where $\widehat{y}$ is replaced by y. Starting from $v_0 = 0$, the t-th iteration is defined by the gradient descent algorithm applied to the dual of problem (**??**) (see Appendix **??** for the definition of dual problem)

$$w_t = \operatorname{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T v_t\big), \quad v_{t+1} = v_t + \gamma(Xw_t - y), \quad u_t = \sum_{k=1}^{t} w_k/t. \tag{5.1}$$

Note that u_t is defined only for $t > 0$.

The distance of the noisy solution from the ideal one is upper bounded as follows:

$$\|\widehat{u}_t - w^\dagger\| \le \|\widehat{u}_t - u_t\| + \|u_t - w^\dagger\|.$$

The term $\|u_t - w^\dagger\|$ is called approximation, but also optimization or regularization error and it vanishes for increasing t. The term $\|\widehat{u}_t - u_t\|$ measures stability and it increases with t and δ. The choice of the early stopping bound t_δ is obtained optimizing the resulting upper bound with respect to $t \in \mathbb{N}$.

The stopping rule for ADGD follows analogously from a general result about convergence of proximal methods in the presence of computational errors [**?**]. In this case, a similar upper bound is derived from a slightly different analysis, where the error is not explicitly split in two terms.

The case of implementation with errors can be treated in a very similar way. Interestingly, computational errors comparable to the noise level does not impact the overall reconstruction capabilities of the proposed method.

5.1. *Proof of Theorem* ??

Here we proof the convergence and stability results for the DGD algorithm and derive the proof Theorem **??**.

Theorem 5.1 (Convergence of DGD). *Consider the iterates generated by the DGD algorithm in* (**??**)*. Assume that there exists $\bar{v} \in \mathbb{R}^p$ such that*

$$-X^T\bar{v} \in \partial R(w^\dagger)$$

and let $v^\dagger$ be a solution of the dual problem. Then

$$\|w_t - w^\dagger\| \leq \frac{\|X\|\|v^\dagger\|}{\alpha\sqrt{t}} \quad \text{and} \quad \|u_t - w^\dagger\| \leq \frac{\|X\|\|v^\dagger\|}{\alpha\sqrt{t}}.$$

Proof. Lemma **??** implies that the dual problem has a solution $v^\dagger$. For every $v \in \mathbb{R}^p$, let $D(v) = R^*(-X^T v) + \langle \widehat{y}, v\rangle$. Then it holds (see for instance [**?**, Theorem 3.1]) that $D(v_t) - D(v^\dagger)$ is a decreasing sequence and

$$\sum_{k=1}^{t} \big(D(v_k) - D(v^\dagger)\big) \leq \frac{\|X\|^2\|v_0 - v^\dagger\|^2}{2\alpha}, \tag{5.2}$$

and therefore

$$D(v_t) - D(v^\dagger) \leq \frac{\|X\|^2\|v_0 - v^\dagger\|^2}{2\alpha t}. \tag{5.3}$$

Next, Lemma **??** yields

$$\frac{\alpha}{2}\|w_t - w^\dagger\|^2 \leq D(v_t) - D(v^\dagger).$$

Combining the two inequalities, and recalling that $v_0 = 0$, we derive

$$\|w_t - w^\dagger\| \leq \frac{\|X\|\|v^\dagger\|}{\alpha\sqrt{t}}.$$

Let $t > 1$. To derive the statement about the averaged iterates (u_t) note that, by convexity of the squared norm and equation (??)

$$\begin{aligned}\|u_t - w^\dagger\|^2 &= \|\sum_{k=1}^{t} w_k/t - w^\dagger\|^2\\ &\leq (1/t)\sum_{k=1}^{t}\|w_k - w^\dagger\|^2\\ &\leq (1/t)\sum_{k=1}^{t}\big(D(v_k) - D(v^\dagger)\big)\\ &\leq \frac{\|X\|^2\|v^\dagger\|^2}{\alpha^2 t}\end{aligned}$$ □

Next we prove the main stability result.

Proposition 5.1 (Stability of DGD). *Consider the sequences generated by DGD. Let $(w_t)_{t\in\mathbb{N}}$ and $(u_t)_{t>0}$ be defined as in (??). Assume $\delta < 1$. Then the following hold:*

i) There exists $t_\delta \in \{\lfloor 1/\delta\rfloor, \ldots, 2\lfloor 1/\delta\rfloor\}$ such that

$$\|w_{t_\delta} - \widehat{w}_{t_\delta}\| \leq 3\|X\|^{-1}\delta t_\delta^{1/2}. \tag{5.4}$$

ii) For every $t \in \mathbb{N}$,

$$\|u_t - \widehat{u}_t\| \leq 3\|X\|^{-1}\delta t^{1/2}. \tag{5.5}$$

Proof.
i): For every $t \in \mathbb{N}$, using the firm nonexpansiveness of $\operatorname{prox}_{\alpha^{-1}F}$ (see Proposition ??) and the definition of γ

$$\begin{aligned}\|\widehat{v}_t - v_t + \gamma X(\widehat{w}_t - w_t)\|^2 &= \|\widehat{v}_t - v_t\|^2 + 2\gamma\langle \widehat{v}_t - v_t, X(\widehat{w}_t - w_t)\rangle + \gamma^2\|X(\widehat{w}_t - w_t)\|^2\\ &\leq \|\widehat{v}_t - v_t\|^2 - 2\gamma\alpha\|\widehat{w}_t - w_t\|^2 + \gamma^2\|X(\widehat{w}_t - w_t)\|^2\\ &\leq \|\widehat{v}_t - v_t\|^2 - \gamma\alpha\|\widehat{w}_t - w_t\|^2\\ &\leq \|\widehat{v}_t - v_t\|^2. \end{aligned}\tag{5.6}$$

Consequently,

$$\begin{aligned}\|\widehat{v}_{t+1} - v_{t+1}\| &= \|\widehat{v}_t - v_t + \gamma X(\widehat{w}_t - w_t) - \gamma(\widehat{y} - y)\|\\ &\leq \|\widehat{v}_t - v_t\| + \gamma\delta\end{aligned}$$

and therefore

$$\|\widehat{v}_{t+1} - v_{t+1}\| \leq \gamma\delta(t+1). \tag{5.7}$$

Moreover,

$$\begin{aligned}\|\widehat{v}_t - v_t + \gamma X(\widehat{w}_t - w_t)\|^2 &= \|\widehat{v}_{t+1} - v_{t+1} + \gamma(\widehat{y} - y)\|^2\\ &= \|\widehat{v}_{t+1} - v_{t+1}\|^2 + \gamma^2\|y - \widehat{y}\|^2 + 2\gamma\langle \widehat{v}_{t+1} - v_{t+1}, \widehat{y} - y\rangle\\ &\geq \|\widehat{v}_{t+1} - v_{t+1}\|^2 + \gamma^2\|y - \widehat{y}\|^2 - 2\gamma\delta\|\widehat{v}_{t+1} - v_{t+1}\|.\end{aligned}$$

Hence, (**??**) and (**??**) yield

$$\begin{aligned}\gamma\alpha\|\widehat{w}_t - w_t\|^2 &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_t - v_t + \gamma X(\widehat{w}_t - w_t)\|^2\\ &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_{t+1} - v_{t+1}\|^2 + 2\gamma\delta\|\widehat{v}_{t+1} - v_{t+1}\|\\ &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_{t+1} - v_{t+1}\|^2 + 2\gamma^2\delta^2(t+1). \qquad (5.8)\end{aligned}$$

Summing the previous inequality for $t \in \{T, \ldots, 2T-1\}$, for some $T \geq 1$, we derive

$$\gamma\alpha\sum_{t=T}^{2T-1}\|\widehat{w}_t - w_t\|^2 \leq \|\widehat{v}_T - v_T\|^2 + 2\gamma^2\delta^2T^2 \leq 3\gamma^2\delta^2T^2 \qquad (5.9)$$

Taking $T = \lfloor c/\delta \rfloor$ it follows that

$$\sum_{t=T}^{2T-1}\|\widehat{w}_t - w_t\|^2 \leq 3\|X\|^{-2}\delta^2\lfloor c/\delta\rfloor^2.$$

Thus there exists at least a $t_\delta \in \{\lfloor c/\delta\rfloor, \ldots, 2\lfloor c/\delta\rfloor\}$ such that

$$\|\widehat{w}_{t_\delta} - w_{t_\delta}\|^2 \leq 3\|X\|^{-2}\delta^2\left\lfloor\frac{c}{\delta}\right\rfloor \leq 3\|X\|^{-2}\delta^2 t_\delta. \qquad (5.10)$$

ii): Summing the inequalities in (**??**) for $t = 1, \ldots, T$ we derive:

$$\gamma\alpha\sum_{t=1}^{T}\|\widehat{w}_t - w_t\|^2 \leq 3\gamma^2\delta^2T^2 \qquad (5.11)$$

Convexity of $\|\cdot\|^2$ implies

$$\|\widehat{u}_T - u_T\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\widehat{w}_t - w_t\|^2 \leq 3\|X\|^{-2}\delta^2T \qquad (5.12)$$

□

Proof. [**of Theorem ??**] Theorem **??** and Proposition **??** imply

$$\|\widehat{u}_t - w^\dagger\| \leq at^{1/2}\delta + bt^{-1/2}.$$

Since $t_\delta = \lceil c\delta^{-1}\rceil$, we have $c\delta^{-1} \leq t_\delta \leq c\delta^{-1} + 1$, therefore

$$\|\widehat{u}_t - w^\dagger\| \leq at^{1/2}\delta + bt^{-1/2} \leq a(c\delta^{-1}+1)^{1/2}\delta + bc^{-1/2}\delta^{1/2}.$$

The statement follows noting that $(c\delta^{-1}+1)^{1/2} \leq (c^{1/2}+1)\delta^{-1/2}$. The statement for the sequence $\widehat{w}_t$ follows analogously. □

5.2. *Proof of Theorem* ??

The following lemma characterizes the asymptotic behavior of the sequence $(\theta_t)_{t\in\mathbb{N}}$.

Lemma 5.1. *Let $(\theta_t)_{t\in\mathbb{N}}$ be the sequence defined in ADGD. Then, for every $t \in \mathbb{N}$*

$$\frac{t+1}{2} \leq \theta_t \leq t+1$$

Proof. We prove the first inequality by induction. The case $t = 0$ is clear since $\theta_0 = 1$. Now suppose that the inequality is true for t. We derive

$$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2} \geq \frac{1+\sqrt{1+(t+1)^2}}{2} \geq \frac{t+2}{2}.$$

For the second inequality, the case $t = 0$ is also clear. Now suppose that the inequality is true for t. We derive

$$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2} \leq \frac{1+\sqrt{1+4(t+1)^2}}{2} \leq \frac{1+1+2(t+1)}{2} = t+2. \qquad \square$$

The following theorem is obtained exploiting existing results on convergence of accelerated forward-backward algorithm in the presence of computational errors. In particular, the result is derived combining [?, Proposition 3.3] (see also [?,?] for related results) with Lemma ?? and the relationship between convergence of the dual objective function and the primal iterates.

Theorem 5.2. *Let $(\widehat{w}_t)_{t\in\mathbb{N}}$ be the sequence generated by ADGD. Then, for every $t \in \mathbb{N}$, $t \geq 3$,*

$$\|\widehat{w}_t - w^\dagger\| \leq \frac{2\|X\|\|v^\dagger\|}{\alpha t} + 4\|X\|^{-1}\delta t. \tag{5.13}$$

Proof. For every $v \in \mathbb{R}^p$, let $D(v) = R^*(-X^T v) + \langle \widehat{y}, v\rangle$. Then Lemma ?? yields

$$(\forall t \in \mathbb{N}) \qquad \frac{\alpha}{2}\|\widehat{w}_t - w^\dagger\|^2 \leq D(\widehat{v}_t) - \min_{v\in\mathbb{R}^p} D(v). \tag{5.14}$$

Proposition 3.3 in [?] and Lemma ?? imply

$$\begin{aligned} D(\widehat{v}_t) - \min_{v\in\mathbb{R}^p} D(v) &\leq \frac{1}{2\gamma\theta_t^2}\left(\|v^\dagger\| + \gamma\delta\sum_{k=0}^{t}\theta_k\right)^2 \qquad (5.15)\\ &\leq \frac{1}{\gamma(t+1)^2}\left(\|v^\dagger\| + \gamma\delta\frac{(t+2)(t+3)}{2}\right)^2 \end{aligned}$$

we derive

$$\|\widehat{w}_t - w^\dagger\| \leq \frac{2\|X\|\|v^\dagger\|}{\alpha t} + \frac{4}{\|X\|}\delta t \tag{5.16}$$

$\square$

Finally, we prove Theorem ??.

Proof. [**of Theorem ??**] Theorem **??** yields

$$\|\widehat{w}_t - w^\dagger\| \leq at\delta + bt^{-1}.$$

Since $t_\delta = \lceil c\delta^{-1/2}\rceil$, we have $c\delta^{-1/2} \leq t_\delta \leq c\delta^{-1/2} + 1$, therefore

$$\|\widehat{u}_t - w^\dagger\| \leq at\delta + bt^{-1} \leq a(c\delta^{-1/2} + 1)\delta + bc^{-1}\delta^{1/2}.$$

The statement follows noting that $(c\delta^{-1/2} + 1) \leq (c+1)\delta^{-1/2}$. □

5.3. *Proof for inexact versions*

In this section we sketch the proof Theorems **??** and **??**. Note that we only need to prove stability results, since the error decomposition will be the same one used for the exact case. We start with a stability result for the IDGD algorithm, whose proof follows the same line as the one of Proposition **??**, but deals with an additional error term due to the inexact computation of the proximal step.

Proposition 5.2 (Stability of IDGD). *Consider the sequences generated by IDGD. Let $(w_t)_{t\in\mathbb{N}}$ and $(u_t)_{t>0}$ be defined as in (**??**). Assume $\delta < 1$. Then the following hold:*

i) There exists $t_\delta \in \{\lfloor 1/\delta\rfloor, \ldots, 2\lfloor 1/\delta\rfloor\}$ such that

$$\|w_{t_\delta} - \widehat{w}_{t_\delta}\| \leq 12(1 + \|X\|)\|X\|^{-1}\delta t_\delta^{1/2}. \tag{5.17}$$

ii) For every $t > 1$,

$$\|u_t - \widehat{u}_t\| \leq 10(1 + \|X\|)\|X\|^{-1}\delta t_\delta^{1/2}. \tag{5.18}$$

Proof.
i): We need to introduce an additional auxiliary iteration,

$$\widetilde{w}_t = \operatorname{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big). \tag{5.19}$$

Since $\widehat{w}_t \approx_\delta \operatorname{prox}_{\alpha^{-1}F}(-\alpha^{-1}X^T\widehat{v}_t)$, equation (**??**) yields

$$(\exists e_t \in \mathbb{R}^p) \quad \widehat{w}_t = \widetilde{w}_t + e_t, \text{ with } \|e_t\| \leq \delta. \tag{5.20}$$

For every $t \in \mathbb{N}$, using the firm nonexpansiveness of $\operatorname{prox}_{\alpha^{-1}F}$ (see Proposition **??**) and the definition of γ

$$\begin{aligned}
\|\widehat{v}_t - v_t + \gamma X(\widetilde{w}_t - w_t)\|^2 &= \|\widehat{v}_t - v_t\|^2 + 2\gamma\langle\widehat{v}_t - v_t, X(\widetilde{w}_t - w_t)\rangle + \gamma^2\|X(\widetilde{w}_t - w_t)\|^2\\
&\leq \|\widehat{v}_t - v_t\|^2 - 2\gamma\alpha\|\widetilde{w}_t - w_t\|^2 + \gamma^2\|X(\widetilde{w}_t - w_t)\|^2\\
&\leq \|\widehat{v}_t - v_t\|^2 - \gamma\alpha\|\widetilde{w}_t - w_t\|^2\\
&\leq \|\widehat{v}_t - v_t\|^2.
\end{aligned} \tag{5.21}$$

Consequently, recalling (**??**)

$$\begin{aligned}
\|\widehat{v}_{t+1} - v_{t+1}\| &= \|\widehat{v}_t - v_t + \gamma X(\widetilde{w}_t - w_t) - \gamma(\widehat{y} - Xe_t - y)\|\\
&\leq \|\widehat{v}_t - v_t\| + \gamma(1 + \|X\|)\delta
\end{aligned}$$

and therefore

$$\|\widehat{v}_{t+1} - v_{t+1}\| \leq \gamma(1 + \|X\|)\delta(t+1). \tag{5.22}$$

Moreover,

$$\begin{aligned}\|\widehat{v}_t - v_t + \gamma X(\widetilde{w}_t - w_t)\|^2 &= \|\widehat{v}_{t+1} - v_{t+1} + \gamma(\widehat{y} - Xe_t - y)\|^2 \\ &= \|\widehat{v}_{t+1} - v_{t+1}\|^2 + \gamma^2\|\widehat{y} - Xe_t - y\|^2 + 2\gamma\langle \widehat{v}_{t+1} - v_{t+1}, \widehat{y} - Xe_t - y\rangle \\ &\geq \|\widehat{v}_{t+1} - v_{t+1}\|^2 + \gamma^2\|\widehat{y} - Xe_t - y\|^2 - 2\gamma(1 + \|X\|)\delta\|\widehat{v}_{t+1} - v_{t+1}\|.\end{aligned}$$

Hence, (**??**) and (**??**) yield

$$\begin{aligned}\gamma\alpha\|\widetilde{w}_t - w_t\|^2 &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_t - v_t + \gamma X(\widetilde{w}_t - w_t)\|^2 \\ &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_{t+1} - v_{t+1}\|^2 + 2\gamma(1 + \|X\|)\delta\|\widehat{v}_{t+1} - v_{t+1}\| \\ &\leq \|\widehat{v}_t - v_t\|^2 - \|\widehat{v}_{t+1} - v_{t+1}\|^2 + 2\gamma^2(1 + \|X\|)^2\delta^2(t+1). \end{aligned} \tag{5.23}$$

Summing the previous inequality for $t \in \{T, \dots, 2T-1\}$, for some $T \geq 1$, we derive

$$\begin{aligned}\gamma\alpha\sum_{t=T}^{2T-1}\|\widehat{w}_t - w_t\|^2 &\leq 2\gamma\alpha\sum_{t=T}^{2T-1}\left(\|\widehat{w}_t - \widetilde{w}_t\|^2 + \|\widetilde{w}_t - w_t\|^2\right) \\ &\leq 2\gamma\alpha T\delta^2 + 2\|\widehat{v}_T - v_T\|^2 + 8\gamma^2(1 + \|X\|^2)\delta^2T^2 \leq 12\gamma^2(1 + \|X\|)^2\delta^2T^2\end{aligned}$$

Taking $T = \lfloor 1/\delta \rfloor$ it follows that

$$\sum_{t=T}^{2T-1}\|\widehat{w}_t - w_t\|^2 \leq 12(1 + \|X\|)^2\|X\|^{-2}\delta^2\lfloor 1/\delta\rfloor^2.$$

Thus there exists at least a $t_\delta \in \{\lfloor 1/\delta\rfloor, \dots, 2\lfloor 1/\delta\rfloor\}$ such that

$$\|\widehat{w}_{t_\delta} - w_{t_\delta}\|^2 \leq 12(1 + \|X\|)^2\|X\|^{-2}\delta^2\left\lfloor\frac{1}{\delta}\right\rfloor \leq 12(1 + \|X\|)^2\|X\|^{-2}\delta^2 t_\delta.$$

ii): Summing the inequalities in (**??**) for $t = 0, \dots, T$, and the fact that $\widehat{w}_0 = w_0$ we derive:

$$\gamma\alpha\sum_{t=0}^{T}\|\widehat{w}_t - w_t\|^2 \leq 10\gamma^2(1 + \|X\|)^2\delta^2T^2 \tag{5.24}$$

Convexity of $\|\cdot\|^2$ implies

$$\|\widehat{u}_T - u_T\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\widehat{w}_t - w_t\|^2 \leq 10\gamma^2(1 + \|X\|)^2\delta^2T. \tag{5.25}$$

□

The following theorem is similar to Theorem **??**. We write the main steps of the proof for completeness.

Theorem 5.3. *Let $(\widehat{w}_t)_{t\in\mathbb{N}}$ be the sequence generated by IADGD. Then, for every $t \in \mathbb{N}$, $t \geq 1$,*

$$\|\widehat{w}_t - w^\dagger\| \leq \frac{2\|X\|\|v^\dagger\|}{\alpha t} + 4(1 + \|X\|)\|X\|^{-1}\delta t. \tag{5.26}$$

Proof. We introduce an auxiliary iteration

$$\widetilde{w}_t = \operatorname{prox}_{\alpha^{-1}F}\big(-\alpha^{-1}X^T\widehat{v}_t\big). \tag{5.27}$$

Since $\widehat{w}_t \approx_\delta \operatorname{prox}_{\alpha^{-1}F}(-\alpha^{-1}X^T\widehat{v}_t)$, equation (**??**) yields

$$(\exists e_t \in \mathbb{R}^p) \quad \widehat{w}_t = \widetilde{w}_t + e_t, \text{ with } \|e_t\| \le \delta. \tag{5.28}$$

and this implies that the IADGD is an accelerated gradient method with errors applied to the dual of the original problem, where at each step the error we make on the true gradient is

$$Xe_t + \widehat{y} - y, \text{ with } \|Xe_t + \widehat{y} - y\| \le (\|X\| + 1)\delta.$$

The proof then proceeds as the one of Theorem **??**, with δ replaced by $(\|X\|+1)\delta$ in (**??**). □

Theorems **??** and **??** can be proved in the same way as Theorems **??** and **??**, respectively.

6. Numerical experiments

In this section we compare our iterative regularization techniques (DGD and ADGD with early stopping) with Tikhonov regularization on three different problems: variable selection, matrix completion, and image deblurring. The performance of the Tikhonov regularization scheme depends of course on the chosen algorithm to solve the regularized problems. We use state of the art techniques: accelerated proximal gradient descent with warm-restart [**?**]. The model selection phase is performed as follows: we first solve the regularized problem with a very large value λ_0, and then for the sequence $\lambda_i = 2^{-i}\lambda_0$. Since in practice the noise level is unknown, we choose λ using holdout cross-validation keeping 1/10 of the available points for validation. For initializing the accelerated gradient descent on the regularized problem we use the warm-restarting trick, which is known (in practice) to dramatically accelerate the computation of the regularization path [**?**]. The comparison relies heavily on the stopping rule used for stopping the iteration computing the minimizer of the Tikhonov regularized functional. We used a very loose stopping rule for the algorithm for a given λ_i to make Tikhonov regularization more competitive. More precisely the iterations were stopped when the distance between two successive iterations was less than $0.001 \cdot \delta$. Since accelerated proximal gradient descent involves steps with the same computational complexity to those of DGD and ADGD, the comparison between the three approaches is made in terms of number of iterations. The number of iterations for Tikhonov regularization is the total number of iterations for all different λ values.

6.1. *Variable selection*

We consider a linear regression problem with $n = 500$ examples and $p = 2000$ variables. We assume that $\widehat{y} \in \mathbb{R}^{500}$ is obtained corrupting with a Gaussian noise of mean zero and variance $\delta/\sqrt{n}$ a measurement Xw_*, where w_* is a vector having a small number of nonzero components (10, 30, or 60, respectively). In this example, the covariates are correlated with a random covariance matrix Σ with $\Sigma = C^T C$, where C is a random matrix with entries drawn independently at random from a gaussian distribution with standard deviation 0.1. To perform variable selection, and obtain a sparse estimator we apply our iterative regularization methods, DGD and ADGD, to the elastic net regularizing function $R(w) = \|w\|_1 + (\alpha/2)\|w\|^2$. We compared the number of iterations of DGD, ADGD, and Tikhonov regularization on 50 different realizations of sample points. The parameters were chosen using a validation set of 100 samples. The results are shown in Table **??**. For Tikhonov regularization, we used a second least squares step on the selected variables to compute the validation score, requiring an extra computation load that we did not quantify here. It is worth noticing that iterative regularization does not require this further step. The results suggest that Tikhonov regularization and iterative regularization algorithms have very similar prediction and variable selection performances. DGD is approximately as fast as state of the art variational regularization, while ADGD is much faster.

6.2. *Matrix completion*

We consider the problem of recovering a low-rank data matrix $W \in \mathbb{R}^{n\times p}$ from a sampling of its entries. We denote by Ω the subset of indices corresponding to sampled entries. We find an approximate solution of this problem by minimizing a strongly convex relaxation [**?**] given by the sum of the nuclear norm with the squared Frobenius norm, that is:

$$\min_{\mathcal{X}W=\widehat{Y}} \|W\|_* + \frac{\alpha}{2}\|W\|_F^2, \tag{6.1}$$

where $\widehat{Y} \in \mathbb{R}^{n\times p}$, is such that, for every $(i,j) \notin \Omega$, $\widehat{Y}_{i,j} = 0$, and $\mathcal{X} \colon \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is such that $(\mathcal{X}W)_{i,j} = W_{i,j}$ if $(i,j) \in \Omega$ and 0 otherwise. DGD applied to this problem is the Singular Value Thresholding (SVT) algorithm described in [**?**] and note that, interestingly, ADGD is its accelerated counterpart. The most expensive computational part is the proximal step, which requires an SVD decomposition [**?**]. While in [**?**] the authors apply the algorithm to noisy data, they then propose as an improvement a different relaxation [**?**]. Here we show that SVT with early stopping is indeed an efficient algorithm to deal with matrix completion of noisy data. We tested the performance on simulated data using a standard procedure described in [**?**]. We multiplied random gaussian matrices with independent entries and variance 1 of size $n \times r$ and $r \times p$ where r is the chosen rank, and then we added an additive gaussian noise. We computed the Root Mean Square Error of

Table 1. Performances of DGD, ADGD, and warm-started Tikhonov regularization with accelerated proximal gradient descent. False positives are the selected irrelevant variables. False negatives are the discarded relevant features. Prediction error is the average prediction error of the estimated solution in percent. The results are averaged over 50 trials with the standard deviation between parentheses.

Noise	Relevant Variables	Algorithm	False Positive	False Negative	Prediction Error	Iterations
0.1	10	DGD	0.10 (0.3)	0.53 (0.7)	3.7 (0.6)	890 (200)
		ADGD	0.40 (0.9)	0.53 (0.7)	3.7 (0.6)	140 (30)
		Tikhonov	0.62 (1)	0.28 (0.5)	3.6 (0.3)	580 (40)
	30	DGD	8.8 (5)	1.8 (1)	4.8 (0.4)	860 (90)
		ADGD	5.0 (5)	1.8 (1)	4.6 (0.4)	110 (16)
		Tikhonov	12 (9)	2.1 (1)	5.4 (0.6)	860(140)
	60	DGD	49 (10)	5.2 (2)	8.1 (0.8)	940 (100)
		ADGD	27 (10)	5.7 (2)	7.4 (0.7)	170 (30)
		Tikhonov	53 (20)	5.7 (2)	7.4 (0.7)	1800 (400)
1	10	DGD	2.2 (3)	2.7 (1)	46 (2)	480 (100)
		ADGD	1.1 (2)	2.8 (1)	45 (2)	92 (40)
		Tikhonov	2.3 (2)	2.9 (1)	48 (4)	360 (90)
	30	DGD	17 (10)	14 (3)	65 (3)	560 (50)
		ADGD	14 (10)	15 (3)	64 (3)	220 (3)
		Tikhonov	8.0 (7)	15 (2)	63 (4)	990 (300)
	60	DGD	40 (10)	33 (4)	77 (3)	560 (10)
		ADGD	40 (20)	33 (6)	77 (3)	220 (3)
		Tikhonov	35 (30)	36 (8)	78 (2)	1700 (500)

the proposed approximation: $\mathrm{RMSE}(\widehat{\mathrm{W}}) = (\sum_{(i,j)\in A}(\widehat{W}_{i,j} - Y_{i,j})^2)^{1/2}/|A|$, where A is the test set. As can be seen in Table **??** ADGD is comparable to state of the art Tikhonov regularization, with a significantly lower computational cost. In addition, we compare DGD with Tikhonov regularization (with accelerated proximal gradient+warm restart) on the MovieLens 100k dataset[b]. We averaged our results over five trials. We left out one tenth of the known entries at each trial and chose the best step/parameter via 2-fold cross validation. The mean RMSE for DGD and Tikhonov was 1.02. It required 250 iterations on average using DGD, and 550 iterations using Tikhonov.

6.3. *Image deblurring*

Finally, we apply ADGD to an image processing problem, namely deblurring, with a strongly convex perturbation of total variation. More precisely, given an image $W \in \mathbb{R}^{256\times 256}$, we consider the regularization function $R(W) = TV(W)_{1,2} + \frac{3}{2}\|W\|^2$, where TV is the discrete total variation. In this application the proximity operator of the total variation penalty is not available in closed form. In our experiments, this is computed at each iteration using 20 steps of accelerated dual forward backward

[b] http://grouplens.org/datasets/movielens/

Table 2. Minimal achieved RMSE and associated cost of ADGD and Tikhonov approach solved with warm-starting, accelerated proximal gradient method, on simulated data with additive gaussian noise of standard deviation δ. We used the ground truth to select the best parameter. The percentage of known entries (knowledge ratio) is 0.12, 0.39 and 0.57 for, respectively, ranks 10, 50 and 100. The matrices are of size 1000x1000. The results were averaged over 5 simulations with the standard deviation between parentheses. For ADGD, the iterative trials were capped at 500 iterations (for noise levels of 0.01) and 250 (noise of 0.1 and 1).

Noise	Rank	RMSE ADGD	RMSE Tikhonov	Iterations ADGD	Iterations Tikhonov
0.01	10	$2.1 \cdot 10^{-3}(3.8 \cdot 10^{-5})$	$7.6 \cdot 10^{-3}(1.5 \cdot 10^{-4})$	500 (0)	527 (3)
	50	$3.2 \cdot 10^{-3}(2.2 \cdot 10^{-5})$	$9.1 \cdot 10^{-3}(4.1 \cdot 10^{-5})$	500 (0)	295 (1)
	100	$4.7 \cdot 10^{-3}(2.9 \cdot 10^{-5})$	$1.1 \cdot 10^{-2}(6.4 \cdot 10^{-5})$	500 (0)	273 (1)
0.1	10	$0.23(4.6 \cdot 10^{-3})$	$0.75(9.0 \cdot 10^{-3})$	250 (0)	539 (3.7)
	50	$0.35(2.1 \cdot 10^{-3})$	$0.95(2.7 \cdot 10^{-3})$	250 (0)	425 (0.49)
	100	$0.48(2.0 \cdot 10^{-3})$	$1.1(4.3 \cdot 10^{-3})$	190 (0)	470 (0.4)
1	10	27(0.28)	76(1.1)	191 (0)	698 (3.6)
	50	41(0.20)	108(0.20)	205 (0.4)	729 (3.6)
	100	55 (0.18)	125(0.11)	210 (0.4)	742 (2.8)

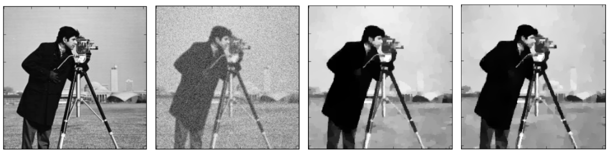

Fig. 1. From left to right: original Cameraman image, noisy blurred image, restored image with Tikhonov regularization, restored image with ADGD.

on the denoising problems corresponding to (**??**), and by warm starting with the previous approximate proximal point. We assume to have access to a noisy image $\widehat{y}$, obtained corrupting the original image with a Gaussian blur of one pixel and an additive Gaussian noise with variance 0.01. We compared the iterative regularization ADGD with early stopping with the solution obtained with the Tikhonov approach corresponding to the best regularization parameter on the cameramen image. The quality of an approximation of the original image is measured in terms of PSNR, and the best results are reported in Figure **??**. On the computational side, for the Tikhonov approach we set $\lambda_0 = 10^5$, and then decreased it by multiplying it by 0.8 at each step. The best solution is obtained for $\lambda = 6.8$, while iterative regularization achieves the best results at the third iteration.

Acknowledgments L. R. and S. V. acknowledge the financial support of the European Research Council (grant SLING 819789), the AFOSR ((European Office of Aerospace Research and Development)) project FA9550-18-1-7009 and FA8655-22-1-7034 , the EU H2020-MSCA-RISE project NoMADS - DLV-777826, the H2020-MSCA-ITN

Project Trade-OPT 2019; L. R. acknowledges the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and IIT. S.V. and C.M. are part of the Indam group "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro applicazioni".

References

[1] S. Arora, N. Cohen, W. Hu,and Y. Luo, Y., Implicit Regularization in Deep Matrix Factorization, In *Advances in Neural Information Processing Systems*, pp. 7413–7424, 2019.

[2] F. J. Aujol and C. Dossal, Stability of over-relaxations for the Forward-Backward algorithm, application to FISTA, *SIAM J. Optim.*, 25, pp. 2408–2433, 2015.

[3] M. Bachmayr and M. Burger, Iterative total variation schemes for nonlinear inverse problems, *Inverse Problems*, 25, pp. 105004, 2009

[4] H. Bausckhe and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[5] F. Bauer, S. Pereverzev, and L. Rosasco, On regularization algorithms in learning theory, *J. Complexity*, 23, pp.52–72, 2007.

[6] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Operations Research Letters*, 31, pp. 167–175, 2003.

[7] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.*, 2, pp. 183–202, 2009.

[8] A. Beck and M. Teboulle, A fast dual proximal gradient algorithm for convex minimization and applications, *Operations Research Letters*, 42, pp.1–6, 2014.

[9] S. Becker, J. Bobin, and E. Candès, NESTA: a fast and accurate first-order method for sparse recovery, *SIAM J. Imaging Sci.* 4, pp. 1–39, 2011.

[10] M. Benning and M. Burger, Modern regularization methods for inverse problems, *Acta Numerica*, 27, pp. 1–111, 2018.

[11] G.Blanchard and N.Krämer, *Optimal learning rates for kernel conjugate gradient regression.* In *Advances in Neural Inf. Proc. Systems*, pp. 226–234, 2010.

[12] S. Bonettini, Inexact block coordinate descent methods with application to non-negative matrix factorization, *IMA J. Numer. Anal.*, 31, pp. 1413–1452, 2011.

[13] R. Bot and T. Hein, Iterative regularization with a general penalty term-theory and application to L^1 and TV regularization, *Inverse Problems*, 28, 104010 (19pp), 2012.

[14] L. Bottou and O. Bousquet, The tradeoffs of large scale learning. In *Optimization for Machine Learning*, pp. 351–368, MIT Press, 2011.

[15] P. Brianzi, F. Di Benedetto and C. Estatico, Preconditioned Iterative regularization in Banach Spaces, *Comput. Optim. Appl.* 54, pp. 263–282, 2013.

[16] M. Burger and S. Osher, *A Guide to the TV Zoo.* Lecture Notes in Mathematics 2090, pp. 1-70, 2013.

[17] M. Burger, E. Resmerita, and L. He, Error estimation for Bregman iterations and inverse scale space methods in image restoration, *Computing*, 81 pp. 109-135, 2007.

[18] M. Burger, A. Sawatzky and G. Steidl, *First Order Algorithms in Variational Image Processing.* In *Splitting Methods in Communication, Imaging, Science, and Engineering* pp. 345?407, 2017

[19] J.-F. Cai, E. Candès, and Z. Shen, A Singular Value Thresholding Algorithm for Matrix Completion, *SIAM J. Optim.*, 20, pp. 1956–1982, 2010.

[20] E. Candès and Y. Plan, Matrix Completion with Noise, *Proceedings of the IEEE*, 98, pp. 925–936, 2010.

[21] A. Caponnetto and E. De Vito, Optimal rates for regularized least-squares algorithm,

Found. Comput. Math., 7, pp. 331–368, 2007.

[22] P. L. Combettes, Quasi-Fejérian analysis of some optimization algorithms, in *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, pp. 115–152, Elsevier, New York, 2001.

[23] P. L. Combettes, D. Dũng, and B. C. Vũ, Dualization of signal recovery problems, *Set-Valued Var. Anal.*, 18, pp. 373–404, 2010.

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

[25] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, *Math. Program.* 146, pp. 37–75, 2014.

[26] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, *Annals Stat.*, 32, pp. 407–499, 2004.

[27] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[28] G. Garrigos, L. Rosasco, and S. Villa, Iterative Regularization via Dual Diagonal Descent, *J. Math. Imaging Vision*, 60, pp. 189–215, 2018.

[29] L. Calatroni, G. Garrigos, L. Rosasco, and S. Villa, Accelerated iIterative Regularization via Dual Diagonal Descent, *SIAM J. Optim.*, 31, pp. 754–784, 2021.

[30] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, Characterizing Implicit Bias in Terms of Optimization Geometry, In *Conference Proceedings of the 35th International Conference on Machine Learning*, 80 pp. 1832–1841, 2018.

[31] S. Gunasekar, B.E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, Implicit regularization in matrix factorization, in *Advances in Neural Information Processing Systems*, 30 pp. 6151–6159, 2017.

[32] D. Soudry, E. Hoffer, M. Shpigel Nacson, S. Gunasekar, and N. Srebro, The Implicit Bias of Gradient Descent on Separable Data *J. Mach. Learn. Res.* 79, pp.1–57, 2018.

[33] E. Hale, W. Yin, and Y. Zhang, Fixed-Point Continuation for ℓ_1-Minimization: Methodology and Convergence, *SIAM J. Optim.*, pp. 1107–1130, 2008.

[34] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, The entire regularization path for support vector machine, *J. Mach. Learn. Res.*, 5, pp. 1391–1415, 2004.

[35] S. Kale, A. Sekhari, K. Sridharan, SGD: The role of implicit regularization, batch-size and multiple-epochs In *Advances in Neural Information Processing Systems*, 34, 2021.

[36] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, De Gruyter, Berlin, New York, 2008.

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

[38] C. Molinari, M. Massias, L. Rosasco, and S. Villa, Iterative regularization for convex regularizers, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR* 130, pp.1684–1692, 2021.

[39] C. Molinari, M. Massias, L. Rosasco, and S. Villa, Iterative regularization for low complexity regularizers, *arXiv*, 2022.

[40] Y. Nesterov, *Introductory Lectures on Convex Optimization. A basic course.* Springer, New York, 2004.

[41] Y. Nesterov, Gradient methods for minimizing composite objective function *CORE Discussion Paper* 2007/76, Catholic University of Louvain, 2007.

[42] A. Neubauer, On Nesterov acceleration for Landweber iteration of linear ill-posed problems, *J. Inverse Ill-posed Problems*, 25, pp. 381–390, 2017.

[43] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, Geometry of optimiza-

tion and implicit regularization in deep learning, *ArXiv 1705.03071*, 2017

[44] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration, *Multiscale Modeling and Simulation*, 4, pp. 460–489, 2005.

[45] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, Sparse Recovery via Differential Inclusions, *Appl. Comput. Harmonic Anal.*, 41, pp. 436–469, 2016.

[46] S. Oymak, M. Soltanolkotabi, and B. Recht, Sharp Time-Data Tradeoffs for Linear Inverse Problems, *IEEE Trans. Inf. Theory*, 64 pp.4129–4158, 2018

[47] B. Polyak, *Introduction to Optimization,* Optimization Software, Inc., New York, 1987.

[48] G. Raskutti, M. Wainwright, and B. Yu, Early Stopping and Non-parametric Regression: An optimal data-dependent stopping rule, *J. Mach. Learn. Res.*, 15, pp. 283–314, 2014.

[49] T. Rockafellar, Monotone operators and the proximal point algorithm *SIAM J. Control Optim.* 14, pp. 877–898, 1976.

[50] L. Rosasco and S. Villa, Learning with incremental iterative regularization, in *Advances in Neural Information Processing Systems* 28, pp. 1630–1638, 2015.

[51] S. Salzo and S. Villa, Accelerated and inexact proximal point algorithm, *J. Convex Anal.*, 19, pp. 1167–1192, 2012.

[52] S. Salzo and S. Villa, Proximal Gradient Methods for Machine Learning and Imaging, in *Harmonic and Applied Analysis. From Radon Transforms to Machine Learning*, pp. 149–244, 2021.

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

[54] M. Schmidt, N. Le Roux, and F. Bach, Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization, in *Advances in Neural Information Processing Systems*, 24, pp. 1458–1466, 2011.

[55] D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, and N. Srebro, The implicit bias of gradient descent on separable data, *J. Mach. Learn. Res.* 19, pp.2822–2878, 2018.

[56] T. Vaskevicius, TV. Kanade, and P. Rebeschini, Implicit Regularization for Optimal Sparse Recovery. In *Advances in Neural Information Processing Systems*, 33, pp. 2972–2983, 2019.

[57] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, Accelerated and inexact forward-backward algorithms, *SIAM J. Optim.*, 23, pp. 1607–1633, 2013.

[58] Y. Yao, L. Rosasco, and A. Caponnetto, On early stopping in gradient descent learning, *Constructive Approximation*, 26, pp. 289–315, 2007.

[59] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, Bregman Iterative Algorithms for ℓ^1-Minimization with Applications to Compressed Sensing, *SIAM J. Imaging Sciences*, 1, pp. 143–168, 2008.

[60] C. Wu and X.-C. Tai, Augmented Lagrangian Method, Dual Methods, and Split Bregman Iteration for ROF, Vectorial TV, and High Order Models, *SIAM J. Imaging Sciences*, 3, pp.300–339, 2010.

[61] T. Zhang and B. Yu, Boosting with early stopping: Convergence and consistency, *Annals Stat.*, 33, pp. 1538–1579, 2005.

[62] X. Zhang, M. Burger, X. Bresson, and S. Osher, Bregmanized Nonlocal Regularization for Deconvolution and Sparse Reconstruction, *SIAM J. Imaging Sciences*, 3, pp. 253–276, 2010

Appendix A. Useful results from convex optimization

Here we report a number of both classical and recent previous results used in our analysis. All the proof is based on duality techniques. We recall the classical definition of Fenchel conjugate of a convex function.

Definition Appendix A.1. Let $\mathcal{X}$ be a Hilbert space and let $f\colon X \to \mathbb{R} \cup \{+\infty\}$ be a proper, and lower semicontinuous function. The Fenchel conjugate of f is the function $f^*\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ defined by

$$(\forall u \in \mathcal{X}) \quad f^*(u) = \sup_{x \in \mathcal{X}} \langle x, u \rangle - f(x).$$

It is well known that f^* is a convex, proper and lower semicontinuous function.

Proposition Appendix A.1. *Let $\mathcal{X}$ be a Hilbert spaces, let $h\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex, and lower semicontinuous function, and let $\sigma > 0$. Define $f = h + (\sigma/2)\|\cdot\|^2$. Then, f^* is differentiable and ∇f^* is σ^{-1} Lipschitz continuous. In addition, $\nabla f^*(v) = \operatorname{prox}_{\sigma^{-1}h}(v/\sigma)$ for all $v \in \mathcal{X}$.*

Proof. Let $v \in \mathcal{X}$. Proposition 12.29 in [**?**] yields

$$\begin{aligned}\nabla R^*(v) &= \alpha^{-1}v - \alpha^{-1}\nabla({}^{\alpha^{-1}}F)(\alpha^{-1}v)\\ &= \alpha^{-1}v - \alpha^{-1}\Big(\alpha\big(\alpha^{-1}v - \operatorname{prox}_{\alpha^{-1}F}(\alpha^{-1}v)\big)\Big)\\ &= \operatorname{prox}_{\alpha^{-1}F}(\alpha^{-1}v). \end{aligned} \tag{A.1} \square$$

Definition Appendix A.2. Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces and let $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous functions. Let $A\colon \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. The Fenchel-Rockafellar dual of the problem $\inf(f + g \circ A)$ is the problem

$$\min_{u \in \mathcal{X}} f^*(-A^*u) + g^*(u).$$

The next result establishes a relation between the dual objective value and the distance from the unique solution of the primal problem in the strongly convex case. Its proof can be found in [**?**, Lemma 5].

Lemma Appendix A.1 (Primal-dual values-iterates bound). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces, let $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a proper, σ-strongly convex, and lower semicontinuous function, let $g\colon \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function, and let $A\colon \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. Let $x^\dagger$ be the unique minimizer of $p := f + g \circ A$, and let $d := f^* \circ (-A^*) + g^*$. Then*

$$\operatorname{argmin} d \neq \varnothing \iff 0 \in \partial f(x^\dagger) + A^*\partial g(Ax^\dagger).$$

In that case, for every $u \in \mathcal{Y}$ and every $x := \nabla f^*(-A^*u)$, we have

$$\frac{\sigma}{2}\|x - x^\dagger\|^2 \leq d(u) - \min_{\mathcal{Y}} d.$$

The following is a classical result about a crucial property of the proximity operator. Its proof is in [**?**, Propositon 12.28].

Proposition Appendix A.2 (Firm nonexpansiveness of prox). *Let $\mathcal{X}$ be a Hilbert space, let $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex, and lower semicontinuous function and let $\sigma > 0$. Then* $\operatorname{prox}_{\sigma f}\colon \mathcal{X} \to \mathcal{X}$ *is firmly nonexpansive, namely*

$$(\forall (w,u) \in \mathcal{X}^2) \qquad \|\operatorname{prox}_{\sigma f}(w) - \operatorname{prox}_{\sigma f}(u)\|^2 \leq \langle w - u, \operatorname{prox}_{\sigma f}(w) - \operatorname{prox}_{\sigma f}(u)\rangle.$$