Iterative regularization with convex regularizers

Cesare Molinari
Istituto Italiano di Tecnologia

Mathurin Massias
MaLGa, DIBRIS
University of Genova

Lorenzo Rosasco
MaLGa, DIBRIS
University of Genova
MIT

Silvia Villa
MaLGa, DIMA
University of Genova

Abstract

We study iterative/implicit regularization for linear models, when the bias is convex but not necessarily strongly convex. We characterize the stability properties of a primal-dual gradient based approach, analyzing its convergence in the presence of worst case deterministic noise. As a main example, we specialize and illustrate the results for the problem of robust sparse recovery. Key to our analysis is a combination of ideas from regularization theory and optimization in the presence of errors. Theoretical results are complemented by experiments showing that state-of-the-art performances can be achieved with considerable computational speed-ups.

1 Introduction

Machine learning often reduces to estimating some model parameters. This approach raises at least two orders of questions: first, multiple solutions may exist, amongst which a specific one must be selected; second, potential instabilities with respect to noise and sampling must be controlled.
A classical way to achieve both goals is to consider explicitly penalized or constrained objective functions. In machine learning, this leads to regularized empirical risk minimization (Shalev-Shwartz and Ben-David, 2014). A more recent approach is based on directly exploiting an iterative optimization procedure for an unconstrained/unpenalized problem.

This approach is shared by several related ideas. One is implicit regularization (Mahoney, 2012; Gunasekar et al., 2017), stemming from the observation that the bias is controlled increasing the number of iterations, just like in penalized methods it is controlled decreasing the penalty parameter. Another one is early stopping (Yao et al., 2007; Raskutti et al., 2014), putting emphasis on the fact that running the iterates to convergence might lead to instabilities in the presence of noise. Yet another, and more classical, idea is iterative regularization, where both aspects (convergence and stability) are considered to be relevant (Engl et al., 1996; Kaltenbacher et al., 2008). This approach naturally blends modeling and numerical aspects, often improving computational efficiency, while retaining good prediction accuracy (Yao et al., 2007). Another reason of interest is that iterative regularization may be one of the mechanisms explaining generalization in deep learning (Neyshabur et al., 2017; Gunasekar et al., 2017; Arora et al., 2019; Vaškevičius et al., 2020).

A classic illustrative example is gradient descent for linear least squares. The latter, if suitably initialized, converges (is biased) to the minimal Euclidean norm solution. Moreover, its stability is controlled along the iterative process, allowing to derive early stopping criterions depending on the noise (Engl et al., 1996; Raskutti et al., 2014). There are a number of developments of these basic results. For example, one line of work has considered extensions to other gradient-based methods, such as stochastic and accelerated gradient descent (Zhang and Yu, 2005; Moulines and Bach, 2011; Rosasco and Villa, 2015; Pagliana and Rosasco, 2019). Another line of work has considered classification problems (Gunasekar et al., 2017; Soudry et al., 2018) and also nonlinear models, such as deep networks (Neyshabur et al., 2017), see also Kaltenbacher et al. (2008) for results for non linear inverse problems.

In this work, we are interested in iterative regularization procedures where the considered bias is not the Euclidean norm but rather a general convex functional. The question is to determine whether or not there exists an iteration analogous to gradient descent for such general bias. This question has been stud-

Proceedings of the 24^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s). Contact: cecio.molinari@gmail.com

ied when the bias is *strongly* convex. In this case, linearized Bregman iterations (a.k.a. mirror descent) can be used (Burger et al., 2007; Gunasekar et al., 2018). For this approach accelerated algorithms (Matet et al., 2017) have also been considered and studied. Finally, approaches have also been studied based on diagonal methods (Garrigos et al., 2018) and their acceleration (Calatroni et al., 2019). The general convex case, even for linear models, is much less understood. There have been studies for ADMM/Bregman iteration, but the procedure requires solving a nontrivial optimization problem at each iteration (Burger et al., 2007). Further, stability and convergence results are proved only in terms of Bregman divergence, which in general is a weak result. Interestingly, various recent results study iterative regularization for sparse recovery, where the bias is defined by an ℓ_1 norm (Agarwal et al., 2012; Osher et al., 2016; Vaškevičius et al., 2019).

In this paper we propose and study an efficient algorithm for general convex bias beyond ℓ_1 norm. Indeed, we adapt the Chambolle and Pock (CP) algorithm, popular in imaging (Chambolle and Pock, 2011), and study its iterative regularization properties. The CP algorithm is a first order primal-dual method, thus easy to implement and requiring only matrix-vector multiplications and proximity operators. In the setting of linear models with worst case errors, our analysis provides dimension free convergence and stability results in terms of both Bregman divergence and approximate feasibility. A combination of these two results allows to derive strong convergence results in the ℓ_1 norm case. The proof relies on results from the analysis of primal-dual methods with errors (Rasch and Chambolle, 2020). From our general results, several special cases can be derived and we discuss as an example sparse recovery, proving dimension free estimates in norm. In the experimental section, we investigate the proposed method and show state-of-the-art performances with significant computation savings compared to the Tikhonov approach.

Notation, The set of integers from 1 to n is $[n]$. Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $J\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous. The subdifferential of J at $w \in \mathbb{R}^p$ is $\partial J(w)$. The Bregman divergence associated to J is denoted $D_J^\theta(w, w') := J(w) - J(w') - \langle \theta, w - w' \rangle$, where $\theta \in \partial J(w')$. The Fenchel-Legendre conjugate of f is $f^\star(\theta) := \sup_w \langle w, \theta \rangle - f(w)$. The indicator function $\iota_{\{\mathbf{y}\}}$ is equal to zero if the argument equals $\mathbf{y}$ and $+\infty$ otherwise. For a proper convex lower semicontinuous function J, $\operatorname{prox}_J(x) = \arg\min_y J(y) + \|x - y\|^2 / 2$.

2 Over-parametrization, implicit and explicit regularization

The basic problem of supervised learning is to find a relationship to predict outputs y from inputs x,

$$x \mapsto f(x) \approx y \ ,$$

given a limited number of pairs $(x_i, y_i)_{i=1}^n$ with, e.g. $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The search for a solution is typically restricted to a set of parametrized functions f_w, with $w \in \mathbb{R}^p$. A prototype example are linear models where $p = d$ and $f_w(x) = \langle w, x \rangle$, or more generally $f_w(x) = \sum_{j=1}^p w^j \phi_j(x)$, for some dictionary $\phi_j : \mathbb{R}^d \to \mathbb{R}$, $\ j = 1, \ldots, p$ (Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014). In modern applications, it is often the case that the number of parameters p is vastly larger than the number of available data points n, a regime called over-parametrized. Excluding degenerate cases, one can then expect to find a solution w capable of interpolating the data, that is satisfying,

$$f_w(x_i) = y_i, \qquad \forall i \in [n] \ . \tag{1}$$

In the sequel we consider the case of a linear $f_w(x) = \langle w, x \rangle$. A popular method to find a solution to (1) is gradient descent on least squares, also called Landweber iteration:

$$w_k = w_{k-1} - \gamma \mathbf{X}^\top (\mathbf{X} w_{k-1} - \mathbf{y}) \ , \tag{2}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the data matrix and the outputs vector, respectively (see Section 3 for more details). It is well known (Engl et al., 1996) that, if initialized at $w_0 = 0$ and with $\gamma < 2 / \|\mathbf{X}\|_{\mathrm{op}}^2$, the iterations of gradient descent converge to a specific solution, namely

$$\operatorname*{arg\,min}_{w \in \mathbb{R}^p} \|w\| \quad \text{s.t.} \quad \mathbf{X} w = \mathbf{y} \ . \tag{3}$$

This means that amongst all solutions, the algorithm is implicitly *biased* towards that with small norm. The bias is implicit in the sense that there is no explicit penalization or constraint in the iterations (2). This approach can be contrasted to explicit penalization (Tikhonov regularization),

$$w^{(\lambda)} = \operatorname*{arg\,min}_{w \in \mathbb{R}^p} \lambda \|w\|^2 + \|\mathbf{y} - \mathbf{X} w\|^2 \ , \tag{4}$$

where the minimal norm solution (3) is obtained for λ going to zero. It is well known that for Tikhonov regularization larger values of λ improve stability. Interestingly, the same effect can also be achieved with gradient descent (2), by *not* running the iterations until convergence, a technique often referred to as early stopping (Engl et al., 1996; Yao et al., 2007). In this view the number of iterations k plays the role of a regularization parameter just like λ in Tikhonov regularization (or rather $1/\lambda$). Iterative regularization is particularly appealing in the large scale setting, where substantial computational savings are expected: early stopping needs a finite number of iterations (2), while Tikhonov regularization requires solving exactly Problem (4) for multiple values of λ.

It is natural to ask whether the above iterative regularization scheme applies to biases beyond the Euclidean norm. For a strongly convex J, an answer is given by considering the mirror descent algorithm (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003) with respect to the Bregman divergence induced by J. The bias J is not used to define an explicit penalization of an empirical risk, but it appears in the mirror descent algorithm, and in this sense is "less implicit". The results in Benning et al. (2016) and Gunasekar et al. (2018) show that mirror descent is implicitly biased towards the solution of the following problem

$$\min_{w\in\mathbb{R}^p} J(w) \quad \text{s.t.} \quad \mathbf{X}w = \mathbf{y} \ , \tag{5}$$

and exhibit similar regularization and stability properties to the one of the gradient descent algorithm. In both Benning et al. (2016) and Gunasekar et al. (2018), the key technical assumption is *strong convexity* of J leaving open the question of dealing with biases that are only convex. In this paper, we take steps to fill in this gap studying an efficient approach for which we characterize the iterative regularization properties.

3 Problem setting and proposed algorithm

We begin describing the algorithm we consider and its derivation. We first set some notation and the main assumption. In the following, $\mathbf{X}$ is an n by p matrix and $\mathbf{y}$ an n-dimensional vector. We consider the case where $\mathbf{y}$ is unknown, and a vector $\mathbf{y}^\delta$ is available such that $\|\mathbf{y}-\mathbf{y}^\delta\| \leqslant \delta$, where $\delta \geqslant 0$ can be interpreted as the noise level.

Assumption 1. *We assume that the bias $J : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, and lower semicontinuous. We also assume that Problem (5) has at least one solution (in particular the linear equation has at least one solution for the exact data $\mathbf{y}$).*

The latter is satisfied, for instance, if J is coercive, $n \leqslant p$ and $\mathbf{X}$ has rank n. Then Problem (5) is feasible and has a solution for every $\mathbf{y}$. Our assumption is weaker: we do not even require feasability of the noisy constraint $\mathbf{X}w = \mathbf{y}^\delta$.

Note that we use a vectorial notation for simplicity but our results are dimension free and sharp for an infinite dimensional setting where $\mathbf{X}$ is a linear bounded operator between separable Hilbert spaces.

3.1 Proposed algorithm

Consider the following iterations, with initialization $w_0 \in \mathbb{R}^p, \theta_0 = \theta_{-1} \in \mathbb{R}^n$, and parameters τ, σ such that $\sigma\tau\|\mathbf{X}\|_{\text{op}}^2 < 1$:

$$\begin{cases} w_{k+1} = \operatorname{prox}_{\tau J}(w_k - \tau\mathbf{X}^\top(2\theta_k - \theta_{k-1})) \ , \\ \theta_{k+1} = \theta_k + \sigma(\mathbf{X}w_{k+1} - \mathbf{y}) \ . \end{cases} \tag{6}$$

If $\theta_0 = 0$, since $\theta_k = \sigma\sum_{i=1}^k(\mathbf{X}w_i - \mathbf{y})$, this algorithm can be rewritten without θ_k:

$$w_{k+1} = \operatorname{prox}_{\tau J}\Big(w_k - \tau\sigma\mathbf{X}^\top\big(\textstyle\sum_{i=1}^k(\mathbf{X}w_i - \mathbf{y}) + \mathbf{X}w_k - \mathbf{y}\big)\Big).$$

In terms of computations the algorithm (6) is very similar to the forward-backward/proximal gradient algorithm (Combettes and Wajs, 2005). The difference is that the gradient term is here replaced by the sum of past gradients. We instantiate algorithm (6) for two popular choices of J. For $J = \|\cdot\|^2$, the updates read:

$$\begin{cases} w_{k+1} = \frac{1}{1+\tau}(w_k - \tau\mathbf{X}^\top(2\theta_k - \theta_{k-1})) \ , \\ \theta_{k+1} = \theta_k + \sigma(\mathbf{X}w_{k+1} - \mathbf{y}) \ . \end{cases}$$

Notice that, though involving very similar computations, the algorithm does not reduce to gradient descent iterations (2).

For $J = \|\cdot\|_1$, denoting by $\mathrm{ST}(\cdot,\tau)$ the soft-thresholding operator of parameter τ, the iterations (6) read:

$$\begin{cases} w_{k+1} = \mathrm{ST}(w_k - \tau\mathbf{X}^\top(2\theta_k - \theta_{k-1}), \tau) \ , \\ \theta_{k+1} = \theta_k + \sigma(\mathbf{X}w_{k+1} - \mathbf{y}) \ . \end{cases}$$

Also in this case, it is similar – yet not equivalent – to a popular algorithm to solve the Tikhonov problem: the Iterative Soft-Thresholding Algorithm (Daubechies et al., 2004).

Proposition 1. *Under Assumption 1, the iterations (6) converge to a point $(w^\star, \theta^\star)$ such that $\mathbf{X}w^\star = \mathbf{y}$. Additionally, $w^\star$ is a minimizer of J amongst all interpolating solutions, meaning that it solves Problem (5).*

Notice that for Landweber iteration, converging to the minimal norm solution (3) requires $w_0 = 0$. Our procedure has no such requirement: for any initialization, the iterates w_k converge to a solution of Problem (5). We now provide some examples of the above setting.

Example 2 (Sparse recovery). *Choosing $J = \|\cdot\|_1$ corresponds to finding the minimal ℓ_1-norm solution to a linear system, and in this case Problem (5) is known as Basis Pursuit (Chen et al., 1998). The relaxed approach of Problem (4) in this case yields the Lasso (Tibshirani, 1996). ℓ_1-based approaches have had a tremendous impact in imaging, signal processing and machine learning in the last decades (Hastie et al., 2015).*

Example 3 (Low rank matrix completion). *In several applications, such as recommendation systems, it is useful to recover a low rank matrix, starting from the observation of a subset of its entries (Candès and Recht, 2009). A convex formulation is:*

$$\min_{W \in \mathbb{R}^{p_1 \times p_2}} \|W\|_* \quad \textit{s.t. } W_{ij} = Y_{ij} \quad \forall (i,j) \in \mathcal{D} \ , \tag{7}$$

where $\|\cdot\|_*$ is the nuclear norm and $\mathcal{D} \subset [p_1] \times [p_2]$ is the set of observed entries of the matrix Y. In that case, $\mathbf{X}$ is not a design matrix: it is a (self-adjoint and linear) masking operator from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^{p_1 \times p_2}$, such that $(XW)_{ij}$ has value W_{ij} if $(i,j) \in \mathcal{D}$ and 0 otherwise; the constraints write $\mathbf{X}W = \mathbf{X}Y$.

Example 4 (Total variation). *In many imaging tasks such as deblurring and denoising, regularization through total variation allows to simultaneously preserve edges whilst removing noise in flat regions Rudin et al. (1992). The problem of Total Variation is $\min_{W \in \mathbb{R}^{p_1 \times p_2}} \|\nabla W\|_1$ s.t. $\mathbf{X}W = Y$, where $\mathbf{X}$ is usually a blurring operator. The problem can be reformulated as: $\min_{\tilde{W}} \Omega(\tilde{W})$ s.t. $\tilde{\mathbf{X}}\tilde{W} = \tilde{Y}$, with $\tilde{W} = \begin{pmatrix} W \\ U \end{pmatrix}$, $\Omega(\tilde{W}) = \|U\|_1$, $\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & 0 \\ \nabla & -\mathrm{Id} \end{pmatrix}$ and $\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$.*

Other examples include ℓ_∞ norm for antiparsimony, ℓ_1/ℓ_2 mixed-norm for group sparsity, or Kullback-Leibler divergence. Finally, for any linear operator $\mathbf{K}$ and convex function $\tilde{J}$ such that $\mathrm{prox}_{\tilde{J}}$ is available, we can handle problems of the form

$$\min_w \tilde{J}(\mathbf{K}w) \qquad \text{s.t.} \qquad \mathbf{X}w = \mathbf{y} \ .$$

Indeed, we can introduce an auxiliary variable as in Example 4, and avoid the computation of $\mathrm{prox}_{\tilde{J} \circ \mathbf{K}}$.

3.2 Chambolle-Pock algorithm

In this section we prove Proposition 1 by casting (6) as an instance of the Chambolle-Pock algorithm (Chambolle and Pock, 2011) which solves:

$$\min_w f(\mathbf{X}w) + g(w) \ .$$

Hence, for $f = \iota_{\{\mathbf{y}\}}$ and $g = J$, it can minimize a convex function on a set defined by linear equalities, as in Problem (5). For this choice of f and g, it instantiates as (6) (see Appendix B.2).

Amongst other assets, algorithm (6) only involves matrix-vector multiplications, and the computation of prox_J, available in closed-form in many cases. The only tunable parameters are two step-sizes, τ and σ, which are easy to set. As usual for this class of methods, called primal-dual, the Lagrangian is a useful tool to establish convergence results. The Lagrangian of Problem (5) is

$$\mathcal{L}(w, \theta) = J(w) + \langle \theta, \mathbf{X}w - y \rangle \ , \tag{8}$$

where $\theta \in \mathbb{R}^n$ is the dual variable. Under a technical condition (Appendix B.1), $w^\star$ is a solution of Problem (5) if and only if there exists a dual variable $\theta^\star$ such that $(w^\star, \theta^\star)$ is a saddle-point for the Lagrangian, namely, iff for every $(w, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$,

$$\mathcal{L}(w^\star, \theta) \leqslant \mathcal{L}(w^\star, \theta^\star) \leqslant \mathcal{L}(w, \theta^\star) \ . \tag{9}$$

The variable θ is in our setting just an auxiliary variable, and we will be interested in convergence properties of w_k towards $w^\star$.

Other algorithms As mentioned in the introduction, other algorithms could be considered, e.g. ADMM/Bregman iteration. However, we are not aware of methods that can be efficiently implemented in our general setting. In Appendix A. we provide an extensive review discussing the connection with a number of different approaches and related works.

4 Theoretical analysis

In this section, we analyze the convergence properties of Algorithm (6). First, we need to choose a suitable criterion to estimate the approximation properties of the iterates. In general, it is not reasonable to expect a rate of convergence for the distance between the iterates and the solution. Indeed, since the problem is only convex, it is well known that the convergence in distance can be arbitrarily slow. In Section 4.1, we explain why a reasonable choice is given by the duality gap together with the residual norm (respectively, $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ and $\|\mathbf{X}w_k - \mathbf{y}\|$). For these two quantities, we derive:

- convergence rates in the exact case, i.e. when the data $\mathbf{y}$ is available (Proposition 6);
- early-stopping bounds in the inexact case, i.e. when the accessible data is only $\mathbf{y}^\delta$ with $\|\mathbf{y}^\delta - \mathbf{y}\| \leqslant \delta$ (Proposition 7 and Corollary 8).

In Section 4.4, we apply our analysis to the specific choice of J equal to the ℓ_1-norm. In this particular case, combining the previous results, we even obtain bounds directly on $\|w_k - w^\star\|$.

4.1 Measure of optimality

To discuss which criterion is significant to study the algorithm convergence, we recall from (9) that, if

$$\mathcal{L}(w', \theta) - \mathcal{L}(w, \theta') \leqslant 0 \tag{10}$$

for every $(w, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$, then (w', θ') is a primal-dual solution. In general, it is difficult to prove that Equation (10) holds for every $(w, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$. Then, given a saddle-point $(w^\star, \theta^\star)$ and a generic $(w', \theta') \in \mathbb{R}^p \times \mathbb{R}^n$, it is popular to consider the quantity

$$\mathcal{L}(w', \theta^\star) - \mathcal{L}(w^\star, \theta') \geqslant 0 \ . \tag{11}$$

To establish the optimality of (w', θ'), it is not enough to ensure $\mathcal{L}(w', \theta^\star) - \mathcal{L}(w^\star, \theta') = 0$. Lemma 5, proved in Appendix C.1, shows that this condition, when coupled with $\mathbf{X}w' = \mathbf{y}$, implies that w' is a solution of Pb (5).

Lemma 5. *Let $(w^\star, \theta^\star)$ be a primal-dual solution and (w', θ') a point in $\mathbb{R}^p \times \mathbb{R}^n$ such that $\mathcal{L}(w', \theta^\star) - \mathcal{L}(w^\star, \theta') = 0$ and $\mathbf{X}w' = \mathbf{y}$. Then $(w', \theta^\star)$ is a primal-dual solution.*

Thus, the quantities $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ and $\|\mathbf{X}w_k - \mathbf{y}\|$, studied together, are a reasonable measure of optimality for the iterate w_k.

Note that $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ is the error measure used in a series of papers dealing with regularization of inverse problems with general convex regularizers, see e.g. Burger et al. (2007). It is well known that if J is strongly convex then this quantity controls the distance in norm (Remark 9) and therefore is a proper measure of convergence. If J is only convex, this measure of error can be quite weak. In Section 4.4 we point out the limitations of this quantity when dealing with $J = \|\cdot\|_1$. For this choice of J, $\mathcal{L}(0, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ is 0 for any θ_k; as shown on Figure 1, this quantity is also 0 when w_k and $w^\star$ have the same support and sign.

4.2 Exact case

First consider the iterates (w_k, θ_k) obtained by applying iterations (6) to the exact problem, namely where the data $\mathbf{y}$ is available. Let $(w^\star, \theta^\star)$ be a saddle-point for the Lagrangian. Denoting the primal-dual variables by $z = (w, \theta)$, we have $z_k = (w_k, \theta_k)$ for the iterates of the algorithm and $z^\star = (w^\star, \theta^\star)$ for the saddle-point. For τ and $\sigma > 0$, define V as the following square weighted norm on $\mathbb{R}^p \times \mathbb{R}^n$:

$$V(z) := \frac{1}{2\tau} \|w\|^2 + \frac{1}{2\sigma} \|\theta\|^2 . \tag{12}$$

For the averaged iterates $\overline{w}^k := \frac{1}{k} \sum_{t=1}^k w_t$ and $\overline{\theta}^k := \frac{1}{k} \sum_{t=1}^k \theta_t$, we have the following rates.

Proposition 6 (Convergence rates). *Under Assumption 1, let $\varepsilon \in (0,1)$ and assume that the step-sizes are such that $\sigma\tau \leqslant \varepsilon / \|\mathbf{X}\|_{\mathrm{op}}^2$. Then*

$$\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k) \leqslant \frac{V(z_0 - z^\star)}{k} \qquad \textit{and}$$
$$\left\|\mathbf{X}\overline{w}^k - \mathbf{y}\right\|^2 \leqslant \frac{2(1+\varepsilon)V(z_0 - z^\star)}{\sigma\varepsilon(1-\varepsilon)k} \ .$$

The first result is classical (see Chambolle and Pock (2011)). Alternatively, it can be obtained by setting $\delta = 0$ in Proposition 7, where we study the more general inexact case. To the best of our knowledge, the second bound is new and can also be derived by setting $\delta = 0$ in Proposition 7. A similar result, in the more specific case of primal-dual coordinate descent, can be found in Fercoq and Bianchi (2019). Note that both results of Proposition 6 are true for every primal-dual solution. On the other hand, the left-hand-side in the second equation does not depend on the selection of $z^\star$ and so the bound can be improved by taking the inf over all primal-dual solutions.

4.3 Inexact case

We now consider the iterates (w_k, θ_k), and their averaged versions $(\overline{w}_k, \overline{\theta}_k)$, obtained by applying iterations (6) to the noisy problem, where $\mathbf{y}$ is replaced by $\mathbf{y}^\delta$ with $\left\|\mathbf{y}^\delta - \mathbf{y}\right\| \leqslant \delta$. In Proposition 7, we derive early-stopping bounds for the iterates, in terms of duality gap $\mathcal{L}(w_k, \theta^\star) - \mathcal{L}(w^\star, \theta_k)$ and residual norm $\|\mathbf{X}w_k - \mathbf{y}\|$. We highlight that, despite the error in the data $\mathbf{y}^\delta$, both quantities are defined in terms of $\mathbf{y}$ and hence related to the *noiseless* problem. In particular, $(w^\star, \theta^\star)$ is a saddle-point for the *noiseless* Lagrangian. We have the following estimates, whose proofs are given in Appendix C.3.

Proposition 7 (Stability). *Under Assumption 1, let $\varepsilon \in (0,1)$ and assume that the step-sizes are such that $\sigma\tau \leqslant \varepsilon / \|\mathbf{X}\|_{\mathrm{op}}^2$. Then,*

$$\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k) \leqslant \tfrac{1}{k} \left(\sqrt{V(z_0 - z^\star)} + \sqrt{2\sigma}\delta k\right)^2 \tag{13}$$

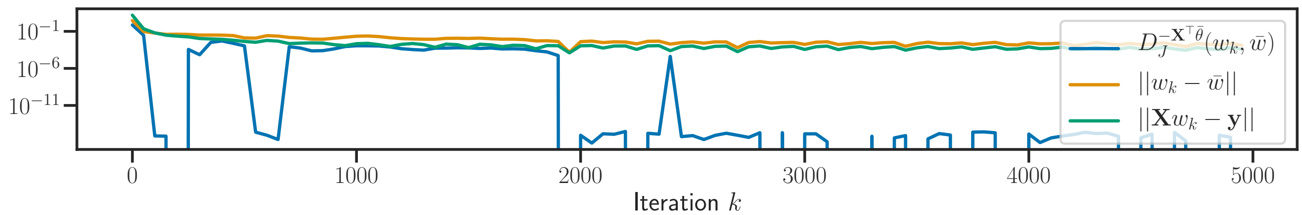

Figure 1: For $J = \|\cdot\|_1$, convergence of primal-dual iterates w_k towards $w^\star$, measured in norm, feasability and Bregman divergence. The Bregman divergence quickly vanishes up to numerical errors (the iterates have the same sign as the solution); yet the iterates are still far from the solution.

and

$$\begin{aligned}\left\|\mathbf{X}\overline{w}^k - \mathbf{y}\right\|^2 \leqslant \frac{2(1+\varepsilon)}{\sigma\varepsilon(1-\varepsilon)}\Bigg[&\sqrt{2\sigma V(z_0 - z^\star)}\delta \\ &+ \frac{\sigma\varepsilon}{1-\varepsilon}\delta^2 + 2\sigma\delta^2 k + \frac{1}{k}V(z_0 - z^\star)\Bigg].\end{aligned} \tag{14}$$

Note that, in the exact case $\delta = 0$, we recover the convergence results stated in Proposition 6. Moreover, we have the following corollary.

Corollary 8 (Early-stopping). *Under the assumptions of Proposition 7, choose $k = c/\delta$ for some $c > 0$. Then there exist constants C, C' and C'' such that*

$$\begin{aligned}&\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k) \ \leqslant\ C\delta \ \text{and} \\ &\left\|\mathbf{X}\overline{w}^k - \mathbf{y}\right\|^2 \leqslant C'\delta + C''\delta^2 \ .\end{aligned}$$

The constants appearing in the Corollary are the ones from Proposition 7. They only depend on the saddle-point $z^\star$, the initialization z_0 and the step-sizes τ, σ. We next add some remarks.

Remark 9. *When J is γ-strongly convex, in particular when $J(\cdot) = \frac{1}{2}\|\cdot\|^2$, both the residual norm and the distance between the averaged iterate and the solution can be controlled by $\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k)$. Indeed, recalling Equation (25),*

$$\begin{aligned}\left\|\mathbf{X}\overline{w}^k - \mathbf{y}\right\|^2 &\leqslant \|\mathbf{X}\|^2 \left\|\overline{w}^k - w^\star\right\|^2 \\ &\leqslant \tfrac{2\|\mathbf{X}\|^2}{\gamma} D_J^{-\mathbf{X}^\top\theta^\star}\left(\overline{w}^k, w^\star\right) \\ &= \tfrac{2\|\mathbf{X}\|^2}{\gamma}\left[\mathcal{L}(\overline{w}^k, \theta^\star) - \mathcal{L}(w^\star, \overline{\theta}^k)\right] \ .\end{aligned}$$

In particular, the previous early-stopping bounds are of the same order of the ones obtained by dual gradient descent in Matet et al. (2017).

Remark 10. *Similar estimates have been obtained in Burger et al. (2007), both for the Tikhonov variational scheme and for the Bregman iteration (also called inverse scale space method) with stopping-criteria given by the discrepancy principle. In the first case (see Theorem 3.1), for a suitable choice of the regularization parameter, the authors get similar estimates for the Tikhonov regularized solution w_λ: $D_J^s(w_\lambda, w^\star) \leqslant C\delta$ and $\|\mathbf{X}w_\lambda - \mathbf{y}\|^2 \leqslant C'\delta^2$, where D_J^s is the symmetric Bregman divergence. For the Bregman iteration (see Theorem 4.2), they get an early-stopping bound on $D_J^{p_k}(w^\star, w_k)$, where $p_k \in \partial J(w_k)$. Note that they do not get any estimate for the quantity $D_J^{-\mathbf{X}^\top\theta^\star}\left(w^k, w^\star\right)$ neither for the residual norm. Moreover, the method requires to solve, at each iteration, an optimization problem with the same complexity of the original one.*

Proof Sketch The proof of Proposition 7 is inspired by Rasch and Chambolle (2020). In this paper, the kind of errors allowed in the prox of the non-extrapolated step (θ_k update) are more general than the ones allowed for the extrapolated step (w_k update). Here, we study stability properties of algorithm (6) when $\mathbf{y}$ is replaced by $\mathbf{y}^\delta$. This change can be read as an inexact proximity operator in the update of θ computation; in order to have this error in the non-extrapolated step, we study algorithm (6), that is CP algorithm applied to the dual problem. We summarize here the main steps. In Lemma 15, we derive a "descent property" for every step t, which we then cumulate summing from $t = 1$ to $t = k$ and using two different approximations (Lemmas 16 and 17). The two bounds that we get are similar, but independent. The first one has the following form,

$$\begin{aligned}\frac{1}{2\sigma}\|\theta_k - \theta^\star\|^2 + \sum_{t=1}^{k}\left[\mathcal{L}(w_t, \theta^\star) - \mathcal{L}(w^\star, \theta_t)\right] \leqslant \\ V(z_0 - z^\star) + \delta\sum_{t=1}^{k}\|\theta_t - \theta^\star\| \ .\end{aligned} \tag{15}$$

We use the latter twice. First we combine it with Lemma 14, a discrete version of Bihari's Lemma. This allows to estimate, for every $1 \leqslant t \leqslant k$, the quantity

$$\|\theta_t - \theta^\star\| \leqslant 2\sigma\delta k + \sqrt{2\sigma V(z_0 - z^\star)} \ . \tag{16}$$

Then we use again Equation (15), joint with the previous information, to find a bound on $\sum_{t=1}^{k} [\mathcal{L}(w_t, \theta^\star) - \mathcal{L}(w^\star, \theta_t)]$. The second inequality (see Lemma 17) has the following form,

$$\frac{\sigma\alpha}{2\eta} \sum_{t=1}^{k} \|\mathbf{X} w_t - \mathbf{y}\|^2 \leqslant V(z_0 - z^\star) + \delta \sum_{t=1}^{k} \|\theta_t - \theta\| + \frac{1}{2}\sigma(\eta - 1)\delta^2 k \ . \qquad (17)$$

Using again the bound on $\|\theta_t - \theta^\star\|$ and choosing $\eta = (1+\varepsilon)/(1-\varepsilon)$, we find an estimate for $\sum_{t=1}^{k} \|\mathbf{X} w_t - \mathbf{y}\|^2$. In both cases, we get the claim on the averaged iterates by Jensen's inequality.

Remark 11. *Following Rasch and Chambolle (2020), our analysis can straightforwardly be extended to the case in which* prox_J *in iterations (6) is computed approximately, with type-2 errors. This leads to bounds similar in spirit to those of Proposition 7, but involving the cumulative prox error up to iteration* k *and thus yielding slightly different stopping criteria.*

4.4 An example: sparse recovery

In the case of sparse recovery ($J = \|\cdot\|_1$), controlling the duality gap and the feasability yields a bound on the distance to the minimizer, thanks to the following result (Grasmair et al., 2011, Lemma 3.10).

Lemma 12. *Let* $(w^\star, \theta^\star)$ *be such that* $\mathbf{X} w^\star = \mathbf{y}$ *and* $-\mathbf{X}^\top \theta^\star \in \partial \|\cdot\|_1 (w^\star)$*. With* $\Gamma := \{j \in [p] : |\mathbf{X}_{:j}^\top \theta^\star| = 1\}$*, assume that* $\mathbf{X}_\Gamma$ *(*$\mathbf{X}$ *restricted to columns whose indices lie in* Γ*) is injective. Let* $m := \max_{j \notin \Gamma} |\mathbf{X}_{:j}^\top \theta^\star| < 1$*. Then, for all* $w \in \mathbb{R}^p$*,*

$$\|w - w^\star\| \leqslant \left\|\mathbf{X}_\Gamma^{-1}\right\|_{\mathrm{op}} \|\mathbf{X} w - \mathbf{y}\| + \frac{1+\left\|\mathbf{X}_\Gamma^{-1}\right\|_{\mathrm{op}} \|\mathbf{X}\|_{\mathrm{op}}}{1-m} D_{\|\cdot\|_1}^{-\mathbf{X}^\top \theta^\star}(w, w^\star) \ . \qquad (18)$$

Note that, under the assumptions of Lemma 12, the primal solution to Problem (19) is unique (see (Grasmair et al., 2011, Thm 4.7)). Combining the latter with Corollary 8 yields a strong early-stopping result.

Corollary 13 (Early-stopping for $J = \|\cdot\|_1$). *Under the assumptions of Proposition 7 and Lemma 12, choose* $k = c/\delta$ *for* $c > 0$*. Then there exist constants* C' *and* C'' *such that*

$$\left\|\overline{w}^k - w^\star\right\| \leqslant C'\sqrt{\delta} + C''\delta \ .$$

The constants C', C'' depend on the saddle-point $z^\star$, the initialization z_0, the step-sizes τ, σ and the norms of $\mathbf{X}$ and $\mathbf{X}_\Gamma^{-1}$. A completely different approach has been considered, for the same problem, in Vaškevičius et al. (2019). A related approach, based on dynamical systems, has been proposed in Osher et al. (2016). Similar results for the Tikhonov regularization approach can be found in Schöpfer and Lorenz (2019b).

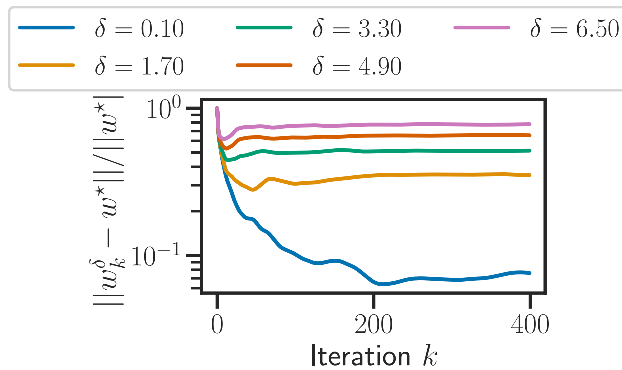

Figure 2: Distance between noisy Basis Pursuit iterates w_k^δ and noiseless solution $w^\star$, for various values of δ. There exists a stopping time: these distances reach a minimum before converging to their limit.

5 Empirical analysis

We stress that there is no implicit regularization result dealing with any non strongly convex J to compare to. The only competitor is therefore the Tikhonov approach. The code with scripts to reproduce the experiments (relying heavily on numpy (Harris et al., 2020) and numba (Lam et al., 2015)) is available at `lcsl.github.io/iterreg`.

5.1 Sparse recovery

Random data for this experiment are generated as follows: $(n, p) = (200, 500)$, columns of $\mathbf{X}$ are Gaussian with $\mathrm{cov}(\mathbf{X}_{:i}, \mathbf{X}_{:j}) = 0.2^{|i-j|}$, $\mathbf{y} = \mathbf{X} w_0$ where w_0 has 75 equal non zero entries, scaled such that $\|\mathbf{y}\| = 20$ (in order to have a meaningful range of values for δ). Note that the linear system $\mathbf{X} w = \mathbf{y}^\delta$ has solutions for any $\mathbf{y}^\delta$, since $\mathbf{X}$ is full-rank. The noiseless solution $w^\star$ is determined by running algorithm (6) up to convergence, on $\mathbf{y}$. For the considered values of δ, $\mathbf{y}^\delta$ is created by adding i.i.d. Gaussian noise to $\mathbf{y}$, so that $\left\|\mathbf{y} - \mathbf{y}^\delta\right\| = \delta$. We denote by w_k^δ the iterates of algorithm (6) ran on $\mathbf{y}^\delta$.

Existence of stopping time. In the first experiment, we highlight the existence of an optimal iterate in terms of distance to $w^\star$. Figure 2 shows *semi-convergence*: before converging to their limit, the iterates get close to $w^\star$. Note that this is stronger than the results of Corollary 13, since the optimal iteration here is the minimizer of the distance, and not some

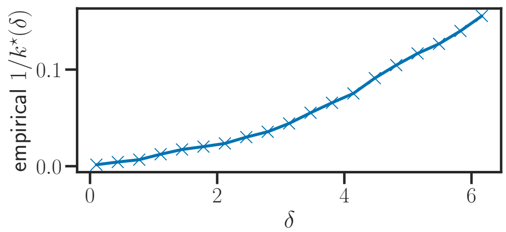

Figure 3: Influence of δ on the stopping time. In agreement with theory, the empirical stopping time roughly scales as $1/\delta$.

iterate for which there exists an upper bound on the distance to $w^\star$. As expected, as δ decreases, the optimal iteration k increases and the optimal iterate w_k^δ is closer to $w^\star$.

Dependency of empirical stopping time on δ. In the same setting as above, for 20 values of δ between 0.1 and 6, we generate 100 values of $\mathbf{y}^\delta$. We run algorithm (6) for 5000 iterations on $\mathbf{y}^\delta$ and determine the empirical best stopping time as $k^\star(\delta) = \arg\min_k \left\|w_k^\delta - w^\star\right\| < +\infty$. Figure 3 shows the mean of the inverse empirical stopping time as a function of δ, where a clear linear trend ($k = c/\delta$) appears as suggested by Proposition 7 and Corollary 13.

In real settings, $w^\star$ and δ are unknown, and so is the stopping time. It can still be evaluated by cross validation or similar procedure, as is usually done for the optimal λ in explicit regularization.

Comparison with Tikhonov approach on real data. The most popular regularization approach is to solve Problem (4) (here, the Lasso) for typically[1] 100 values of λ geometrically chosen as $\lambda_t = 10^{-3t/99}\left\|\mathbf{X}^\top \mathbf{y}\right\|_\infty$ for $t = 0, \dots, 99$. In Figure 4 we compare the Lasso regularization path to the Basis Pursuit optimization path of the Chambolle-Pock algorithm. The dataset for this experiment is rcv1-train from libsvm (Fan et al., 2008), with $(n, p) = (20\,242, 26\,683)$. The figure of merit is the prediction mean squared error on left out data, using 4-fold cross validation (dashed color lines), with the average across the folds in black. The horizontal line marks the λ (resp. the iteration k) for which the Lasso path (resp. the optimization path of iterations (6)) reaches its minimum MSE on the test fold.

The first observation is that the Basis Pursuit solution (both the end of the optimization ($k = +\infty$) and regularization paths ($\lambda = 0$)) performs very poorly, having a MSE greater than the one obtained by the 0 solution For the bottom plot, this would also be visible if

[1]default grid in scikit-learn (Pedregosa et al., 2011) and GLMNET (Friedman et al., 2010) packages

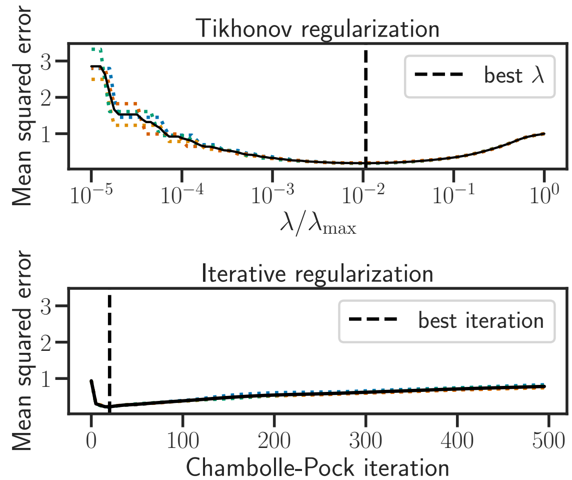

Figure 4: Comparison of Tikhonov regularization path (top) and optimization path of Algorithm (6) (bottom) on rcv1 with 4-fold cross-validation. Minimal value reached: 0.19 (top), 0.21 (bottom). Computation time up to optimal parameter: 50 s (top); 0.5 s (bottom).

the number of iterations of Algorithm (6) was picked greater than 500, which we do not do for readability of the figure. It is therefore necessary to early stop. The second observation is that the minimal MSEs on both paths are similar: 0.19 for Lasso path, 0.21 for optimization path of Algorithm (6). The main point is however that it takes 20 iterations of algorithm (6) to reach its best iterate, while the optimal λ for the Lasso is around $\lambda_{\max}/100$. If the default grid of 100 values between $\lambda_{\max}$ and $\lambda_{\max}/1000$ was used, this means that 66 Lassos must be solved, each one needing hundreds or thousands of iterations to converge. This is reflected in the timings: 0.5 s for Algorithm (6) vs 50 s for Tikhonov, eventhough we use a state-of-the-art coordinate descent + working set approach to solve the Lasso, with warm-start (using the solution for λ_{t-1} as initialization for problem with λ_t).

5.2 Low rank matrix completion

Random data for this experiment is generated as follows: the matrices are $d \times d$ with $d = 20$. $\mathbf{Y}$ is equal to $UV^\top$ with U and V of size $d \times 5$, whose entries are i.i.d Gaussian ($\mathbf{Y}$ is rank 5). We scale $\mathbf{Y}$ such that $\|\mathbf{Y}\| = 20$. Recall that in low rank matrix completion (Example 3), $\mathbf{X}$ corresponds to a masking operator (the observed entries); to determine which entries are observed, we uniformly draw $d^2/5$ observed couples $(i, j) \in [d] \times [d]$. Figure 5 shows the same type of results as Figure 2: iterates first approach the noiseless solution, then get further away, justifying early stopping of the iterates. For this experiment, we use higher values for δ to better highlight the semiconver-

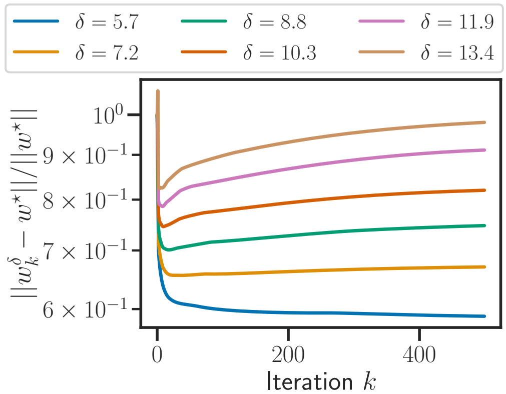

Figure 5: Distance between low rank matrix completion iterates w_k^δ and noiseless solution $w^\star$, for various values of δ. There exists a stopping time: a minimum before distance is reached before the limit.

gence, as curves get flatter for e.g. $\delta = 5.7$. Note that in that case, the algorithm can still be early stopped to save computations.

6 Conclusion

We have studied implicit regularization for convex bias, not necessarily strongly convex nor smooth. We proposed to use the Chambolle-Pock algorithm and we analyzed both convergence and stability to deterministic worst case noise. Our general analysis was specialized, as an example, to the problem of sparse recovery. The approach was investigated empirically both for sparse recovery and matrix completion, showing great timing improvements over relaxation approaches. A future development is to consider more specific noise models than the worst-case, such as stochastic noise. We emphasize again that our results hold in infinite dimension. It would be interesting to specialize our analysis in the finite dimensional setting, when the noisy solution always exists (in the least-square sense) and so the iterates produced by the algorithm are bounded. Moreover, it would be interesting to consider additional assumptions such as sparsity. Considering the role of initialization or nonlinear models would also be of interest. Finally, it would complete the analysis to obtain lower bounds for this class of problems, to confirm the sharpness of our results.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and the Italian Institute of Technology. Part of this work has been carried out at the Machine Learning Genoa (MaLGa) center, Università di Genova (IT). L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826. S. V. acknowledges the support of INDAM-GNAMPA, Project 2019: “Equazioni integro-differenziali: aspetti teorici e applicazioni”.

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *NeurIPS*, pages 1538–1546, 2012.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *NeurIPS*, pages 7413–7424, 2019.

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.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Martin Benning, Marta M Betcke, Matthias J Ehrhardt, and Carola-Bibiane Schönlieb. Gradient descent in a generalised Bregman distance framework. *arXiv preprint arXiv:1612.02506*, 2016.

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. Mach. Learn.*, 3:1–122, 2010.

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.

L. Calatroni, G. Garrigos, L. Rosasco, and S. Villa. Accelerated iterative regularization via dual diagonal descent. *arXiv preprint arXiv:1912.12153*, 2019.

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

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

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 V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

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.

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

R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

O. Fercoq and P. Bianchi. A coordinate descent primal-dual algorithm with large step size and possibly non separable functions. *SIAM Journal on Optimization*, 29(1):100–134, 2019.

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

J. Friedman, T. J. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1, 2010.

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

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. *arXiv preprint arXiv:1802.08246*, 2018.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with numpy. *arXiv preprint arXiv:2006.10256*, 2020.

T. J. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009.

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

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

S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

M. W. Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 143–154, 2012.

S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Image Process.*, 41:3397–3415, 1993.

S. Matet, L. Rosasco, S. Villa, and B. L. Vu. Don't relax: early stopping for convex regularization. *arXiv preprint arXiv:1707.05422*, 2017.

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.

A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

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. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

J. Peypouquet. *Convex Optimization in Normed Spaces. Theory, Methods and Examples.* Springer, 2015.

J. Rasch and A. Chambolle. Inexact first-order primal–dual algorithms. *Computational Optimization and Applications*, 76(2):381–430, 2020.

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):335366, 2014.

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

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

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 Journal on Optimization*, 22(4):1206–1223, 2012.

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

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

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge University Press, 2014.

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(1):2822–2878, 2018.

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58 (1):267–288, 1996.

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. *arXiv preprint arXiv:2002.00189*, 2020.

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

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.

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

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.