Iterative regularization with convex regularizers

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; Vaskevičius 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 methods (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-
Iterative regularization with convex regularizers

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,

$$f_w(x_i) = y_i, \quad \forall i \in [n]. \quad (1)$$

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:

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

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 $\gamma < 2/\|\mathbf{X}\|_{op}$, the iterations of gradient descent converge to a specific solution, namely

$$\arg \min_{w \in \mathbb{R}^p} \|w\| \quad \text{s.t.} \quad \mathbf{X} w = \mathbf{y}. \quad (3)$$

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),

$$w(\lambda) = \arg \min_{w \in \mathbb{R}^p} \lambda \|w\|^2 + \|\mathbf{y} - \mathbf{X} w\|^2 , \quad (4)$$

where the minimal norm solution (3) is obtained for $\lambda$ going to zero. It is well known that for Tikhonov regularization larger values of $\lambda$ 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 $\lambda$ 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 $\lambda$.

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

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

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, $X$ is an $n$ by $p$ matrix and $y$ an $n$-dimensional vector. We consider the case where $y$ is unknown, and a vector $y^\delta$ is available such that $\|y-y^\delta\| \leq \delta$, where $\delta \geq 0$ 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 $y$).

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

Note that we use a vectorial notation for simplicity but our results are dimension free and sharp for an infinite dimensional setting where $X$ 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 $\tau, \sigma$ such that $\sigma \tau \|X\|_{op}^2 < 1$:

$$\begin{aligned}
  w_{k+1} &= \text{prox}_{\tau J}(w_k - \tau X^T (2\theta_k - \theta_{k-1})) \\
  \theta_{k+1} &= \theta_k + \sigma (X w_{k+1} - y).
\end{aligned}$$

(6)

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

$$w_{k+1} = \text{prox}_{\tau J} \left( w_k - \tau \sigma X^T(\sum_{i=1}^k (Xw_i - y) + Xw_k - y) \right).$$

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 $J = \|\cdot\|_2^2$, the updates read:

$$\begin{aligned}
  w_{k+1} &= \frac{1}{1+\tau} (w_k - \tau X^T (2\theta_k - \theta_{k-1})) \\
  \theta_{k+1} &= \theta_k + \sigma (X w_{k+1} - y).
\end{aligned}$$

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

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

$$\begin{aligned}
  w_{k+1} &= \text{ST}(w_k - \tau X^T (2\theta_k - \theta_{k-1}), \tau) \\
  \theta_{k+1} &= \theta_k + \sigma (X w_{k+1} - y).
\end{aligned}$$

(6)

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 $X w^\star = y$. 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 $\ell_1$-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). $\ell_1$-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:

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

where $\| \cdot \|_*$ is the nuclear norm and $D \subseteq [p_1] \times [p_2]$ is the set of observed entries of the matrix $Y$. In that case, $X$ 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 $(i, j) \in D$ and 0 otherwise; the constraints write $XW = XY$.

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. $XW = Y$, where $X$ is usually a blurring operator. The problem can be reformulated as: $\min_{W} \Omega(W)$ s.t. $XW = \hat{Y}$, with $W = \begin{pmatrix} W \\ U \end{pmatrix}$, $\Omega(W) = \|U\|_1$, $\hat{X} = \begin{pmatrix} X & 0 \\ \nabla & -1d \end{pmatrix}$ and $\hat{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$.

Other examples include $\ell_2$ norm for antiparsimony, $\ell_1/\ell_2$ mixed-norm for group sparsity, or Kullback-Leibler divergence. Finally, for any linear operator $K$ and convex function $J$ such that $\text{prox}_J$ is available, we can handle problems of the form

$$\min_w J(Kw) \quad \text{s.t.} \quad Xw = y.$$

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

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:

$$\min_w f(Xw) + g(w).$$

Hence, for $f = \iota_{\{y\}}$ 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 $\text{prox}_J$, available in closed-form in many cases. The only tunable parameters are two step-sizes, $\tau$ and $\sigma$, 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

$$\mathcal{L}(w, \theta) = J(w) + \langle \theta, Xw - y \rangle,$$

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

$$\mathcal{L}(w^*, \theta) \leq \mathcal{L}(w^*, \theta^*) \leq \mathcal{L}(w, \theta^*) \quad (9).$$

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

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, $\mathcal{L}(w_k, \theta^*) - \mathcal{L}(w^*, \theta_k)$ and $\|Xw_k - y\|$). For these two quantities, we derive:

- convergence rates in the exact case, i.e. when the data $y$ is available (Proposition 6);
- early-stopping bounds in the inexact case, i.e. when the accessible data is only $y^\delta$ with $\|y^\delta - y\| \leq \delta$ (Proposition 7 and Corollary 8).
In Section 4.4, we apply our analysis to the specific choice of $J$ equal to the $\ell_1$-norm. In this particular case, combining the previous results, we even obtain bounds directly on $\|w_k - w^*\|$.

### 4.1 Measure of optimality

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

$$\mathcal{L}(w', \theta) - \mathcal{L}(w, \theta) \leq 0$$  \hspace{1cm} (10)

for every $(w, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$, then $(w, \theta)$ 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 $(w^*, \theta^*)$ and a generic $(w', \theta') \in \mathbb{R}^p \times \mathbb{R}^n$, it is popular to consider the quantity

$$\mathcal{L}(w', \theta^*) - \mathcal{L}(w^*, \theta') \geq 0.$$  \hspace{1cm} (11)

To establish the optimality of $(w', \theta')$, it is not enough to ensure $\mathcal{L}(w', \theta^*) - \mathcal{L}(w^*, \theta') = 0$. Lemma 5, proved in Appendix C.3, shows that, when coupled with $Xw' = y$, implies that $w'$ is a solution of Pb (5).

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

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

Note that $\mathcal{L}(w_k, \theta^*) - \mathcal{L}(w^*, \theta_k)$ 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 $J$, $\mathcal{L}(0, \theta^*) - \mathcal{L}(w^*, \theta_k)$ is 0 for any $\theta_k$; as shown in Figure 1, this quantity is also 0 when $w_k$ and $w^*$ have the same support and sign.

### 4.2 Exact case

First consider the iterates $(w_k, \theta_k)$ obtained by applying iterations (6) to the exact problem, namely where the data $y$ is available. Let $(w^*, \theta^*)$ 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 $z^* = (w^*, \theta^*)$ for the saddle-point. For $\tau$ and $\sigma > 0$, define $V$ as the following square weighted norm on $\mathbb{R}^p \times \mathbb{R}^n$:

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

For the averaged iterates $\bar{w}^k := \frac{1}{k} \sum_{t=1}^k w_t$ and $\bar{\theta}^k := \frac{1}{k} \sum_{t=1}^k \theta_t$, 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 $\sigma \tau \leq \varepsilon / \|X\|_{op}^2$. Then

$$\mathcal{L}(\bar{w}^k, \bar{\theta}^k) - \mathcal{L}(w^*, \theta^*) \leq \frac{V(z_0 - z^*)}{k}$$  \hspace{1cm} and

$$\|X\bar{w}^k - y\|^2 \leq \frac{2(1 + \varepsilon)V(z_0 - z^*)}{\varepsilon(1 - \varepsilon)k}.$$  \hspace{1cm}

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 $z^*$ 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 $(w_k, \theta_k)$, and their averaged versions $(\bar{w}^k, \bar{\theta}^k)$, obtained by applying iterations (6) to the noisy problem, where $y$ is replaced by $y^d$ with $\|y^d - y\| \leq \delta$. In Proposition 7, we derive early-stopping bounds for the iterates, in terms of duality gap $\mathcal{L}(w_k, \theta^*) - \mathcal{L}(w^*, \theta_k)$ and residual norm $\|Xw_k - y\|$. We highlight that, despite the error in the data $y^d$, both quantities are defined in terms of $y$ and hence related to the noiseless problem. In particular, $(w^*, \theta^*)$ 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 $\sigma \tau \leq \varepsilon / \|X\|_{op}^2$. Then,

$$\mathcal{L}(\bar{w}^k, \bar{\theta}^k) - \mathcal{L}(w^*, \theta^*) \leq \frac{1}{\varepsilon} \left( \sqrt{V(z_0 - z^*)} + \sqrt{2\sigma\delta k} \right)^2.$$  \hspace{1cm} (13)
and
\[
\|Xw_k^k - y\|^2 \leq \frac{2(1 + \varepsilon)}{\sigma \varepsilon (1 - \varepsilon)} \left[ \sqrt{2\sigma V(z_0 - z^*)} \delta + \frac{\sigma \varepsilon}{1 - \varepsilon} \delta^2 + 2\sigma \delta^2 k + \frac{1}{k} V(z_0 - z^*) \right].
\] (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
\[
L(\overline{w}_k, \theta^*) - L(w^*, \overline{\theta}) \leq C\delta \quad \text{and} \quad \|Xw_k^k - y\|^2 \leq C''\delta + C''\delta^2.
\]

The constants appearing in the Corollary are the ones from Proposition 7. They only depend on the saddle-point \( z^* \), the initialization \( z_0 \) and the step-sizes \( \tau, \sigma \). We next add some remarks.

**Remark 9.** When \( J \) is \( \gamma \)-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 \( L(\overline{w}_k, \theta^*) - L(w^*, \overline{\theta}) \). Indeed, recalling Equation (25),
\[
\|Xw_k^k - y\|^2 \leq \|X\|^2 \|w_k^k - w^*\|^2 \leq 2\|X\|^2 \frac{\gamma}{\sigma} J(\overline{w}_k, \theta^*) - J(w^*, \overline{\theta}) \leq 2\|X\|^2 \frac{\gamma}{\sigma} [L(\overline{w}_k, \theta^*) - L(w^*, \overline{w}_k)].
\]

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 \( w_\lambda \): \( D_\lambda (w_\lambda, w^*) \leq C\delta \) and \( \|Xw_\lambda - y\|^2 \leq C'\delta^2 \), where \( D_\lambda \) is the symmetric Bregman divergence. For the Bregman iteration (see Theorem 4.2), they get an early-stopping bound on \( D_{\lambda_k} (w^*, w_k) \), where \( p_k \in cJ(w_k) \). Note that they do not get any estimate for the quantity \( D_{\lambda_k} (w^*, w_k) \) 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 (\( \theta_k \) update) are more general than the ones allowed for the extrapolated step (\( w_k \) update). Here, we study stability properties of algorithm (6) when \( y \) is replaced by \( y^* \). This change can be read as an inexact proximity operator in the update of \( \theta \) 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 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,
\[
\frac{1}{2\sigma} \|\theta_k - \theta^*\|^2 + \sum_{t=1}^{k} [L(w_t, \theta^*) - L(w^*, \theta_t)] \leq V(z_0 - z^*) + \delta \sum_{t=1}^{k} \|\theta_t - \theta^*\|.
\] (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 \( 1 \leq t \leq k \), the quantity
\[
\|\theta_t - \theta^*\| \leq 2\sigma \delta k + \sqrt{2\sigma V(z_0 - z^*)}.
\] (16)
Then we use again Equation (15), joint with the previous information, to find a bound on $\sum_{t=1}^{k} [\mathcal{L}(w_t, \theta^*) - \mathcal{L}(w^*, \theta_t)]$. The second inequality (see Lemma 17) has the following form,

$$\frac{\sigma \alpha}{2\eta} \sum_{t=1}^{k} \|Xw_t - y\|^2 \leq V(z_0 - z^*) + \delta \sum_{t=1}^{k} \|\theta_t - \theta\|$$

$$+ \frac{1}{2} \sigma (\eta - 1) \delta^2 k . \quad (17)$$

Using again the bound on $\|\theta_t - \theta^*\|$ and choosing $\eta = (1 + \varepsilon)/(1 - \varepsilon)$, we find an estimate for $\sum_{t=1}^{k} \|Xw_t - y\|^2$. 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 $\text{prox}_J$ 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 feasibility yields a bound on the distance to the minimizer, thanks to the following result (Grasmair et al., 2011, Lemma 3.10).

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

$$\|w - w^*\| \leq \|X_{\Gamma}^{-1}\|_{op} \|Xw - y\|$$

$$+ \frac{1 + |X_{\Gamma}^{-1}|}{1 - m} \|X_{\Gamma}^{-1}\|_{op} D_{\| \cdot \|_1} \|X^T \theta^* (w, w^*)\| . \quad (18)$$

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

$$\|w^k - w^*\| \leq C' \sqrt{\delta} + C'' \delta .$$

The constants $C'$ and $C''$ depend on the saddle-point $z^*$, the initialization $z_0$, the step-sizes $\tau, \sigma$ and the norms of $X$ and $X_{\Gamma}^{-1}$. 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).

### 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 $X$ are Gaussian with $\text{cov}(X_{i}, X_{j}) = 0.2^{i-j}$, $y = Xw_0$ where $w_0$ has 75 equal non zero entries, scaled such that $\|y\| = 20$ (in order to have a meaningful range of values for $\delta$). Note that the linear system $Xw = y^\delta$ has solutions for any $y^\delta$, since $X$ is full-rank. The noiseless solution $w^*$ is determined by running algorithm (6) up to convergence, on $y$. For the considered values of $\delta$, $y^\delta$ is created by adding i.i.d. Gaussian noise to $y$, so that $\|y - y^\delta\| = \delta$. We denote by $w^k$ the iterates of algorithm (6) ran on $y^\delta$.

**Existence of stopping time.** In the first experiment, we highlight the existence of an optimal iterator in terms of distance to $w^*$. Figure 2 shows semi-convergence: before converging to their limit, the iterates get close to $w^*$. 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
iterate for which there exists an upper bound on the distance to $w^*$. As expected, as $\delta$ decreases, the optimal iteration $k$ increases and the optimal iterate $w_k^*$ is closer to $w^*$.

**Dependency of empirical stopping time on $\delta$.**

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

In real settings, $w^*$ and $\delta$ 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 $\lambda$ 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 $\lambda$ geometrically chosen as $\lambda_t = 10^{-3t/99} \|X^T y\|_{\infty}$ for $t = 0, \ldots, 99$. 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 $(n, p) = (20242, 26683)$. 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 $\lambda$ (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 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 $\lambda$ for the Lasso is around $\lambda_{\text{max}}/100$. If the default grid of 100 values between $\lambda_{\text{max}}$ and $\lambda_{\text{max}}/1000$ 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 $\lambda_{t-1}$ as initialization for problem with $\lambda_t$).

### 5.2 Low rank matrix completion

Random data for this experiment is generated as follows: the matrices are $d \times d$ with $d = 20$. $Y$ is equal to $UV^T$ with $U$ and $V$ of size $d \times 5$, whose entries are i.i.d Gaussian ($Y$ is rank 5). We scale $Y$ such that $\|Y\| = 20$. Recall that in low rank matrix completion (Example 3), $X$ 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 $\delta$ to better highlight the semiconver-
Figure 5: Distance between low rank matrix completion iterates $w^k_\delta$ and noiseless solution $w^*$, for various values of $\delta$. 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


Iterative regularization with convex regularizers


