Local Q-Linear Convergence and Finite-time Active Set Identification of ADMM on a Class of Penalized Regression Problems

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, Gaël Varoquaux

LOCAL Q-LINEAR CONVERGENCE AND FINITE-TIME ACTIVE SET IDENTIFICATION OF ADMM ON A CLASS OF PENALIZED REGRESSION PROBLEMS

Elvis Dohmatob *Michael Eickenberg* *Bertrand Thirion* *Gael Varoquaux*

Parietal Team, INRIA / CEA, Université de Paris-Saclay

ABSTRACT

We study the convergence of the ADMM (Alternating Direction Method of Multipliers) algorithm on a broad range of penalized regression problems including the Lasso, Group-Lasso and Graph-Lasso,(isotropic) TV-L1, Sparse Variation, and others. First, we establish a fixed-point iteration –via a nonlinear operator– which is equivalent to the ADMM iterates. We then show that this nonlinear operator is Fréchet-differentiable almost everywhere and that around each fixed point, Q-linear convergence is guaranteed, provided the spectral radius of the Jacobian of the operator at the fixed point is less than 1 (a classical result on stability). Moreover, this spectral radius is then a rate of convergence for the ADMM algorithm. Also, we show that the support of the split variable can be identified after finitely many iterations. In the anisotropic cases, we show that for sufficiently large values of the tuning parameter, we recover the optimal rates in terms of Friedrichs angles, that have appeared recently in the literature. Empirical results on various problems are also presented and discussed.

***Index Terms*—** ADMM, activity identification, linear convergence, structured sparsity, convex optimization

1 Introduction

ADMM [1, 2, 3] is an operator-splitting optimization method which is easy to implement and well-adapted for large-scale optimization problems [4]. For penalized regression problems with complicated composite penalties, such as for example analysis sparse problems, ADMM can provide a distinctive advantage over proximal gradient methods such as FISTA [5] when there is no closed-form expression for the *proximal operator* (see [6] for an overview of proximal calculus). Indeed, ADMM can avoid this difficulty by introducing a "split" variable, for which the proximal operator results in updates computable in closed-form. This is typically the case in *analysis sparsity* regularization, that impose sparsity on a transformation of the optimization variable [7]. However, the theory of the convergence rate of ADMM is not complete [4].

The present manuscript contributes to the understanding of the ADMM algorithm on penalized regression problems of the form

$$\underset{(w,z)\in\mathbb{R}^p\times\mathbb{R}^q}{\text{minimize}} \ \frac{1}{2}\|Xw-y\|^2 + \lambda\Omega(z) \text{ subject to } Kw - z = 0, \quad (1)$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix; $y \in \mathbb{R}^n$ is a vector of measurements or classification targets; $K \in \mathbb{R}^{q\times p}$ is linear operator; $\lambda > 0$ is the regularization parameter; and $\Omega : \mathbb{R}^p \to (-\infty, +\infty]$ is the penalty, which is assumed to be a *closed proper convex* function. Our main results are summarized in Theorem 1 (section 2), where in the case where Ω is an $\ell_{2,1}$ mixed-norm (as in Group-Lasso and Sparse Variation[8]), or a *concatenation* of such (i.e different norms acting on different blocks of coordinates of the same vector, as in TV-L1[9, 10]) , we derive –under mild conditions– an analytic formula for *a* Q-linear convergence rate (see e.g [11] for a precise definition) for the ADMM algorithm in terms of the spectral radius of certain Jacobian matrices, and also show finite-time recovery of the support of the "split" variable z, i.e of Kw.

The first author was funded by the EU FP7/2007-2013 under grant agreement no. 604102 (HBP), and also the iConnectome Digiteo grant.

1.1 Notation and terminology

For a positive integer n, denote $[n] := \{1, 2, ..., n\}$. The identity map will be denoted Id, and its domain of definition will be clear from the context. This same notation will be used for the identity matrix. As usual, $p \in [1, +\infty]$ the ℓ_p-norm of a vector $v \in \mathbb{R}^n$ will be denoted $\|v\|_p$. The *euclidean* / ℓ_2-norm will be denoted $\|v\|$ without subscript. Given $a \in \mathbb{R}^n$ and $\kappa \geq 0$, the closed ball (w.r.t the euclidean norm) centered at a with radius κ, is denoted $\bar{\mathbb{B}}_n(a, \kappa)$. When the center is 0, we will simply write $\bar{\mathbb{B}}_n(\kappa)$ for $\bar{\mathbb{B}}_n(0, \kappa)$. The *euclidean projection* onto a convex subset $C \subseteq \mathbb{R}^n$ will be denoted P_C. If A is square (i.e $m = n$), $\lambda(A)$ denotes the set of all its *eigenvalues*, and its *spectral radius*, denoted $r(A)$, corresponds to its largest absolute value of its eigenvalues. For any matrix A, its nonzero *singular values* are defined to be the square roots of the nonzero the eigenvalues of A^TA (or of AA^T). $\|A\|$ is the *spectral norm* of A, and is the largest of its singular-values. If $A \neq 0$, $\sigma_{\min^*}(A)$ denotes its smallest nonzero singular value.

1.2 The ADMM iterates for problem (1)

Consider the ADMM algorithm [1, 2, 3, 4] applied to problem (1). Let $\mu \in \mathbb{R}^q$ be the dual variable and $\rho > 0$ be the penalty parameter on the splitting residual. The augmented Lagrangian is:

$$\mathcal{L}_\rho(w, z, \mu) = \frac{1}{2}\|Xw-y\|^2 + \lambda\Omega(z) + \mu^T(Kw-z) + \frac{1}{2}\rho\|Kw-z\|^2.$$

Further, introducing the scaled dual variable $u := \mu/\rho$, which we will use instead of μ from here on, the ADMM iterates for problem (1) are given by the following equations:

$$\left.\begin{aligned}
w^{(n+1)} &\leftarrow \underset{w}{\text{argmin}}\, \mathcal{L}_\rho(w, z^{(n)}, u^{(n)}) = \\
&\qquad (\rho K^TK + X^TX)^{-1}(\rho K^T(z^{(n)} - u^{(n)}) + X^Ty) \\
z^{(n+1)} &\leftarrow \underset{z}{\text{argmin}}\, \mathcal{L}_\rho(w^{(n+1)}, z, u^{(n)}) = \\
&\qquad \text{prox}_{(\lambda/\rho)\Omega}(Kw^{(n+1)} + u^{(n)}) \\
u^{(n+1)} &\leftarrow u^{(n)} + Kw^{(n+1)} - z^{(n+1)}.
\end{aligned}\right\} \quad (2)$$

Assumptions. We will assume that the matrix sum $\rho K^TK + X^TX$ is invertible. This assumption is equivalent to $\ker K^TK \cap \ker X^TX = \{0\}$ (see e.g [12, Theorem 1]), which is reasonable in the context of regularization. Indeed, the idea behind this assumption is that, in high-dimensional problems ($n \ll p$), X typically has a large kernel, and so one would naturally choose K to act on it.

1.3 Examples: Some instances of problem (1)

Problem (1) covers a broad spectrum of problems encountered in pattern recognition and image processing. Here are a few:

Classical examples. We have $\Omega = \frac{1}{2}\|.\|^2$ for Ridge regression; $\Omega = \|.\|_1 : z \mapsto \sum_{j\in[p]} |z_j|$ for Lasso and Fused-Lasso [13]. For all but the last of these examples, we have $K = \text{Id}$. For Group-Lasso, we have $K = \text{Id}$, $\Omega =$ the *mixed-norm* $\ell_{2,1} = \|.\|_{2,1} : z \mapsto \sum_{j\in[d]} \|z_{j:j+c-1}\|$, where there are $d \geq 1$ blocks $z_{j:j+c-1} := (z_j, z_{j+1}, ..., z_{j+c-1})$ each of size $c \geq 1$.

Isotropic TV-L1 and Sparse Variation. These extensions of TV (Total Variation) proposed in the context of brain imaging, enforce sparse and structured (see Fig. 1) weights, w. We have $K = [\beta \, \mathrm{Id}, \; (1-\beta)\nabla]^T \in \mathbb{R}^{4p \times p}$, where ∇ is the discrete (multi-dimensional) spatial gradient operator and $\beta \in [0,1]$ is a mixing parameter. For TV-L1 [9, 10], the penalty is given by $\Omega(z) = \sum_{j \in [p]} |z_{j,1}| + \sum_{j \in [p]} \|z_{j,2:4}\|$ (i.e an ℓ_1 norm on the first p coordinates of z and an $\ell_{2,1}$ mixed-norm on the last $3p$ coordinates). In particular, the case $\beta = 1$ corresponds to the usual ℓ_1 norm, while $\beta = 0$ corresponds to the isotropic TV semi-norm.

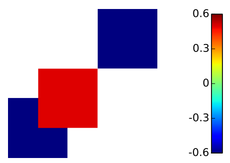

Fig. 1: Structured and sparse weights w

In Sparse Variation [8], the penalty is modified to simply be an $\ell_{2,1}$ mixed-norm on $d = p$ blocks of size $c = 4$ each, i.e $\Omega(z) = \sum_{j \in [p]} \|z_{j,1:4}\|$. TV-L1 and Sparse Variation combine sparsity (due to the the ℓ_1-norm) and structure (due to the isotropic TV term) to extract local concentrations of spatially correlated features from the data. Fig. 1 is a good illustration of the kinds of patterns one can learn using TV-L1 and Sparse Variation models.

2 Our contributions

2.1 Preliminaries

In the spirit of [11], let us start with a simple lemma (proof omitted) which rewrites the ADMM iterates (2) as a Picard fixed-point process in terms of the (z, u) pair of variables.

Lemma 2.1. *Define the following objects:*

$$G_\rho := K(K^TK + \rho^{-1}X^TX)^{-1}K^T, \;\; A_\rho := [G_\rho \;\; \mathrm{Id} - G_\rho],$$
$$b_\rho := \rho^{-1}K(K^TK + \rho^{-1}X^TX)^{-1}X^Ty, \; \tilde{A}_\rho := A_\rho(.) + b_\rho,$$
$$T_\rho := \left(\mathrm{prox}_{(\lambda/\rho)\Omega} \circ \tilde{A}_\rho, (\mathrm{Id} - \mathrm{prox}_{(\lambda/\rho)\Omega}) \circ \tilde{A}_\rho\right).$$

Then the z and u updates in the ADMM iterates (2) *can be jointly written as a Picard fixed-point iteration for the operator T_ρ, i.e*

$$(z^{(n+1)}, u^{(n+1)}) \leftarrow T_\rho(z^{(n)}, u^{(n)}). \tag{3}$$

In the special case where $\mathrm{prox}_{(\lambda/\rho)\Omega}$ is a linear transformation –as in Ridge regression or the nonnegative Lasso, for example– the operator T_ρ is linear so that the fixed-point iteration (3) is a linear dynamical system. Moreover, in such cases one can derive closed-form formulae for the spectral radius $r(T_\rho)$ of T_ρ as function of ρ, and thus recover the results of [11] and [14]. In the latter simple situations, a strategy for speeding up the ADMM algorithm is then to choose the parameter ρ so that the spectral radius of the linear part of the then affine transformation T_ρ is minimized. The following Corollary is immediate. Due to lack of space, we omit the proof, which is obtainable via the *Spectral Mapping Theorem*.

Corollary 2.2. *Let G_ρ, A_ρ, $\tilde{A}_\rho$, and T_ρ be defined as in Lemma 2.1. Then the following hold:*

(a) $\max(\|G_\rho\|, \|\mathrm{Id} - G_\rho\|) \le 1$, $\sigma_{\min^*}(A_\rho) \ge 1/\sqrt{2}$, *and* $\|A_\rho\| \le 1$ *with equality in the last inequality iff at least one of G_ρ and $\mathrm{Id} - G_\rho$ is singular.*

(b) T_ρ is $\|A_\rho\|$-Lipschitz. That is, $\forall (x_1, x_2) \in \mathbb{R}^{q+q} \times \mathbb{R}^{q+q}$,

$$\|T_\rho(x_1) - T_\rho(x_2)\| \le \|A_\rho\| \|x_1 - x_2\|. \tag{4}$$

In particular, if $\|A_\rho\| < 1$, then T_ρ is a contraction and the ADMM iterates (2) *converge globally Q-linearly to a solution of* (1). *Moreover, this solution is unique.*

According to Corollary 2.2, T_ρ is an $\|A_\rho\|$-contraction in case $\|A_\rho\| < 1$, and so we have global Q-linear convergence of the ADMM iterates (2) at the rate $\|A_\rho\|$. This particular case is analogous to the results obtained in [15] when the loss function or the penalty is strongly convex. But what if $\|G_\rho\| = \|\mathrm{Id} - G_\rho\| = \|A_\rho\| = 1$? Can we still have Q-linear convergence, –at least locally ? These questions are answered in the sequel.

2.2 Behavior of ADMM around fixed-points

Henceforth, we consider problem (1) in situations where the penalty Ω is an $\ell_{2,1}$ mixed-norm. Note that the ℓ_1-norm is a special case of the $\ell_{2,1}$ mixed-norm with $c = 1$ feature per block, and corresponds to the anisotropic case. The results presented in Theorem (1) carry over effortlessly to the case where the Ω is the concatenation of $\ell_{2,1}$ norms, for example as in the the TV-L1 semi-norm.The following theorem –inspired by a careful synthesis of the arguments in [16] and [17]– is our main result.

Theorem 1. *Consider the ADMM algorithm* (2) *on problem* (1), *where Ω is an $\ell_{2,1}$ mixed-norm on $d \ge 1$ blocks each of size $c \ge 1$, for a total of $q = d \times c$ features. Let the operators A, $\tilde{A}$, and T be defined as in Lemma 2.1, with the ρ subscript dropped for ease of notation. For $x \in \mathbb{R}^{q+q}$, define $\mathrm{supp}_z(x) := \{j \in [d] \mid x_{j:j+c-1} \neq 0\}$, $\mathcal{A}_z(x) := \{v \in \mathbb{R}^{q+q} | \, \mathrm{supp}_z(v) = \mathrm{supp}_z(x)\}$, $\tilde{X} = (\tilde{X}_j)_{j \in [d]}$, with $\tilde{X}_j = \tilde{A}(x)_j \in \mathbb{R}^c$, $\kappa := \lambda/\rho$, and $\epsilon(x) := \min_{j \in [d]} |\|\tilde{X}_j\| - \kappa|$.*

Then the following hold:

(a) ***Attractivity of supports.*** *For all $x \in \mathbb{R}^{q+q}$, we have*

$$T(\bar{\mathbb{B}}_{2q}(x, \epsilon(x)/\|A\|)) \subseteq \bar{\mathbb{B}}_{2q}(T(x), \epsilon(x)) \cap \mathcal{A}_z(T(x)).$$

In particular, if x^* is a fixed-point of T, then

$$T(\bar{\mathbb{B}}_{2q}(x^*, \epsilon(x^*)/\|A\|)) \subseteq \bar{\mathbb{B}}_{2q}(x^*, \epsilon(x^*)) \cap \mathcal{A}_z(x^*).$$

(b) ***Fréchet-differentiability.*** *If $x \in \mathbb{R}^{q+q}$ with $\epsilon(x) > 0$, then T is Fréchet-differentiable at x with derivative*

$$T'(x) = F_xA \in \mathbb{R}^{2q \times 2q}, \tag{5}$$

where $F_x := [D_x \;\; \mathrm{Id} - D_x]^T$ and $D_x \in \mathbb{R}^{q \times q}$ is a block-diagonal matrix with block $D_{x,j} \in \mathbb{R}^{c \times c}$ given by

$$D_{x,j} = \begin{cases} \mathrm{Id} - \frac{\kappa}{\|\tilde{X}_j\|} P_{\langle \tilde{X}_j \rangle^\perp}, & \text{if } j \in \mathrm{supp}_z(T(x)), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

In particular, when $c = 1$, each $D_{x,j}$ reduces to a bit $\in \{0, 1\}$ which indicates whether the jth feature is active, and D_x reduces to a diagonal projector matrix with only 0s and 1s.

(c) Let $x^* \in \mathbb{R}^{q+q}$ be any fixed-point of T.

(1) ***Finite-time identification of active set.*** *If the closed ball $\bar{\mathbb{B}}_{2q}(x^*, \epsilon(x^*)/\|A\|)$ contains any point of the sequence of iterates $x^{(n)}$, then the active set $\mathcal{A}_z(x^*)$ is identified after finitely many iterations, i.e*

$$\exists N_{x^*} \ge 0 \text{ s.t } x^{(n)} \in \mathcal{A}_z(x^*) \forall t \ge N_{x^*}. \tag{7}$$

In particular, (7) *holds if $x^{(n)}$ converges to x^*.*

(2) ***Local Q-linear convergence.*** *If $\epsilon(x^*) > 0$ and $r(T'(x^*)) < 1$, then the iterates $x^{(n)}$ converge locally Q-linearly to x^* at the rate $r(T'(x^*))$.*

(3) ***Optimal rates in the anisotropic case.*** *If $c = 1$ and ρ is large, then the optimal rate of convergence rate is the cosine of the Friedrichs angle between $\mathrm{Im}K$ and $\mathrm{Im}D_{x^*}$ = the canonical projection of $\mathcal{A}_z(x^*)$ onto $\mathbb{R}^q$. If in addition $K = \mathrm{Id}$ (as in the Lasso, sparse Spike-deconvolution, etc.), then the whole algorithm converges in finite time.*

Proof of Theorem 1. Recall notation and terminology from 1.1.

Part (a). For $x \in \mathbb{R}^{q+q}$ and any block index $j \in [d]$, observe that $T(x)_{j:j+c-1} = \text{soft}_\kappa(\tilde{X}_j)$, where soft_κ is the c-dimensional *soft-thresholding* operator, with threshold κ, defined by

$$\text{soft}_\kappa(v) := (1 - \kappa/\|v\|)_+ v = v - P_{\mathbb{B}_c(\kappa)}(v). \tag{8}$$

Now, one notes that Corollary 2.2*(b)* guarantees the set-inclusion $T(\bar{\mathbb{B}}_{2q}(x, \epsilon(x)/\|A\|)) \subseteq \bar{\mathbb{B}}_{2q}(T(x), \epsilon(x))$. It remains to show that $T(\bar{\mathbb{B}}_{2q}(x, \epsilon(x)/\|A\|)) \subseteq \mathcal{A}_z(T(x))$. Suppose on the contrary that there exists $x' \in \bar{\mathbb{B}}_{2q}(x, \epsilon(x)/\|A\|)$ such that $T(x') \notin \mathcal{A}_z(T(x))$. Then simultaneously, $\|x' - x\| \le \epsilon(x)/\|A\|$ and there exists an index $j \in [q]$ such that exactly one of $T(x)_{j:j+c-1} = \text{soft}_\kappa(\tilde{X}_j)$ and $T(x')_{j:j+c-1} = \text{soft}_\kappa(\tilde{X}'_j)$ is zero. Thus by the definition of soft_κ, we have $\min(\|\tilde{X}'{}_j\|, \|\tilde{X}_j\|) \le \kappa < \max(\|\tilde{X}'_j\|, \|\tilde{X}_j\|)$, from which $|\|\tilde{X}'_j\| - \|\tilde{X}'_j\|| > |\|\tilde{X}_j\| - \kappa| \ge \epsilon(x)$. Hence

$$\sum_{k \in [d]} \|\tilde{X}'_k - \tilde{X}_k\| \ge \|\tilde{X}'_j - \tilde{X}_j\| \ge |\|\tilde{X}'_j\| - \|\tilde{X}_j\|| > \epsilon(x). \tag{9}$$

On the other hand, by definition of $\tilde{X}$ and $\tilde{X}'$, we have

$$\sum_{k \in [d]} \|\tilde{X}'_k - \tilde{X}_k\| = \|A(x' - x)\| \le \|A\|\|x' - x\| \le \epsilon(x),$$

which is contradicted by (9). This proves the claim.

Part (b). Let $x \in \mathbb{R}^{q+q}$ with $\epsilon(x) > 0$. Then for any $j \in [d]$, we have $\|\tilde{X}^*_j\| \ne \kappa$, and so by [16, Theorem 2], the euclidean projection $P_{\bar{\mathbb{B}}_c(\kappa)}$ is differentiable in a neighborhood of $\tilde{X}_j$. Thus for small a perturbation $h \in \mathbb{R}^{2q}$ on x, and for any block $j \in [d]$, we have

$$\begin{aligned}
&(T(x+h) - T(x))_j = \text{soft}_\kappa(\tilde{X}_j + (Ah)_j) - \text{soft}_\kappa(\tilde{X}_j) \\
&= (\text{Id} - P_{\bar{\mathbb{B}}_c(\kappa)})(\tilde{X}_j + (Ah)_j) - (\text{Id} - P_{\bar{\mathbb{B}}_c(\kappa)})(\tilde{X}_j) \\
&= (Ah)_j - (P_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j + (Ah)_j) - P_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j)) \\
&= (\text{Id} - \text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j))(Ah)_j + o(\|h\|).
\end{aligned} \tag{10}$$

Now, invoking [16, equation (4.1)] and the ensuing paragraph therein, we compute $\text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j) = \kappa\|\tilde{X}_j\|^{-1} P_{\langle \tilde{X}_j \rangle^\perp}$ if $\|\tilde{X}_j\| > \kappa$, and $\text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j) = \text{Id}$ if $\|\tilde{X}_j\| < \kappa$. So, using the fact that $\|\tilde{X}_j\| > \kappa$ iff $j \in \text{supp}_z(T(x))$, we get $\text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j) = \kappa\|\tilde{X}_j\|^{-1} P_{\langle \tilde{X}_j \rangle^\perp}$ if $j \in \text{supp}_z(T(x))$ and $\text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j) = \text{Id}$ otherwise. Thus, from the definition of $D_{x,j}$ in the claim, we recognize $\text{proj}'_{\bar{\mathbb{B}}_c(\kappa)}(\tilde{X}_j) = \text{Id} - D_{x,j}$, and plugging into (10) yields

$$(T(x+h) - T(x))_j = D_{x,j}(Ah)_j + o(\|h\|).$$

Using the last equation and the definition of T, it follows that

$$\begin{aligned}
&(T(x+h) - T(x))_{j+d} = (Ah - (T(x+h) - T(x)))_j = \\
&\quad (\text{Id} - D_{x,j})(Ah)_j + o(\|h\|).
\end{aligned}$$

Putting everything together then yields

$$T(x+h) - T(x) - [D_x \;\; \text{Id} - D_x]^T Ah = o(\|h\|),$$

thus proving that T is Fréchet-differentiable at x with derivative $T'(x) = [D_x \;\; \text{Id} - D_x]^T A$. In particular, if $c = 1$, then $P_{\langle \tilde{X}_j \rangle^\perp} = 0$, and so $D_{x,j}$ reduces to a bit which is active iff $j \in \text{supp}_z(T(x))$.

Part (c1). Let x^* be as in the hypothesis. Indeed w.l.o.g, suppose $x^{(0)} \in \bar{\mathbb{B}}_{2q}(x^*, \epsilon(x^*)/\|A\|)$ and observe that $\|x^{(n)} - x^*\| = \|T(x^{(t-1)}) - T(x^*)\| \overset{(4)}{\le} \|A\|^{k-1}\|x^{(0)} - x^*\| \overset{\text{Theorem 2.1}(b)}{\le} \|x^{(0)} - x^*\| \le \epsilon(x^*)/\|A\|, \forall t > 0$. Thus we may choose $N_{x^*} = 0$ and the result (7) then follows from parts *(a)*. Now, suppose $x^{(n)} \overset{t \to \infty}{\longrightarrow} x^*$. Then every open neighborhood of x^* contains all but finitely many terms of the sequence. In particular, there exists $N_{x^*} \ge 0$ such that $\|x^{(n)} - x^*\| < \epsilon(x^*)/\|A\|$ for all $t \ge N_{x^*}$. The result (7) then follows from part *(a)* and the previously concluded argument.

Part (c2). Since $\epsilon(x^*) > 0$ by hypothesis, it follows from part *(b)* that T is Fréchet-differentiable at x^* with derivative $T'(x^*)$ given by (5). Also, since $r(T'(x^*)) < 1$ by hypothesis, we then deduce from [18, Theorem 4.3] (a refinement of [19, Theorem 10.1.4]) that the sequence of iterates $x^{(n)}$ converges to x^* locally Q-linearly at a rate $r(T'(x^*))$, which concludes the proof.

Part (c3). By the *Woodbury identity*, for large ρ we have

$$\begin{aligned}
G &= KK^+ - (XK^+)^T(\rho \text{I}_n + X(K^TK)^{-1}X^T)^{-1}XK^+ \\
&= KK^+ + o(\rho^{-1}\|XK^+\|^2) = P_{\text{Im}\,K} + o(\rho^{-1}\|XK^+\|^2).
\end{aligned} \tag{11}$$

Thus setting $U := \text{Im}\,K$, $V := \text{Im}\,D_{x^*}$, and using (11), we get

$$AF_{x^*} = P_U P_V + P_{U^\perp} P_{V^\perp} + o(\rho^{-1}\|XK^+\|^2). \tag{12}$$

Noting that $T'(x^*)^{n+1} = (F_{x^*}A)^{n+1} = F_{x^*}(AF_{x^*})^n A$ and invoking [20, Theorem 3.10], it follows that for large ρ, the matrix powers $T'(x^*)^n$ converge Q-linearly and the cosine of the Friedrichs angle between the subspaces U and V is the optimal rate of convergence. If in addition $K = \text{Id}$, then $U = \mathbb{R}^q$, so that $\cos\theta_F(U, V) = 0$ and the whole algorithm converges in finitely many iterations.

□

3 Relation to prior work

Recently, there have been a number of results on the local linear convergence of ADMM on particular classes of problems. Below, we outline the corresponding major works.

3.1 Ridge, QP, nonnegative Lasso

On problems like Ridge regression, quadratic programming (QP), and nonnegative Lasso, [11] demonstrated local linear convergence of ADMM under certain rank conditions which are equivalent to requiring that the p.s.d matrix G_ρ (defined in (3)) be invertible. The same paper prescribed explicit formulae for optimally selecting the tuning parameter ρ for ADMM on these problems. We note that these results can be recovered from our Lemma 2.1 and Corollary 2.2 as they correspond to the case where $\text{prox}_{(\lambda/\rho)\Omega}$ is a linear operator. Using similar spectral arguments, [14] demonstrated similar local convergence results for quadratic and linear QP problems.

3.2 Fréchet-differentiable nonlinear systems

In the SISTA algorithm [17], the authors linked the rate of convergence of their multi-band ISTA (refer to [21] and the references therein, for the original ISTA algorithm) scheme to the spectral radius of a certain Jacobian matrix related to the problem data and dependent on the fixed-point [17, Propositions 6 and 7], provided this spectral radius is less than 1. Most importantly, the authors show [17, Proposition 8] how their algorithm can be made as fast as possible by choosing the shrinkage parameter per sub-band to be "as large as possible". Finally, analogous to our Theorem 1*(a)*, Lemma 2 of [17] shows that the SISTA iteration projects points sufficiently close to fixed-points onto the support of these fixed-points.

3.3 Partly-smooth functions and Friedrichs angles

In the recent work [22] which focuses on Douglas-Rachford/ADMM, and [23] which uses the same ideas as in [22] but with a forward-backward scheme [24], the authors consider a subclass PSS (refer to definition 2.2 of [23]) of the class of so-called partly-smooth

(PS) penalties and general $\mathcal{C}^2$ loss functions with Lipschitz gradient. Under nonlinear complementarity requirements analogous to the non-degeneracy assumption "$\epsilon(x^*) > 0$" of Theorem 1*(b)*, and rank constraints analogous to the requirement that the Jacobian matrix $T'(x^*)$ have spectral radius less than 1 (in Theorem 1*(c2)*), the authors of [22, 23] prove finite-time activity identification and local Q-linear convergence at a rate given in terms of *Friedrichs angles*, via direct application of [20, Theorem 3.10]. The authors show that their arguments are valid for a broad variety of problems, for example the *anisotropic* TV penalty. Still in the framework of partly-smooth penalties, [25] showed local Q-linear convergence of the Douglas-Rachford algorithm on the Basis Pursuit problem.

Comparison with [22, 23]. The works which are most comparable to ours are [22] and [23], already presented above. Let us point out some similarities and differences between these papers and ours. First, though our constructions are entirely different from the techniques developed in [22, 23], one notes that both approaches are ultimately rooted in the same idea, namely the work of B. Holmes [16] on the smoothness of the euclidean projection onto convex sets, and other related functionals (Minkowski gauges, etc.). Indeed, Theorem 1 builds directly upon [16], whilst, [23] and [22] are linked to [16] via [26], which builds on [27], and the latter builds on [16].

Second, part *(c1)* of Theorem 1 (finite-time identification of active set) of the theorem can be recovered as a consequence of the results established in [22, 23]. However, the rest of our results, notably part *(c2)* (Q-linear convergence) cannot be recovered from the aforementioned works, at least on models like isotropic TV-L1, Sparse Variation, etc., since these models are not PSS. Indeed, the convergence rates in [22, 23] do not extend from anisotropic to isotropic TV, for example. Success in the former case is due to the fact that the anisotropic TV semi-norm is polyhedral and therefore is of class PSS at each point. By contrast, our framework can handle isotropic TV and similar "entangled" penalty types like isotropic TV-L1, Sparse Variation, etc., but suffers complementary limitations; for example, we were unable to generalize it beyond the squared-loss setting and we can only handle penalties which are a composition of a $\ell_{2,1}$ mixed-norm (or a concatenation of such) and a linear operator.

Lastly, the convergence rates in [22, 23] are tight and given in terms of Friedrichs angles [20], whilst our rates are given in terms of spectral radii, and will be suboptimal in certain cases. An exception are the anisotropic cases, where we proved in part *(c3)* of Theorem 1 that we recover the optimal rates obtained in [22, 23] in terms of Friedrichs angles. Moreover, for the Lasso, we showed the whole algorithm converges after only finitely many iterations.

4 Numerical experiments and results

Here, we present results for a variety of experiments. Each experiment is an instance of problem (1) with an appropriate choice of the linear operators X, K, and the penalty function Ω which can be the ℓ_1-norm the $\ell_{2,1}$ mixed-norm, or a mixture of the two (as in TV-L1).

Setting. We use a grid of 20 values of ρ, evenly spaced in log-space from 10^{-3} to 10^6. For each problem model (see below), the iteration process (3) is started with $x^{(0)} = 0 \in \mathbb{R}^{q \times q}$, and iterated $N = 1500$ times. The final point $x^{(N)}$ is approximately a fixed-point $x^{(*)}$ of the operator T_ρ. Now, the iteration process is run again (starting with the same initial $x^{(0)}$) and the distance $\|x^{(k)} - x^{(N)}\|$ is record on each iteration k, producing a curve. This procedure is run for each value of ρ from the aforementioned grid. Except otherwise stated, the n rows of design matrix X where drawn from a p-dimensional standard Gaussian. The measurements variable y is then computed as $y = Xw_0 + \text{noise}$, where w_0 is the true signal.

Simple models. As discussed in section 3, the local Q-linear convergence of ADMM on a variety of particular problems has been studied in the literature (for example [11, 15, 22, 23]). We validated empirically our linear convergence results (Theorem 1) by reproducing experiments from [22, 23]. For each of these experiments the regularization parameter λ was set to 1. Viz,

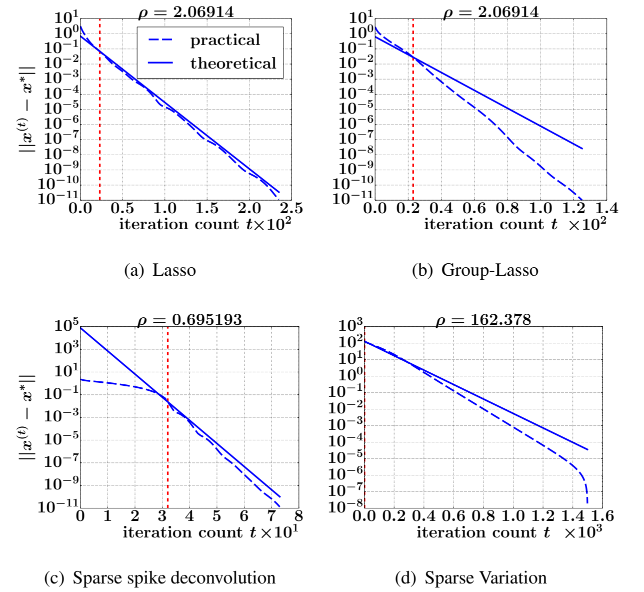

Fig. 2: Experimental results: Local Q-linear convergence for ADMM on problem (1). The "theoretical" line is the exponential curve $t \mapsto \|x^{(0)} - x^*\| r(T'(x^*))^t$. The red broken vertical line marks the instant $\mathcal{A}_z(x^*)$ is identified. We can see from figure that the upper bound for the local convergence rate (Theorem 1) is satisfied. Each shown thumbnail is for the value of ρ for which the spectral radius $r(T'_\rho(x^*))$ was smallest.

(a) Lasso: Here the problem is an instance of (1) with $K = \text{Id}$ and $\Omega = \|\cdot\|_1$; $n = 32$, $q = p = 128$, and w_0 is 8-sparse.

(b) Group-Lasso: Here $K = \text{Id}$ and $\Omega = \|\cdot\|_{2,1}$, $n = 48$, $p = 128$, number of blocks $d = 32$, block size = $c = 4$, $q = d \times c = 128$, w_0 is has 2 non-zero blocks.

(c) Sparse spikes deconvolution: Here, $K = \text{Id}$, X is a projector onto low Fourier frequencies (Dirichlet kernel) and the penalty Ω is the ℓ_1-norm; $n = p = 200$ (with $\operatorname{rank} X = 40$). The true signal w_0 is a 20-sparse vector (of length p), containing randomly distributed spikes with Gaussian values at a minimum pairwise distance of 5.

Sparse Variation: Going beyond known results. The data (refer to Fig. 1) is a simulation of $n = 200$ images of size $p = 12^3$ voxels each, with a set of 3 overlapping ROIs (Regions of Interests) each worth 5^3 voxels. Each ROI has a fixed weight which can be ± 0.5. The resulting images are then smoothed with a Gaussian kernel of width 2 voxel. This data is a toy model for brain activity. A Sparse Variation model (refer to subsection 1.3) with $\lambda = 10^2$ and $\beta = 0.5$ was then fitted on the data. It should be noted that the SV model is not PSS, and so the convergence rates in [22, 23] donnot apply.

The results for all the experiments are shown in Fig. 2.

5 Concluding remarks

We have derived a fixed-point iteration which is equivalent to the ADMM iterates for a broad class of penalized regression problems (1). Exploiting the formulation so obtained, we have established detailed qualitative properties of the algorithm around solution points (Theorem 1). Most importantly, under mild conditions, local Q-linear convergence is guaranteed and we have provided an explicit formula for this rate of convergence. Finally, Theorem 1 –implicitly– opens the possibility of speeding up the ADMM algorithm on problem (1) by selecting the tuning parameter ρ so as to minimize the spectral radius (an inverted mexican-hat-shaped curve, as ρ varies from 0 to $+\infty$) of the Jacobian matrix $T'_\rho(x^*)$.

6 References

[1] Roland Glowinski and A Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problemes de Dirichlet non linéaires," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 9, 1975.

[2] Daniel Gabay and Bertrand Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, 1976.

[3] Jonathan Eckstein and Dimitri P Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, 1992.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, 2011.

[5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183, 2009.

[6] Patrick L Combettes and Jean-Christophe 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.

[7] Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux, "Benchmarking solvers for tv-l1 least-squares and logistic regression in brain imaging," in *PRNI*. IEEE, 2014.

[8] Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux, "Total variation meets sparsity: statistical learning with segmenting penalties," in *MICCAI*. 2015.

[9] Luca Baldassarre, Janaina Mourao-Miranda, and Massimiliano Pontil, "Structured sparsity models for brain decoding from fMRI data," in *PRNI*, 2012, p. 5.

[10] Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux, "Identifying predictive regions from fMRI with TV-L1 prior," in *PRNI*, 2013.

[11] Euhanna Ghadimi, André Teixeira, Iman Shames, and Mikael Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems," *arXiv preprint arXiv:1306.2454*, 2013.

[12] R. Piziak, P.L. Odell, and R. Hahn, "Constructing projections on sums and intersections," *Pergamon, Computers and Mathematics with Applications*, 1999.

[13] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society Series B*, pp. 91–108, 2005.

[14] Daniel Boley, "Local linear convergence of the Alternating Direction Method of Multipliers on quadratic or linear programs," *SIAM Journal on Optimization*, vol. 23, 2013.

[15] Robert Nishihara, Laurent Lessard, Benjamin Recht, Andrew Packard, and Michael I Jordan, "A general analysis of the convergence of admm," *arXiv preprint arXiv:1502.02009*, 2015.

[16] Richard B. Holmes, "Smoothness of certain metric projections of Hilbert space," *Trans. Amer. Math. Soc.*, vol. 184, 1973.

[17] Ilker Bayram and Ivan W Selesnick, "A Subband Adaptive Iterative Shrinkage/Thresholding Algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, 2010.

[18] Werner C. Rheinboldt, "Iterative methods for nolinear systems," *https://www-m2.ma.tum.de/foswiki/pub/M2/Allgemeines/SemWs09/nonlin-it.pdf*, year $\geq$ 2004.

[19] James M. Ortega and Werner C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Computer science and applied mathematics. Academic Press, 1970.

[20] Heinz H Bauschke, JY Cruz, Tran TA Nghia, Hung M Phan, and Xianfu Wang, "Optimal rates of convergence of matrices with applications," *arXiv preprint arXiv:1407.0671*, 2014.

[21] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413, 2004.

[22] Jingwei Liang, Jalal Fadili, Gabriel Peyré, and Russell Luke, "Activity identification and local linear convergence of Douglas–Rachford/ADMM under partial smoothness," *arXiv preprint arXiv:1412.6858*, 2014.

[23] Jingwei Liang, Jalal Fadili, and Gabriel Peyré, "Activity identification and local linear convergence of inertial forward-backward splitting," *arXiv preprint arXiv:1503.03703*, 2015.

[24] Patrick L Combettes and Valérie R Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, 2005.

[25] Laurent Demanet and Xiangxiong Zhang, "Eventual linear convergence of the Douglas-Rachford iteration for basis pursuit," *CoRR*, vol. abs/1301.0542, 2013.

[26] S.J Wright, "Identifiable surfaces in constrained optimization," *SIAM Journal on Control and Optimization*, vol. 31, pp. 49–67, 1993.

[27] S. Fitzpatrick and R. R. Phelps, "Differentiability of the metric projection in hilbert space," *Trans. Am. Math. Soc.*, vol. 270, pp. 483–501, 1982.