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

To cite this version:


HAL Id: hal-01265372
https://hal.archives-ouvertes.fr/hal-01265372
Submitted on 1 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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. Then we show that this nonlinear operator is Fréchet-differentiable almost everywhere and that around each fixed point, Q-linear convergence is guaranteed, provided the operator is Fréchet-differentiable almost everywhere and that around

The present manuscript contributes to the understanding of the convergence rate (see e.g [11] for a precise definition) for the ADMM algorithm in terms of the spectral radius of the identity matrix [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

\[ \min_{(u,v)} \frac{1}{2} \| Xw - y \|^2 + \lambda \Omega(z) \text{ subject to } Kw - z = 0, \]

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; \( K \in \mathbb{R}^{n \times p} \) 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 \( \Omega \) 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 \( Kw \).

1.1 Notation and terminology

For a positive integer \( n \), denote \([n] := \{1, 2, ..., n\} \). The identity map will be denoted \( \text{Id} \), 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 \( \ell_p \)-norm of a vector \( v \in \mathbb{R}^n \) will be denoted \( \|v\|_p \). The euclidean / \( \ell_2 \)-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 \( \kappa \), is denoted \( B_a(a, \kappa) \). When the center is 0, we will simply write \( B_a(0, \kappa) \).

We will assume that the matrix sum \( AB \) is square (i.e \( n = n \)), \( \lambda(A) \) denotes the set of all its eigenvalues, and its spectral radius, \( \rho(A) \), corresponds to the square roots of the nonzero eigenvalues of \( A^T A \) (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).

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

\[ u^{(n+1)} \leftarrow \arg\min_{u} \frac{\rho K^T (K X T X)^{-1} (u^{(n)} - K X v^{(n)}) + Y^T y^{(n)}}{\|K w^{(n)} - z^{(n)}\|^2} \]

Assumptions. We will assume that the matrix sum \( \rho K^T K + X^T X \) is invertible. This assumption is equivalent to ker \( K^T K \cap \ker X^T X = \{0\} \) (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 \( \Omega = \frac{1}{2} \|z\|^2 \) for Ridge regression; \( \Omega = \|z\|_1 : z \mapsto \sum_{j \in [p]} |z_j| \) for Lasso and Fused-Lasso [13]. For all but the last of these examples, we have \( K = \text{Id} \), \( \Omega = \) the mixed-norm \( \ell_2,1 = \|z\|_2,1 : z \mapsto \sum_{j \in [d]} |z_j|^{\frac{1}{2}} \), where there are \( d \geq 1 \) blocks \( z_{j+1} := (z_j, z_{j+1}, ..., z_{j+c-1}) \) 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 u, w. We have K = [β Id, (1 − β)∇]T ∈ Rn×p, where ∇ is the discrete (multi-dimensional) spatial gradient operator and β ∈ [0, 1] is a mixing parameter. For TV-L1 [9, 10], the penalty is given by Ω(z) = ∑j∈[p]∥zj∥1 + ∑j∈[p]∥zj,2,4∥ (i.e an l1 norm on the first p coordinates of z and an l2,4 norm on the last 3p coordinates). In particular, the case β = 1 corresponds to the usual l1 norm, while β = 0 corresponds to the isotropic TV semi-norm.

In Sparse Variation [8], the penalty is modified to simply be an l2,1 mixed-norm on d = p blocks of size c = 4 each, i.e Ω(z) = ∑j∈[p]∥zj,1,4∥. TV-L1 and Sparse Variation combine sparsity (due to the the l1-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 (u, v) pair of variables.

Lemma 2.1. Define the following objects:

Gρ := K(KT + ρ−1XTX)−1KT, Ap := Gρ Id − Gρ, bρ := ρ−1K(KT + ρ−1XTX)−1XTy, ᾱρ := Ap(·) + bρ,

Tρ := (proxλ/(ρ,Ω) ◦ ᾱρ, Id − proxλ/(ρ,Ω) ◦ ᾱρ).

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

zz(n+1), u(n+1) ← Tρ(zz(n), u(n)). (3)

In the special case where proxλ/(ρ,Ω) is a linear transformation –as in Ridge regression or the nonnegative Lasso, for example– the operator Tρ 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(Tρ) of Tρ as function of ρ, and thus recover the results of [11] and [14]. In the latter simple situations, a strategy for speeding up the algorithm is then to choose the parameter ρ so that the spectral radius of the linear part of the then affine transformation Tρ 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 Gρ, Ap, ᾱρ, and Tρ be defined as in Lemma 2.1. Then the following hold:

(a) max(∥Gρ∥, 1/∥Id − Gρ∥) ≤ 1, σmin(Ap) ≥ 1/√n, and ∥Ap∥ ≤ 1 with equality in the last inequality if at least one of Gρ and Id − Gρ is singular.

(b) Tρ is |Aρ|L-Lipschitz. That is, ∀(x1, x2) ∈ Rn×p × Rn×p,

∥Tρ(x1) − Tρ(x2)∥ ≤ ∥Aρ∥∥x1 − x2∥. (4)

In particular, if ∥Aρ∥ < 1, then Tρ 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, Tρ is an |Aρ|−contract in case ∥Aρ∥ < 1, and so we have global Q-linear convergence of the ADMM iterates (2) at the rate |Aρ|−. This particular case is analogous to the results obtained in [15] when the loss function or the penalty is strongly convex. But what if ∥Gρ∥ = 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 l2,1 mixed-norm. Note that the l1-norm is a special case of the l2,1 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 l2,1 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 l2,1 mixed-norm on d ≥ 1 blocks of size c ≥ 1, for a total of q = dc features. Let the operators A, A, and T be defined as in Lemma 2.1, with the ρ subscript dropped for ease of notation. For x ∈ Rn×q, define suppρ(x) := {j ∈ [d] | xj,1,1−c−1 ≠ 0}, Aρ(x) := {v ∈ Rn×q | suppρ(v) = suppρ(x)}, X = (Xj)j∈[d], with Xj = Aρ(xj) ∈ Rn, ρ := λ/ρ, and ε(x) := minj∈[d]∥∥Xj∥∥ − κ.

Then the following hold:

(a) Attractivity of supports. For all x ∈ Rn×q, we have

\[ T(B_{2q}(x, ε(x)/|A|)) ⊆ B_{2q}(T(x), ε(x)) \cap A(T(x)). \]

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

\[ T(B_{2q}(x∗, ε(x∗)/|A|)) ⊆ B_{2q}(x∗, ε(x∗)) \cap A(x∗) \]

(b) Fréchet-differentiability. If x ∈ Rn×q with ε(x) > 0, then T is Fréchet-differentiable at x with derivative T ′(x) = F(x, A) ∈ Rn×q, where F(x, A) := [Dx Id − D∗x]T and Dx ∈ Rn×q is a block-diagonal matrix with block Ds,j ∈ Rn×n given by

\[ Ds,j := \begin{cases} \text{Id} - \frac{n}{πXj}P(Xj) & \text{if } j \text{ is supp}ρ(T(x)), \\ 0, & \text{otherwise.} \end{cases} \]

In particular, when c = 1, each Ds,j reduces to a δ ∈ [0, 1], which indicates whether the jth feature is active, and Ds reduces to a diagonal projector matrix with only 0s and 1s.

(c) Let x∗ ∈ Rn×q be any fixed-point of T.

(1) Finite-time identification of active set. If the closed ball B2q(x∗, ε(x∗)/|A|) contains any point of the sequence of iterates x(n), then the active set Aρ(x∗) is identified after finitely many iterations, i.e.

\[ \exists N ϵ, ρ \geq N \Rightarrow x(n) ∈ Aρ(x∗) \forall n \geq N ϵ, \]

(7) In particular, (7) holds if x(n) converges to x∗.

(2) Local Q-linear convergence. If ε(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 ImK and ImDx∗ = the canonical projection of Aρ(x∗) onto Rn. In addition K = 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 $x \in \mathbb{R}^{q+d}$ and any block index $j \in [d]$, observe that $T(x)_{j,j+c-1} = \text{soft}_x(X_j)$, where \text{soft}_x is the $c$-dimensional soft-thresholding operator, with threshold $\kappa$, defined by

$$\text{soft}_x(v) := (1 - \kappa/\|v\|)_+ v = v - F_{\kappa/\|v\|}(v). \quad (8)$$

Now, one notes that Corollary 2.2(b) guarantees the set-inclusion $T(\mathbb{B}_2(x, \epsilon \|x\|/\|A\|)) \subseteq \mathbb{B}_2(T(x), \epsilon \|x\|)$. It remains to show that $T(\mathbb{B}_2(x, \epsilon \|x\|/\|A\|)) \subseteq A_c(T(x))$. Suppose on the contrary that there exists $x' \in \mathbb{B}_2(x, \epsilon \|x\|/\|A\|)$ such that $T(x') \notin A_c(T(x))$. Then, since $x' \in \mathbb{B}_2(x, \epsilon \|x\|/\|A\|)$, there exists an index $j \in [q]$ such that exactly one of $T(x)_{j,j+c-1} = \text{soft}_x(X_j)$ and $T(x')_{j,j+c-1} = \text{soft}_x(X_j)$ is zero. Thus by the definition of $\text{soft}_x$, we have $\min(jX_j, \|X_j\|) \leq \kappa < \max(jX_j, \|X_j\|)$ from which $\|X_j\| - \|X_j\| > \|X_j\| - \kappa \geq \epsilon(x)$. Hence

$$\sum_{k \in [d]} \|\tilde{X}_k - X_k\| \geq \|\tilde{X}_j - X_j\| > \|\tilde{X}_j - X_j\| > \epsilon(x). \quad (9)$$

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

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

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

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

$$(T(x + h) - T(x))_{j,j} = \text{soft}_x(\tilde{X}_j + (Ah)_j) - \text{soft}_x(\tilde{X}_j)$$

$$= (\text{Id} - P_{\kappa/\|v\|}(X_j + (Ah)_j)) - (\text{Id} - P_{\kappa/\|v\|}(X_j))$$

$$= (Ah)_j - (P_{\kappa/\|v\|}(X_j + (Ah)_j) - P_{\kappa/\|v\|}(X_j))$$

$$= (\text{Id} - \text{proj}_{\kappa/\|v\|}(X_j))(Ah)_j + o(\|h\|). \quad (10)$$

Now, invoking [16, equation (4.1)] and the ensuing paragraph therein, we compute $\text{proj}_{\kappa/\|v\|}(X_j) = \kappa(\|X_j\|^{-1}P_{\kappa}(X_j))$. If $\|X_j\| > \kappa$, and $\text{proj}_{\kappa/\|v\|}(X_j)$ is Id if $\|X_j\| < \kappa$. So, using the fact that $\|X_j\| > \kappa$ iff $j \in \text{supp}(T(x))$, we get $\text{proj}_{\kappa/\|v\|}(X_j) = \kappa(\|X_j\|^{-1}P_{\kappa}(X_j))$. If $j \in \text{supp}(T(x))$ and $\text{proj}_{\kappa/\|v\|}(X_j) = \text{Id}$ otherwise. Thus, from the definition of $D_{x,j}$ in the claim, we recognize $\text{proj}_{\kappa/\|v\|}(X_j) = \text{Id} - D_{x,j}$, and plugging into (10) yields

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

Thus $\text{proj}_{\kappa/\|v\|}(X_j)$ is Fréchet-differentiable at $h = 0$. If $x = 0$, then $T = 0$, and so $D_{x,j}$ reduces to a bit which is active iff $j \notin \text{supp}(T(x))$.

Part (c). Let $x^*$ be as in the hypothesis. Indeed w.l.o.g. suppose $x^{(0)} \in \mathbb{B}_2(x^*, \epsilon(x^*)/\|A\|)$ and observe that $\|x^{(n)} - x^*\|^2 = \|T(x^{(n-1)}) - T(x^*)\| \leq \|A\|^{k-1}\|x^{(0)} - x^*\|^2 \leq \|x^{(0)} - x^*\|^2 \leq \epsilon(x^*)/\|A\|$, $\forall t > 0$. Thus we may choose $N_t = 0$ and the result (7) then follows from parts (a). Now, suppose $x^{(n)} \notin \mathbb{B}_2(x^*, \epsilon(x^*)/\|A\|)$. Then every open neighborhood of $x^*$ contains all but finitely many terms of the sequence. In particular, there exists $N_{\epsilon(x^*)} \geq 0$ such that $\|x^{(n)} - x^*\| < \epsilon(x^*)/\|A\|$ for all $t \geq N_{\epsilon(x^*)}$. 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 $\epsilon(T'(x^*)) < 1$ 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 $\mu(T'(x^*))$, which concludes the proof.

Part (c3). By the Woodbury identity, for large $\rho$ we have

$$G = KK^+ - (XK^T)^T\rho_\kappa + (XK^T)^{-1}XK^+$$

$$= KK^+ + o((\rho^{-1}\|XK^+\|^2) = \rho_{\kappa} + o((\rho^{-1}\|XK^+\|^2). \quad (11)$$

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

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

Noting that $T'(x^{(n)})^{+1} = (F_{x^*}A^{+})^{+1} = F_{x^*}(A_{x^*}^{+})^{+1}$ and invoking [20, Theorem 3.10], it follows that for large $\rho$, the matrix powers $T'(x^{(n)})$ converge Q-linearly and the cosine of the Friedichs angles 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 $\cos \theta_{PU}(U, V) = 0$ 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 $G_\kappa$ (defined in (3)) be invertible. The same paper prescribed explicit formulae for optimally selecting the tuning parameter $\rho$ 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 $\text{prox}_{\kappa/\|v\|}$ 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 Friedichs 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 $\ell^2$ loss functions with Lipschitz gradient. Under nonlinear complementarity requirements analogous to the non-degeneracy assumption $r(x^*) \geq 0$ 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(c)), 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 (c) 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 $\Omega$ which can be the $\ell^1$-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 $\rho$, evenly spaced in log-space from $10^{-3}$ to $10^{6}$. For each problem model (see below), the iteration process (3) is started with $x^{(0)} = 0 \in \mathbb{R}^{n \times k}$, and iterated $N = 1500$ times. The final point $x^{(N)}$ is approximately a fixed-point $x^*$ of the operator $T\rho$. 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 $\rho$ from the aforementioned grid. Except otherwise stated, the $n$ rows of design matrix $X$ drawn from a $p$-dimensional standard Gaussian. The measurements variable $y$ is then computed as $y = X w_0 + \text{noisy}$, where $w_0$ is the true signal.

Simple models. As discussed in section 3, the local Q-linear convergence of [22, 23] under a anisotropic TV semi-norm 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 $\lambda$ was set to 1. Viz,

For each problem model (see below), the data (refer to Fig. 1) is a simulation of $y = X w_0 + \text{noisy}$. The red broken vertical line marks the instant $\| A_t x^* \|$ 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 $\rho$ for which the spectral radius $r(T^*_{\rho}(x^*))$ is smallest.

Fig. 2: Experimental results: Local Q-linear convergence for ADMM on problem (1). The “theoretical” line is the exponential curve $t \mapsto \|x^{(0)} - x^*\| / \|T(x^*)\|$. The red broken vertical line marks the instant $A_t x^*$ 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 $\rho$ for which the spectral radius $r(T^*_{\rho}(x^*))$ is smallest.

(a) Lasso: Here the problem is an instance of (1) with $K = \text{Id}$ and $\Omega = \| \cdot \|_1$; $\rho = 92$, $p = p = 128$, 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 $c = 4$, $q = d \times c = 128$, $w_0$ 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 $\Omega$ is the $\ell_1$-norm; $n = p = 200$ (with 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 $\pm 0.5$. 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 $\lambda = 10^3$ 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 $\rho$ so as to minimize the spectral radius (an inverted mexican-hat-shaped curve, as $\rho$ varies from 0 to $\infty$) of the Jacobian matrix $T_{\rho}(x^*)$. 

}\]
6 References


