Exploiting regularity in sparse Generalized Linear Models

Mathurin Massias†, Samuel Vaiter‡, Alexandre Gramfort†, Joseph Salmon⋆
† INRIA, Palaiseau, France
‡ CNRS & Institut de Mathématiques de Bourgogne, Dijon, France
⋆ IMAG, Univ Montpellier, CNRS, Montpellier, France
Email: mathurin.massias@inria.fr

***Abstract*— Generalized Linear Models (GLM) are a wide class of regression and classification models, where the predicted variable is obtained from a linear combination of the input variables. For statistical inference in high dimensions, sparsity inducing regularization have proven useful while offering statistical guarantees. However, solving the resulting optimization problems can be challenging: even for popular iterative algorithms such as coordinate descent, one needs to loop over a large number of variables. To mitigate this, techniques known as *screening rules* and *working sets* diminish the size of the optimization problem at hand, either by progressively removing variables, or by solving a growing sequence of smaller problems. For both of these techniques, significant variables are identified by convex duality. In this paper, we show that the dual iterates of a GLM exhibit a Vector AutoRegressive (VAR) behavior after sign identification, when the primal problem is solved with proximal gradient descent or cyclic coordinate descent. Exploiting this regularity one can construct dual points that offer tighter control of optimality, enhancing the performance of screening rules and helping to design a competitive working set algorithm.**

I. Introduction

Since the introduction of the Lasso [30], sparsity inducing penalties have had a tremendous impact on Machine Learning [3]. They have been applied to a variety of regression and classification tasks: sparse logistic regression [19], Group Lasso [36], multitask Lasso [25]. These estimators fall under the framework of Generalized Linear Models, where the prediction is drawn from an exponential family distribution whose mean is a linear combination of the input variables. The key property of ℓ_1 regularization is that it allows to perform jointly feature selection and prediction, which is particularly useful in high dimensional settings. Amongst the algorithms proposed to solve these, coordinate descent[1] [33, 14] is the most popular in Machine Learning scenarios [10, 13].

Since only a fraction of the coefficients are non-zero in the optimal parameter vector, a recurring idea to speed up solvers is to limit the size of the optimization problem by ignoring features which are not included in the solution. To do so, two approaches can be distinguished:

- *screening rules*, introduced by [9] and later developed [34, 35, 7], progressively remove features from the problems in a backward approach,
- *working sets* techniques [11, 26, 31, 17] solve a sequence of smaller problems restricted to a growing number of features.

Both the current state-of-art methods for screening (Gap Safe rules, [12, 24]) and working sets (Blitz, [17, 18]) rely heavily on a dual point to identify useful features. However, although a lot of attention has been devoted to creating a sequence of iterates in the primal converging fast to the primal optimum [13], the construction of dual iterates has not been scrutinized, and the standard approach [21], although robust and converging, is crude.

In this paper, we propose a principled way to construct a sequence of dual points converging faster to the dual optimum. Based on an extrapolation procedure [29], its cost is exactly the same as the classical approach, which allows to retain the stability and convergence guarantees of the latter. We define, quantify and prove the asymptotic VAR behavior of dual iterates for sparse GLMs solved with proximal gradient descent or cyclic coordinate descent. The resulting new construction:

- provides a tighter optimality control through duality gap,
- improves uniformly the performance of Gap safe rules,
- improves the performance of the working set policy proposed in [22], thanks to better feature identification,

We introduce the framework of ℓ_1-regularized Generalized Linear Models, Gap safe rules in Section II. We justify and generalize dual extrapolation, orginally introduced for the Lasso [23] to any ℓ_1-regularized GLM in Section III. We show how to use dual extrapolation to create efficient working sets in Section IV. Results of Section V illustrate the benefits of dual extrapolation.

II. Notation and framework

a) Notation: For any integer $d \in \mathbb{N}$, we denote by $[d]$ the set $\{1, \dots, d\}$. The design matrix $X \in \mathbb{R}^{n \times p}$ is composed of observations $\mathbf{x}_i \in \mathbb{R}^p$ stored row-wise, and whose j-th column is $x_j \in \mathbb{R}^n$; the vector $y \in \mathbb{R}^n$ (resp. $\{-1, 1\}^n$) is the response vector for regression (resp. binary classification). The support of $\beta \in \mathbb{R}^p$ is $\mathcal{S}(\beta) = \{j \in [p] : \beta_j \neq 0\}$. For $\mathcal{W} \subset [p]$, $\beta_\mathcal{W}$ and $X_\mathcal{W}$ are β and X restricted to features in $\mathcal{W}$. As much as possible, exponents between parenthesis (*e.g.,* $\beta^{(t)}$) denote iterates and subscripts (*e.g.,* β_j) denote vector entries or matrix columns. The sign function is $\operatorname{sign} : x \mapsto x/|x|$ with the convention $0/0 = 0$. The sigmoid function is $\sigma : x \mapsto 1/(1+e^{-x})$. The soft-thresholding of x at level u is $\mathrm{ST}(x, u) = \operatorname{sign}(x) \cdot \max(0, |x| - u)$. Applied to vectors, sign, σ and $\mathrm{ST}(\cdot, u)$ (for $u \in \mathbb{R}_+$) act element-wise. Element-wise product between vectors of same length is denoted by $\odot$. The vector of size n whose entries are all equal to 0 (resp. 1) is denoted by $\mathbf{0}_n$ (resp. $\mathbf{1}_n$). On square matrices, $\|\cdot\|_2$ is the spectral norm. For a symmetric postive definite matrix H, $\langle x, y \rangle_H = x^\top H y$ is the H-weighted inner product, whose associated norm is denoted $\|\cdot\|_H$. We extend the small-o notation to vector valued functions in the following way: for $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$, $f = o(g)$ if and only if $\|f\| = o(\|g\|)$, *i.e.,* $\|f\|/\|g\|$ tends to 0 when $\|g\|$ tends to 0. For a convex and proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, its Fenchel-Legendre conjugate $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined by $f^*(u) = \sup_{x \in \mathbb{R}^n} u^\top x - f(x)$.

[1] throughout the paper, this means *cyclic and proximal* coordinate descent unless specified otherwise

Definition 1 (Sparse Generalized Linear Model)**.** *We consider the following problem:*

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \underbrace{\sum_{i=1}^{n} f_i(\beta^\top \mathbf{x}_i) + \lambda \|\beta\|_1}_{\mathcal{P}(\beta)} \ , \tag{1}$$

where all f_i are convex functions with $1/\gamma$-Lipschitz gradients. Instances of Pb. (1) *are the Lasso ($f_i(t) = \frac{1}{2}(y_i - t)^2$, $\gamma = 1$) and Sparse Logistic regression ($f_i(t) = \log(1 + \exp(-y_i t))$, $\gamma = 4$).*

Proposition 2. *A dual formulation of Problem* (1) *reads:*

$$\hat{\theta} = \arg\max_{\theta \in \Delta_X} \underbrace{\left(- \sum_{i=1}^{n} f_i^*(-\lambda \theta_i) \right)}_{\mathcal{D}(\theta)} \tag{2}$$

where $\Delta_X = \{\theta \in \mathbb{R}^n \ : \ \|X^\top \theta\|_\infty \leq 1\}$. $\hat{\theta}$ is unique, because the f_i's are γ-strongly convex. The KKT conditions read:

$$\forall i \in [n],\ \hat{\theta}_i = -f_i'(\hat{\beta}^\top \mathbf{x}_i) \qquad \textbf{\textit{(link equation)}} \tag{3}$$

$$\forall j \in [p],\ x_j^\top \hat{\theta} \in \partial |\cdot|(\hat{\beta}_j) \qquad \textbf{\textit{(subdifferential inclusion)}} \tag{4}$$

If for $u \in \mathbb{R}^n$ we write $F(u) \overset{\text{def.}}{=} \sum_{i=1}^n f_i(u_i)$, the link equation reads $\hat{\theta} = -\nabla F(X\hat{\beta})/\lambda$.

Remark 3. *For any $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$, one has $\mathcal{D}(\theta) \leq \mathcal{P}(\beta)$. Denoting $\mathcal{P}(\beta) - \mathcal{D}(\theta)$ the duality gap, it can be used as an upper bound for the sub-optimality of the current β: for any $\epsilon > 0$, any $\beta \in \mathbb{R}^p$, and any (feasible) $\theta \in \Delta_X$:*

$$\mathcal{P}(\beta) - \mathcal{D}(\theta) \leq \epsilon \Rightarrow \mathcal{P}(\beta) - \mathcal{P}(\hat{\beta}) \leq \epsilon \ . \tag{5}$$

This shows, that even though $\hat{\beta}$ is unknown in practice and the sub-optimality gap cannot be evaluated, creating a dual (feasible) point $\theta \in \Delta_X$ allows to guarantee an ϵ-solution is reached, and it can therefore be used to get a tractable stopping criterion.

By design of the ℓ_1 penalty, $\hat{\beta}$ is sparse, and the larger λ is, the sparser $\hat{\beta}$ gets. Thus, a key principle to speed up these PG or CD is to identify the support of $\hat{\beta}$ so that features outside of it can be ignored, which leads to a smaller and easier problem to solve. Removing features when it is guaranteed that they are not in the support of the solution is at the heart of the so-called *Gap Safe Screening rules* [12, 24]:

Proposition 4 (Gap Safe Screening rule)**.** *The Gap Safe screening rule for Problem* (1) *reads: $\forall j \in [p], \forall \theta \in \Delta_X$,*

$$|x_j^\top \theta| < 1 - \|x_j\| \sqrt{\tfrac{2}{\gamma \lambda^2}(\mathcal{P}(\beta) - \mathcal{D}(\theta))} \implies \hat{\beta}_j = 0 \ . \tag{6}$$

Therefore, while running an iterative solver and computing the duality gap at iteration t, the criterion (6) can be tested for all features j, and the features guaranteed to be inactive at optimum can be ignored.

Equations (5) and (6) do not require a specific choice of θ. Because of the link equation $\hat{\theta} = -\nabla F(X\hat{\beta})/\lambda$, a natural way to construct a dual feasible point $\theta^{(t)} \in \Delta_X$ at iteration t, when only a primal vector $\beta^{(t)}$ is available, is:

$$\theta^{(t)}_{\text{res}} = -\nabla F(X\beta^{(t)}) / \max(\lambda, \|X^\top \nabla F(X\beta^{(t)})\|_\infty) \ . \tag{7}$$

This was coined *residuals rescaling* following the terminology used of the Lasso case where $-\nabla F(X\beta) = y - X\beta$ [21].

To improve the control of sub-optimality, and to better identify useful features, the aim of *dual extrapolation* is to obtain a better dual point (*i.e.,* closer to the optimum $\hat{\theta}$). The idea is to do it at a low computational cost by exploiting the structure of the sequence of dual iterates $(X\beta^{(t)})_{t \in \mathbb{N}}$; we explain what is meant by "structure", and how to exploit it, in the following definition and proposition:

Definition 5. *We say that $(r^{(t)})_{t \in \mathbb{N}} \in (\mathbb{R}^d)^{\mathbb{N}}$ is a Vector AutoRegressive (VAR) sequence (of order 1) if there exists $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$ such that for $t \in \mathbb{N}$:*

$$r^{(t+1)} = Ar^{(t)} + b \ . \tag{8}$$

We also say that the sequence $(r^{(t)})_{t \in \mathbb{N}}$, converging to $\hat{r}$, is an asymptotic VAR sequence if

$$r^{(t+1)} - Ar^{(t)} - b = o(r^{(t)} - \hat{r}) \ . \tag{9}$$

Definition 6 (Vector AutoRegressive sequence)**.** *We say that $(r^{(t)})_{t \in \mathbb{N}} \in (\mathbb{R}^n)^{\mathbb{N}}$ is a Vector AutoRegressive (VAR) sequence (of order 1) if there exists $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ such that for $t \in \mathbb{N}$:*

$$r^{(t+1)} = Ar^{(t)} + b \ . \tag{10}$$

We also say that the sequence $(r^{(t)})_{t \in \mathbb{N}}$, converging to $\hat{r}$, is an asymptotic VAR sequence if there exist $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ such that for $t \in \mathbb{N}$:

$$r^{(t+1)} - Ar^{(t)} - b = o(r^{(t)} - \hat{r}) \ . \tag{11}$$

We can now introduce formally the extrapolation procedure, as formalized for optimization tasks in [29] although the idea dates back to [1] and [2, 8] in the vector case.

Proposition 7 (Extrapolation for VAR sequences [29, Thm 3.2.2])**.** *Let $(r^{(t)})_{t \in \mathbb{N}}$ be a VAR sequence in $\mathbb{R}^n$, satisfying $r^{(t+1)} = Ar^{(t)} + b$ with $A \in \mathbb{R}^{n \times n}$ a symmetric positive definite matrix with $\|A\|_2 < 1$, and $b \in \mathbb{R}^n$. Let $K < n$, and for $t \geq K$ let:*

$$U^{(t)} = [r^{(t-K)} - r^{(t-K+1)}, \dots, r^{(t-1)} - r^{(t)}] \in \mathbb{R}^{n \times K} \ , \tag{12}$$

$$(c_1, \dots, c_K) = \frac{(U^{(t)\top} U^{(t)})^{-1} \mathbf{1}_K}{\mathbf{1}_K^\top (U^{(t)\top} U^{(t)})^{-1} \mathbf{1}_K} \in \mathbb{R}^K \ , \tag{13}$$

$$r_{extr} = \sum_{k=1}^{K} c_k r^{(t-K-1+k)} \in \mathbb{R}^n \ . \tag{14}$$

Then,

$$\|Ar_{extr} - b - r_{extr}\| \leq \mathcal{O}(\rho^K) \ , \tag{15}$$

where $\rho = \frac{1 - \sqrt{1 - \|A\|}}{1 + \sqrt{1 - \|A\|}} < 1$.

The justification for this approach is the following: for $t \in \mathbb{N}$, we have $r^{(t+1)} - \hat{r} = A(r^{(t)} - \hat{r})$. Let $(a_0, \dots, a_n) \in \mathbb{R}^{n+1}$ be the coefficients of A's characteristic polynomial. By Cayley-Hamilton's theorem, $\sum_{k=0}^n a_k A^k = 0$. Since $\|A\|_2 < 1$, 1 is not an eigenvalue of A and $\sum_{k=0}^n a_k \neq 0$, so we can normalize these coefficients to have $\sum_{k=0}^n a_k = 1$. If $t \geq n$, we have:

$$\sum_{k=0}^{n} a_k (r^{(t-n+k)} - \hat{r}) = \left(\sum_{k=0}^{n} a_k A^k \right) (r^{(t-n)} - \hat{r}) = 0 \ , \tag{16}$$

$$\sum_{k=0}^{n} a_k r^{(t-n+k)} = \sum_{k=0}^{n} a_k \hat{r} = \hat{r}. \tag{17}$$

Hence, $\hat{r} \in \text{Span}(r^{(t-n)}, \dots, r^{(t)})$. Therefore, it is natural to approximate $\hat{r}$ as an affine combination of the $(n+1)$ last iterates $(r^{(t-n)}, \dots, r^{(t)})$. Using $(n+1)$ iterates might be costly for large n, so one might rather consider only a smaller number K, *i.e.,* find $(c_1, \dots, c_K) \in \mathbb{R}^K$ s.t. $\sum_{k=1}^K c_k r^{(t-K-1+k)}$ approximates $\hat{r}$. Since $\hat{r}$ is a fixed point of $r \mapsto Ar + b$, $\sum_{k=1}^K c_k r^{(t-K-1+k)}$ should be one too. Under a normalizing condition $\sum_{k=1}^K c_k = 1$, this means that

$$r_{extr} - Ar_{extr} - b = \sum_{k=1}^{K} c_k (r^{(t-K-1+k)} - r^{(t-K+k)})$$

should be as close to $\mathbf{0}_n$ as possible; minimizing the norm of the RHS under $c^\top \mathbf{1}_K = 1$ admits a closed-form solution:

$$\hat{c} = \frac{(U^{(t)\top} U^{(t)})^{-1} \mathbf{1}_K}{\mathbf{1}_K^\top (U^{(t)\top} U^{(t)})^{-1} \mathbf{1}_K} \ , \tag{18}$$

where $U^{(t)} = [r^{(t-K+1)} - r^{(t-K)}, \dots, r^{(t)} - r^{(t-1)}] \in \mathbb{R}^{n \times K}$.

Finally, to exhibit VAR sequences, we will use the following result on sign identification for sparse GLMs.

Theorem 8 (Sign identification)**.** *Let $(\beta^{(t)})_{t\in\mathbb{N}}$ be the sequence of iterates converging to $\hat{\beta}$ and produced by PG or CD when solving Problem* (1) *(the solution might not be unique, but the algorithms converge to a unique, well-defined value, which we call $\hat{\beta}$). There exists $T \in \mathbb{N}$ s.t. : $\forall j \in [p], t \geq T \implies \text{sign}(\beta_j^{(t)}) = \text{sign}(\hat{\beta}_j)$. The smallest epoch T for which this holds is when* sign identification *is achieved.*

Proof For lighter notation in this proof, denote $l_j = \|x_j\|^2/\gamma$ and $h_j(\beta) = \beta_j - \frac{1}{l_j} x_j^\top \nabla F(X\beta)$. The first order optimality conditions for the sparse GLM model defined in Eq. (1) are:

$$\forall j \in [p], \quad \frac{x_j^\top \nabla F(X\hat{\beta})}{\lambda} \in \begin{cases} \{1\}, & \text{if } \hat{\beta}_j > 0, \\ \{-1\}, & \text{if } \hat{\beta}_j < 0, \\ [-1, 1], & \text{if } \hat{\beta}_j = 0. \end{cases} \tag{19}$$

Motivated by these conditions, the *equicorrelation set* [32] is:

$$E \overset{\text{def.}}{=} \{j \in [p] \,:\, |x_j^\top \nabla F(X\hat{\beta})| = \lambda\} = \{j \in [p] \,:\, |x_j^\top \hat{\theta}| = 1\} \ .$$

We introduce the *saturation gap* associated to Problem (1):

$$\hat{\delta} \overset{\text{def.}}{=} \min \left\{ \frac{\lambda}{l_j} \left(1 - \frac{|x_j^\top \nabla F(X\hat{\beta})|}{\lambda} \right) : j \notin E \right\} \ . \tag{20}$$

As $\hat{\theta} = \nabla F(X\hat{\beta})|/\lambda$ is unique, $\hat{\delta}$ is well-defined, and strictly positive by definition of E. By (19), the support of any solution is included in the equicorrelation set, with equality when the solution is unique [32]. We will also need the following technical results about the soft-thresholding operator.

Lemma 9. *For any $x, y \in \mathbb{R}$, and any $\nu > 0$:*

$$|\,\text{ST}(x, \nu) - \text{ST}(y, \nu)| \leq |x - y| \tag{21}$$

$$|x| > \nu, |y| < \nu \Rightarrow |\,\text{ST}(x, \nu)| \leq |x - y| - (\nu - |y|) \tag{22}$$

$$|y| \geq \nu, \text{sign}\, x \neq \text{sign}\, y \Rightarrow |\text{ST}(x, \nu) - \text{ST}(y, \nu)| \leq |x - y| - \nu \tag{23}$$

Proof The first result in Lemma 9 comes from the nonexpansiveness of proximal operators [4, Theorem 6.42]. For the other ones, see [15, Lemma 3.2]. ∎

We start by showing a weaker result: the coefficients outside the equicorrelation eventually vanish. The proof requires to study the primal iterates after each update (instead of after each epoch). Hence, we use the notation $\tilde{\beta}^{(s)}$ to denote the primal iterate after the s-th update of coordinate descent. This update only modifies the j-th coordinate, with $s \equiv j-1 \mod p$:

$$\tilde{\beta}_j^{(s+1)} = \text{ST}\left(h_j(\tilde{\beta}^{(s)}), \tfrac{\lambda}{l_j} \right) \ . \tag{24}$$

Note that at optimality, for every $j \in [p]$, one has:

$$\hat{\beta}_j = \text{ST}\left(h_j(\hat{\beta}), \tfrac{\lambda}{l_j} \right). \tag{25}$$

Let us consider an update $s \in \mathbb{N}$ of CD such that the updated coordinate j verifies $\tilde{\beta}_j^{(s+1)} \neq 0$ and $j \notin E$, hence, $\hat{\beta}_j = 0$. Then, the following holds true, using Eq. (22):

$$\begin{aligned} |\tilde{\beta}_j^{(s+1)} - \hat{\beta}_j| &= \left| \text{ST}\left(h_j(\tilde{\beta}^{(s)}), \tfrac{\lambda}{l_j} \right) - \text{ST}\left(h_j(\hat{\beta}), \tfrac{\lambda}{l_j} \right) \right| \\ &\leq \left| h_j(\tilde{\beta}^{(s)}) - h_j(\hat{\beta}) \right| - \left(\tfrac{\lambda}{l_j} - |h_j(\hat{\beta})| \right) \ . \end{aligned} \tag{26}$$

Now notice that by definition of the saturation gap (20), and since $j \notin E$:

$$\begin{aligned} &\frac{\lambda}{l_j} \left(1 - \frac{|x_j^\top \nabla F(X\hat{\beta})|}{\lambda} \right) \geq \hat{\delta} \ , \\ \text{that is,} \quad &\frac{\lambda}{l_j} - |h_j(\hat{\beta})| \geq \hat{\delta} \quad (\text{using } \hat{\beta}_j = 0) \ . \end{aligned} \tag{27}$$

Combining Equations (26) and (27) yields

$$|\tilde{\beta}_j^{(s+1)} - \hat{\beta}_j| \leq \left| h_j(\tilde{\beta}^{(s)}) - h_j(\hat{\beta}) \right| - \hat{\delta} \ . \tag{28}$$

This can only be true for a finite number of updates, since otherwise taking the limit would give $0 \leq -\hat{\delta}$. Therefore, after a finite number of updates, $\tilde{\beta}_j^{(s)} = 0$ for $j \notin E$.

We can now show the sign identification result for $j \in E$. First observe that for all $j \in E$, $|h_j(\hat{\beta})| \geq \frac{\lambda}{l_j}$. Indeed, if $j \in \mathcal{S}(\hat{\beta})$, $0 \neq \hat{\beta}_j = \text{ST}(h_j(\hat{\beta}), \frac{\lambda}{l_j})$ so $|h_j(\hat{\beta})| > \frac{\lambda}{l_j}$. If $j \in E \setminus \mathcal{S}(\hat{\beta})$, $|h_j(\hat{\beta})| = |\frac{1}{l_j} x_j^\top \nabla F(X\hat{\beta})| = \frac{\lambda}{l_j}$.

Now let $s \in \mathbb{N}$ and $j \in E$ be such that $\text{sign}\, \tilde{\beta}_j^{(s+1)} \neq \text{sign}\, \hat{\beta}_j$.

$$\begin{aligned} \left| \tilde{\beta}_j^{(s+1)} - \hat{\beta}_j \right|^2 &= \left(\text{ST}\left(h_j(\tilde{\beta}^{(s)}), \tfrac{\lambda}{l_j} \right) - \text{ST}\left(h_j(\hat{\beta}), \tfrac{\lambda}{l_j} \right) \right)^2 \\ &\leq \left(\left| h_j(\tilde{\beta}^{(s)}) - h_j(\hat{\beta}) \right| - \frac{\lambda}{l_j} \right)^2 \qquad \text{using (23)} \\ &\leq \left(\left| h_j(\tilde{\beta}^{(s)}) - h_j(\hat{\beta}) \right|^2 - \frac{\lambda^2}{l_j^2} \right), \end{aligned} \tag{29}$$

because since $|h_j(\hat{\beta})| \geq \frac{\lambda}{l_j}$ and $\text{sign}\, h_j(\hat{\beta}) = \text{sign}\, \hat{\beta}_j \neq \text{sign}\, \tilde{\beta}_j^{(s+1)} = \text{sign}\, h_j(\tilde{\beta}^{(s)})$, we have $|h_j(\tilde{\beta}^{(s)}) - h_j(\hat{\beta})| \geq \frac{\lambda}{l_j}$. Equation (29) can only happen for a finite number of updates, otherwise taking the limit would yield a contradiction.

The proof for proximal gradient descent is a result of [15, Theorem 4.1], who give the bound $T \leq \|\tilde{\beta}^{(s)} - \hat{\beta}\|_2^2 / \hat{\delta}^2$. ∎

Note that Theorem 8 does not require an hypothesis on the uniqueness of the solution. Even if there are multiple solutions, CD or PG will converge to one of them [15], and identify its sign in a finite number of iterations.

III. Generalized models

A. Coordinate descent for ℓ_1 regularization

Dual extrapolation was introduced for the Lasso [23]: we now generalize it to Problem (1).

Theorem 10 (VAR for coordinate descent and Sparse GLM)**.** *When Problem* (1) *is solved by cyclic coordinate descent, the dual iterates* $(X\beta^{(t)})_{t\in\mathbb{N}}$ *form an asymptotical VAR sequence.*

Proof We place ourselves in the identified sign regime, and consider only one epoch t of CD: let $\tilde{\beta}^{(0)}$ denote the value of the primal iterate at the beginning of the epoch ($\tilde{\beta}^{(0)} = \beta^{(t)}$), and for $s \in [S]$, $\tilde{\beta}^{(s)} \in \mathbb{R}^p$ denotes its value after the j_s coordinate has been updated ($\tilde{\beta}^{(S)} = \beta^{(t+1)}$). Recall that in the framework of Problem (1), the data-fitting functions f_i have $1/\gamma$-Lipschitz gradients, and $\nabla F(u) = (f_1'(u_1), \dots, f_n'(u_n))$. For $s \in [S]$, $\tilde{\beta}^{(s)}$ and $\tilde{\beta}^{(s-1)}$ are equal everywhere except at entry j_s, for which the coordinate descent update with fixed step size $\frac{\|x_{j_s}\|^2}{\gamma}$ is

$$\begin{aligned}\tilde{\beta}^{(s)}_{j_s} &= \mathrm{ST}\left(\tilde{\beta}^{(s-1)}_{j_s} - \tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}^\top \nabla F(X\tilde{\beta}^{(s-1)}), \tfrac{\gamma}{\|x_{j_s}\|^2}\lambda\right) \ , \\ &= \tilde{\beta}^{(s-1)}_{j_s} - \tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}^\top \nabla F(X\tilde{\beta}^{(s-1)}) - \tfrac{\gamma}{\|x_{j_s}\|^2}\lambda \operatorname{sign}(\hat{\beta}_{j_s}) \ .\end{aligned}$$

Therefore,

$$\begin{aligned}X\tilde{\beta}^{(s)} - X\tilde{\beta}^{(s-1)} &= x_{j_s}(\tilde{\beta}^{(s)}_{j_s} - \tilde{\beta}^{(s-1)}_{j_s}) \ , \\ &= -\tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}\left(x_{j_s}^\top \nabla F(X\tilde{\beta}^{(s)}) + \lambda \operatorname{sign}(\hat{\beta}_{j_s})\right)\end{aligned}$$

Using point-wise linearization of ∇F around $X\hat{\beta}$, we have, with $D \overset{\text{def.}}{=} \operatorname{diag}(f_1''(\hat{\beta}^\top \mathbf{x}_1), \dots, f_n''(\hat{\beta}^\top \mathbf{x}_n)) \in \mathbb{R}^{n\times n}$:

$$\begin{aligned}D^{1/2}X\tilde{\beta}^{(s)} = &\underbrace{\left(\mathrm{Id}_n - \tfrac{\gamma}{\|x_{j_s}\|^2} D^{1/2} x_{j_s} x_{j_s}^\top D^{1/2}\right)}_{A_s} D^{1/2} X\tilde{\beta}^{(s-1)} \\ &+ \underbrace{\tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}^\top (DX\hat{\beta}) D^{1/2} x_{j_s}}_{b_s} + o(X\tilde{\beta}^{(s)} - X\hat{\beta}) \ . \quad (30)\end{aligned}$$

Thus $(D^{1/2}X\beta^{(t)})_{t\in\mathbb{N}}$ is an asymptotical VAR sequence, and so is $(X\beta^{(t)})_{t\in\mathbb{N}}$: $X\beta^{(t+1)} = AX\beta^{(t)} + b + o(X\beta^{(t)} - X\hat{\beta})$, with $A = D^{-\frac{1}{2}} A_S \dots A_1 D^{\frac{1}{2}}$ and $b = D^{-\frac{1}{2}}(b_S + \dots + A_S \dots A_2 b_1)$.

The proof for PG follows similar ideas and is omitted due to space constraint; see [23] for the Lasso case. ∎

Theorem 10 show that we can construct an extrapolated dual point for any sparse GLM, by using extrapolation applied to the sequence $(r^{(t)} = X\beta^{(t)})_{t\in\mathbb{N}}$, with the guarantees of Proposition 7.

B. Multitask Lasso

Let $q \in \mathbb{N}$ be a number of tasks, and consider an observation matrix $Y \in \mathbb{R}^{n\times q}$, whose i-th line is $\mathbf{y}_i \in \mathbb{R}^q$. For $\mathrm{B} \in \mathbb{R}^{p\times q}$, let $\|\mathrm{B}\|_{2,1} = \sum_1^p \|\mathrm{B}_j\|$ (with $\mathrm{B}_j \in \mathbb{R}^{1\times q}$ the j-th line of B).

Definition 11. *The multitask Lasso estimator is defined as the solution of:*

$$\hat{\mathrm{B}} \in \operatorname*{arg\,min}_{\mathrm{B}\in\mathbb{R}^{n\times q}} \frac{1}{2}\|Y - X\mathrm{B}\|_F^2 + \lambda \|\mathrm{B}\|_{2,1} \ . \quad (31)$$

Although we are unable to show that $(X\beta^{(t)})_{t\in\mathbb{N}}$ is an asymptotic VAR sequence, empirical results of Section V show that dual extrapolation still provides a tighter dual point in the identified support regime. Celer empirical adaptation to multitask Lasso consists in using $d_j^{(t)} = (1 - \|x_j^\top \Theta^{(t)}\|)/\|x_j\|$ with the dual iterate $\Theta^{(t)} \in \mathbb{R}^{n\times q}$. The inner solver is cyclic block coordinate descent, and extrapolation takes $r^{(t)} \in \mathbb{R}^{nq}$ equal to the stacked columns of $X\mathrm{B}^{(t)}$. The linear combination $\sum c_k r^{(t-k)} \in \mathbb{R}^{nq}$ is mapped to $\mathbb{R}^{n\times q}$ by unstacking it.

IV. Working sets

Being able to construct a better dual point leads to a tighter gap and a smaller upper bound in Equation (6), hence to more features being discarded and a better Gap Safe screening rules. As we detail in this section, it also helps to better prioritize features, and to design an efficient working set policy.

A. Improved working sets policy

Working set methods start by solving Problem (1) restricted to a small set of features $\mathcal{W}^{(0)} \subset [p]$ (the working set), then define iteratively new working sets $\mathcal{W}^{(t)}$ and solve a sequence of growing problems [6, 27]. It is easy to see that when $\mathcal{W}^{(t)} \subsetneq \mathcal{W}^{(t+1)}$ and when the subproblems are solved up to the precision required for the whole problem, then working sets techniques converge.

It was shown by [22] that Gap Safe rules allow to define a working set policy. The value $d_j(\theta) \overset{\text{def.}}{=} \frac{1-|x_j^\top \theta|}{\|x_j\|}$ can be seen as measuring the importance of feature j, and so given an initial size $p^{(1)}$ the first working set can be defined as:

$$\mathcal{W}^{(1)} = \{j_1^{(1)}, \dots, j_{p^{(1)}}^{(1)}\} \ ,$$

where the selected features are the indices of the $p^{(1)}$ smallest values of $d(\theta)$. New primal and dual iterates are returned as solution of the first subproblem, which allow to recompute d_j's and define iteratively:

$$\mathcal{W}^{(t+1)} = \{j_1^{(t+1)}, \dots, j_{p^{(t+1)}}^{(t+1)}\} \ , \quad (32)$$

where we impose $d_j(\theta) = -1, \forall j \in \mathcal{W}^{(t)}$ to ensure nested working sets, *i.e.,* $\mathcal{W}^{(t)} \subset \mathcal{W}^{(t+1)}$. Combined with CD as an inner solver, this algorithm was coined Celer (Constraint Elimination for the Lasso with Extrapolated Residuals). The results of Section III justify the use of dual extrapolation for any sparse GLM, thus enabling us to generalize Celer to the whole class of models (Line 16).

Theorem 12. *Celer as defined in Line 16 converges as long as the inner solver converges.*

Proof Since by construction $\mathcal{W}^{(t)} \subset \mathcal{W}^{(t+1)}$ and $|\mathcal{W}^{(t+1)}| = \max(2|\mathcal{W}^{(t)}|, p)$, if $t \geq (\log p - \log p^{(1)})/\log 2 + 1$, then the working set contains all features. Since subproblems are solved to precision ϵ, this guarantees convergence. ∎

However, using a monotonic growth may lead to too large working sets, especially if the first size $p^{(1)}$ is chosen too big. Solving all subproblems to precision ϵ may also be a waste of computing time. In practice, as in [23], we introduce a simple WS variant coined *pruning*: the growth policy is $p^{(t+1)} = \min(p, 2\|\beta^{(t)}\|_0)$, and the stopping criterion for the inner solver on $\mathcal{W}^{(t)}$ is to reach a gap lower than a fraction ρ of the duality gap for the whole problem, $\mathcal{P}(\beta^{(t)}) - \mathcal{D}(\theta^{(t)})$. In practice, we set $\rho = 0.3$.

B. Newton-Celer

For the Lasso and multitask Lasso, the Hessian does not dependent on the current iterate. This is however not true for

Algorithm 1 Celer for Problem (1)

input : $X, y, \lambda, \beta^{(0)}$
param: $p_{\text{init}} = 100, \epsilon, \text{max\_it}$
init : $\theta^{(0)} = \theta^{(0)}_{\text{inner}} = \mathbf{0}_n$, $\mathcal{W}^{(0)} = \emptyset$
1 **if** $\beta^{(0)} \neq \mathbf{0}_p$ **then** $p^{(1)} = |\mathcal{S}(\beta^{(0)})|$ // warm start
2 **else** $p^{(1)} = p_{\text{init}}$
3 **for** $t = 1, \dots, \text{max\_it}$ **do**
4 compute $\theta^{(t)}_{\text{res}}$
5 $\theta^{(t)} = \arg\max_{\theta \in \{\theta^{(t-1)}, \theta^{(t-1)}_{\text{inner}}, \theta^{(t)}_{\text{res}}\}} \mathcal{D}(\theta)$
6 $g^{(t)} = \mathcal{P}(\beta^{(t-1)}) - \mathcal{D}(\theta^{(t)})$ // global gap
7 **if** $g^{(t)} \leq \epsilon$ **then** break
8 **for** $j = 1, \dots, p$ **do**
9 **if** $j \in \mathcal{W}^{(t-1)}$ **then** $d_j^{(t)} = -1$ // monotonicity
10 **else** $d_j^{(t)} = (1 - |x_j^\top \theta^{(t)}|)/\|x_j\|$
11 **if** $t \geq 2$ **then** $p^{(t)} = \min(2p^{(t-1)}, p)$
12 $\mathcal{W}^{(t)} = \left\{ j \in [p] : d_j^{(t)} \text{ among } p^{(t)} \text{ smallest values of } d^{(t)} \right\}$
// Solver is CD or prox-Newton, uses dual extrapolation:
13 get $\tilde{\beta}^{(t)}, \theta^{(t)}_{\text{inner}}$ with subproblem solver applied to $(X_{\mathcal{W}^{(t)}}, y, \lambda, (\beta^{(t-1)})_{\mathcal{W}^{(t)}}, \epsilon)$
14 set $\beta^{(t)} = \mathbf{0}_p$ and $(\beta^{(t)})_{\mathcal{W}^{(t)}} = \tilde{\beta}^{(t)}$
15 $\theta^{(t)}_{\text{inner}} = \theta^{(t)}_{\text{inner}} / \max(\lambda, \|X^\top \theta^{(t)}_{\text{inner}}\|_\infty)$
16 **return** $\beta^{(t)}, \theta^{(t)}$

other datafitting terms, *e.g.,* Logistic regression, for which taking into account the second order information proves to be very useful for fast convergence [16]. To leverage this information, we can use a prox-Newton method [20, 28] as inner solver; an advantage of dual extrapolation is that it can be combined with *any* inner solver, as we detail below. Contrary to CD, Newton steps do not lead to an asymptotic VAR, which is needed to apply dual extrapolation. To address this issue, we propose to compute K passes of cyclic CD after the Prox-Newton step. The K values of $X\beta$ obtained allow for the computation of θ_{accel} along with θ_{res}. When Line 16 is used with this method as inner solver, we refer to it as the Newton-Celer variant.

V. Experiments

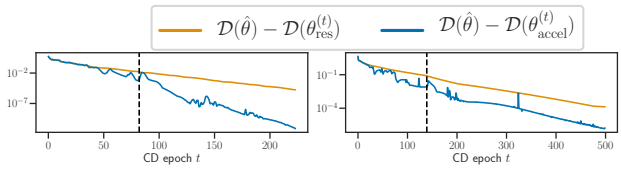
Implementation is done in Python and Cython [5] for the low-level critical parts. The solvers exactly follow the `scikit-learn` API, so that Celer can be used as a drop-in replacement in existing code. The package is available under BSD3 license at https://github.com/mathurinm/celer.

a) Lasso: Figure 1a shows the improved dual objective of θ_{accel}, after sign identification.

Figure 3 shows the time to compute a Lasso path for Celer, Gap Safe rules (w. and w/o. dual extrapolation) [12] and Blitz [17]. Dual extrapolation improves the performance of Gap Safe rules, and the working set policy of Celer makes it efficient for both dense and coarse grids of λ.

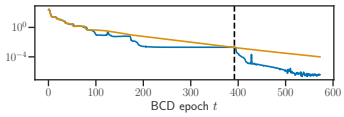
b) Logistic regression: Figure 1b shows that even for an asymptotic VAR, the dual extrapolated point θ_{accel} gives a better dual objective than the classical approach θ_{res}, after sign identification. Experiment for second order methods (Blitz, Newton-Celer) are ommitted due to space constraints.

c) Multitask Lasso: Figure 1c shows that for the Multitask Lasso, where we replace sign by support identification, the dual extrapolation still gives an improved duality gap even if we have not proved the VAR behavior of dual iterates.



(a) Lasso, on *leukemia* for $\lambda = \lambda_{\max}/5$.

(b) Logistic regression, on *leukemia* for $\lambda = \lambda_{\max}/10$.



(c) Multitask Lasso, on M/EEG data for $\lambda = \lambda_{\max}/20$.

Fig. 1: Dual objectives with classical and proposed approach, for Lasso (top), Logistic regression (middle), Multitask Lasso (bottom). The dashed lign marks sign identification (support identification for MTL)

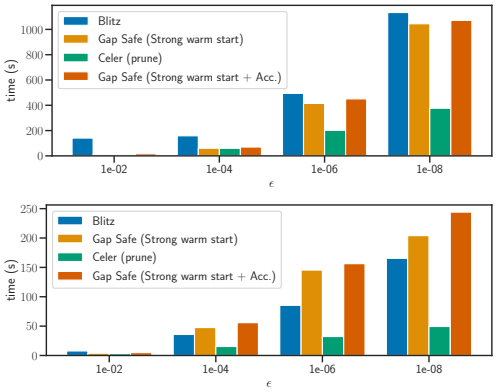


Fig. 2: Time to compute a Lasso path from $\lambda_{\max}$ to $\lambda_{\max}/100$ on the *news20* dataset. Top: grid of 100 values. Bottom: grid of 10 values ($\lambda_{\max}$ is the smallest value resulting in a 0 solution)

Figure 4 shows that the working set policy of Celer does better than Gap Safes rules with strong active warm start on magneto-electroencephalographic data from MNE (no public implementation of Blitz for this problem).

Conclusion

In this work, we generalize the dual extrapolation procedure for the Lasso (Celer) to any l_1-regularized GLM, in particular sparse Logistic regression. Theoretical guarantees based on *sign identification* of coordinate descent are provided. Experiments show that dual extrapolation yields more efficient Gap Safe screening rules and working sets solver. Finally, we adapt Celer

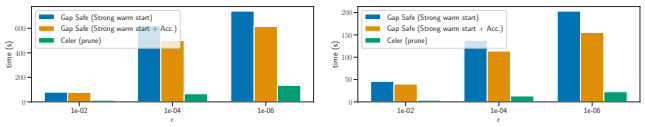


Fig. 3: Time to compute a Logistic regression path from $\lambda_{\max}$ to $\lambda_{\max}/100$ on the *news20* dataset. Top: grid of 100 values. Bottom: grid of 10 values

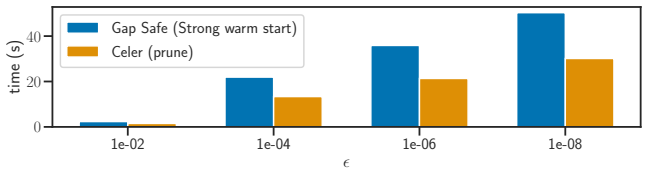


Fig. 4: Time to compute a Multitask Lasso path from $\lambda_{\max}$ to $\lambda_{\max}/100$ on the M/EEG data (grid of 10 values). $n = 305, p = 7498$.

to make it compatible with prox-Newton solvers, and empirically demonstrate its applicability to the Multi-task Lasso.

Acknowledgment

This work was funded by the ERC Starting Grant SLAB ERC-YStG-676943 and by the Chair Machine Learning for Big Data at Télécom ParisTech.

References

[1] A. Aitken. On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46:289–305, 1926.

[2] D. G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 1965.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[4] A. Beck. *First-Order Methods in Optimization*. SIAM, 2017.

[5] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith. Cython: The best of both worlds. *Computing in Science Engineering*, 13(2):31 –39, 2011.

[6] A. Boisbunon, R. Flamary, and A. Rakotomamonjy. Active set strategy for high-dimensional non-convex sparse optimization problems. In *ICASSP*, pages 1517–1521, 2014.

[7] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. A dynamic screening principle for the lasso. In *EUSIPCO*, 2014.

[8] R.P. Eddy. Extrapolating to the limit of a vector sequence. *Information Linkage between Applied Mathematics and Industry*, 1979.

[9] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[11] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.

[12] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, 2015.

[13] O. Fercoq and P. Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM J. Optim.*, 25(3):1997 – 2013, 2015.

[14] J. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

[15] E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008.

[16] C.-J Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. QUIC: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, 15:2911–2947, 2014.

[17] T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, pages 1171–1179, 2015.

[18] T. B. Johnson and C. Guestrin. A fast, principled working set algorithm for exploiting piecewise linear structure in convex problems. *arXiv preprint arXiv:1807.08046*, 2018.

[19] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *J. Mach. Learn. Res.*, 8(8):1519–1555, 2007.

[20] J. Lee, Y. Sun, and M. Saunders. Proximal Newton-type methods for convex optimization. In *NIPS*, 2012.

[21] J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, École normale supérieure de Cachan, 2010.

[22] M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster lasso-type solvers. In *NIPS-OPT*, 2017.

[23] M. Massias, A. Gramfort, and J. Salmon. Celer: a fast solver for the Lasso with dual extrapolation. In *ICML*, 2018.

[24] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18(128):1–33, 2017.

[25] G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[26] V. Roth and B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML*, pages 848–855, 2008.

[27] M. De Santis, S. Lucidi, and F. Rinaldi. A fast active set block coordinate descent algorithm for ℓ_1-regularized least squares. *SIAM J. Optim.*, 26(1):781–809, 2016.

[28] K. Scheinberg and X. Tang. Complexity of inexact proximal Newton methods. *arXiv preprint arxiv:1311.6547*, 2013.

[29] D. Scieur. *Acceleration in Optimization*. PhD thesis, École normale supérieure, 2018.

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

[31] R. Tibshirani, J. Bien, J. Friedman, T. J. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(2):245–266, 2012.

[32] R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013.

[33] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[34] J. Wang, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. *arXiv preprint arXiv:1211.3966*, 2012.

[35] Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening tests for lasso problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99), 2016.

[36] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.