Résolution rapide de problèmes de type Lasso: des règles de Safe Screening aux Working sets

Mathurin MASSIAS, Alexandre GRAMFORT, Joseph SALMON

[1]LTCI, Télécom ParisTech, Université Paris Saclay
37 rue Barrault, 75013 Paris, France
prénom.nom@telecom-paristech.fr

Résumé – Les régularisations convexes parcimonieuses sont un outil de choix pour régulariser les problèmes inverses en imagerie ou en traitement du signal, ainsi que dans de nombreuses applications d'apprentissage statistique. Par construction, elles produisent des solutions possédant peu de coefficients non-nuls, qui correspondent à des contraintes saturées dans le problème dual. Les stratégies de Working Sets (WS) constituent une famille de méthodes d'optimisation qui consistent à résoudre une série de problèmes plus simples, en considérant seulement un sous-ensemble de contraintes, ces dernières formant le WS. Ces méthodes comportent donc deux itérations imbriquées : la boucle externe correspond à la définition du WS et la boucle interne consiste à résoudre le sous-problème de manière itérative.

Dans le cas du Lasso, un WS est un ensemble de prédicteurs, et dans celui du Group Lasso c'est un ensemble de groupes. Dans cet article, nous montrons que la règle de Gauss-Southwell (une stratégie glouton pour la descente par bloc de coordonnées) permet l'implémentation de solvers rapides dans ce cas. Combinée à une stratégie de WS reposant sur un usage agressif des règles de Gap Safe screening, nous proposons un algorithme obtenant des performances du niveau de l'état de l'art pour des problèmes parcimonieux. Les résultats sont notamment présentés pour un problème inverse d'imagerie cérébrale, la localisation de sources à partir de données magneto- et électroencéphalographiques (M/EEG).

Abstract – Convex sparsity-promoting regularizations are ubiquitous to regularize inverse problems in imaging or signal processing as well as for machine learning applications. By construction, they yield solutions with few non-zero coefficients, which correspond to saturated constraints in the dual optimization formulation. Working set (WS) strategies are generic optimization techniques that consist in solving simpler problems that only consider a subset of constraints, whose indices form the WS. Working set methods therefore involve two nested iterations: the outer loop corresponds to the definition of the WS and the inner loop calls a solver for the subproblems. For the Lasso estimator a WS is a set of features, while for a Group Lasso it refers to a set of groups. In practice,Here we show that the Gauss-Southwell rule (a greedy strategy for block coordinate descent techniques) leads to fast solvers in this case. Combined with a WS strategy based on an aggressive use of so-called Gap Safe screening rules, we propose a solver achieving state-of-the-art performance on sparse learning problems. Results are presented on an inverse problem relevant for neuroscience, namely the problem of neural source localization from magneto- and electroencephalography (M/EEG).

1 Introduction

Les régularisations de type ℓ_1 ont eu un impact considérable en traitement du signal, tant en terme d'applications que de résultats théoriques [4]. De telles régularisations ont cependant un prix, puisqu'elles nécessitent d'utiliser des algorithmes spécifiques pour résoudre des problèmes contraints ou non-lisses en grande dimension [1]. Cependant, la résolution de ces problèmes peut être accélérée en utilisant la parcimonie de leur solution.

De nombreuses stratégies d'optimisation ont été proposées pour d'accélérer la résolution de problèmes comme le Lasso, la régression logitique régularisée ℓ_1, le Lasso multi-tâches ou le Group Lasso avec régularisation ℓ_1/ℓ_2 [12]. Nous regrouperons ces problèmes sous l'appellation "de type Lasso" [1]. Pour ces problèmes, les algorithmes de descente par (bloc de) coordonnées (BCD), qui consistent à modifier séquentiellement la valeur d'une coordonnée ou d'un bloc de coordonnées, ont rencontré un succès massif [15]. Il existe différentes règles de choix de bloc pour la BCD : cyclique, aléatoire ou gloutonne (connue sous le nom de Gauss-Southwell (GS) [13]).

Une idée récurrente pour accélerer la résolution des problèmes de type Lasso est de réduire la taille du problème traité. Cette idée est notamment au coeur des *strong rules* [14], et figure également dans des travaux récents comme BLITZ [8]. Parallèlement à ces méthodes de WS, qui utilisent un algorithme de BCD pour résoudre une suite de sous-problèmes de taille réduite, les *safe rules* ont été proposées par [5]. Alors qu'un algorithme de WS inclut et exclut itérativement des prédicteurs (qui forment les WS successifs), les safe rules excluent définitivement du problème les prédicteurs dès qu'il est certain qu'ils sont inutilisés à convergence. La version la plus récente de ces règles est appelée Gap Safe screening [6].

Les contributions de cet article sont 1) l'introduction d'une stratégie de WS reposant sur un usage agressif des règles de Gap Safe screening, et 2) la preuve que la règle de GS, de pair avec le précalcul de matrices de Gram, est compétitive pour la résolution des sous-problèmes en terme de temps d'éxécution.

La structure de l'article est la suivante : la Section 2 présente l'usage des règles de Gap Safe screening pour construire des WS. Nous montrons ensuite comment la règle de Gauss-Southwell peut être utilisée pour réduire les temps de calcul des sous-problèmes. Les expériences, présentées en Section 4, montrent une amélioration de l'état de l'art sur un problème inverse réel, la reconstruction de sources M/EEG.

Modèle et notations

$[d]$ désigne l'ensemble $\{1, \dots, d\}$ pour tout $d \in \mathbb{N}$. Pour tout vecteur $u \in \mathbb{R}^d$ et $\mathcal{C} \subset [d]$, le support de u est noté $\mathcal{S}_u = \{i \in [d] \,:\, u_i \neq 0\}$, $(u)_\mathcal{C}$ est le vecteur composé des éléments de u dont l'indice est dans $\mathcal{C}$, et $\bar{\mathcal{C}}$ est le complémentaire de $\mathcal{C}$ dans $[d]$. $\mathcal{S}^r_B \subset [p]$ est le support ligne de la matrice $\mathrm{B} \in \mathbb{R}^{p \times q}$ (*i.e.,* les indices des lignes non nulles de B). Soient n et $p \in \mathbb{N}$ les nombres d'observations et de prédicteurs respectivement. Soient $X \in \mathbb{R}^{n \times p}$ la matrice de design et $Y \in \mathbb{R}^{n \times q}$ la matrice des observations, où q est le nombre de tâches ou de classes du problème. La norme euclidienne (resp. Frobenius) sur les vecteurs (resp. les matrices) est noté $\|\cdot\|$ (resp. $\|\cdot\|_F$, et la i-ème ligne (resp. j-ème colonne) de B est notée $\mathrm{B}_{i,:}$ (resp. $\mathrm{B}_{:,j}$). La norme de groupe par colonne $\ell_{2,1}$ d'une matrice B s'écrit $\|\mathrm{B}\|_{2,1} = \sum_i \|\mathrm{B}_{i,:}\|$. Pour toute norme Ω, Ω_* est sa norme duale ; dans le cas de $\|\cdot\|_{2,1}$, la norme duale est ℓ_∞ / ℓ_2 : $\|B\|_{2,\infty} = \max_i \|\mathrm{B}_{i,:}\|$. $\|\mathrm{B}\|_{2,0}$ est le nombre de lignes non-nulles de B, *i.e.,* le cardinal de $\mathcal{S}^r_B$.

Le problème de régression multi-tâches que nous considérons ici est le suivant :

$$\hat{\mathrm{B}}^{(\lambda)} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \underbrace{\tfrac{1}{2} \|Y - X\mathrm{B}\|_F^2 + \lambda \Omega(\mathrm{B})}_{\mathcal{P}^{(\lambda)}(\mathrm{B})} \;, \tag{1}$$

avec $\lambda > 0$ le paramètre contrôlant l'impact de la régularisation. Le problème dual associé est (voir par exemple [9])

$$\hat{\Theta}^{(\lambda)} = \underset{\Theta \in \Delta_X}{\arg\max} \underbrace{\tfrac{1}{2} \|Y\|_F^2 - \tfrac{\lambda^2}{2} \left\|\Theta - \tfrac{Y}{\lambda}\right\|_F^2}_{\mathcal{D}^{(\lambda)}(\Theta)} \;, \tag{2}$$

où $\Delta_X = \{\Theta \in \mathbb{R}^{n \times q} \,:\, \Omega_*(X^\top \Theta) \leq 1\}$. Le saut de dualité pour (1) est $\mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta) := \mathcal{P}^{(\lambda)}(\mathrm{B}) - \mathcal{D}^{(\lambda)}(\Theta)$. Lorsque la dépendance en X est nécessaire, nous écrivons $\mathcal{G}^{(X,\lambda)}(\mathrm{B}, \Theta)$ à la place de $\mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)$.

2 Du screening aux working sets

Le principe des règles de screening sûr est d'éliminer certains prédicteurs du problème (1) dès qu'il est possible de garantir que les coefficients associés seront nuls à convergence. Pour la régression multi-tâches, les règles de Gap Safe screening proposées par [6] sont les suivantes : pour tout couple de variables primale et duale (faisable) (B, Θ), le prédicteur j peut être retiré du problème d'optimisation (1) si :

$$\left\|X_{:,j}^\top \Theta\right\| + \|X_{:,j}\| \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} < 1 \;, \tag{3}$$

de manière équivalente, il est nécessaire de considérer j si :

$$d_j(\Theta) = \frac{1 - \left\|X_{:,j}^\top \Theta\right\|}{\|X_{:,j}\|} \leq \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} \;. \tag{4}$$

En d'autres termes, le saut de dualité permet de définir un seuil auquel $d_j(\Theta)$ doit être comparé pour savoir s'il est possible de ne plus considérer le prédicteur j. Une idée naturelle pour éliminer un plus grand nombre de prédicteurs est de sacrifier la sûreté de cette règle et d'utiliser les d_j pour quantifier l'importance de chaque prédicteur. Ainsi, on peut introduire $r \in [0, 1]$ et résoudre un sous-problème en se limitant aux prédicteurs tels que :

$$d_j(\Theta) \leq r \sqrt{\frac{2}{\lambda^2} \mathcal{G}^{(\lambda)}(\mathrm{B}, \Theta)} \;. \tag{5}$$

Cette méthode de construction d'un WS peut être utilisée dans une stratégie itérative de la manière suivante. À partir d'une initialisation de B^0 (*e.g.,* $0 \in \mathbb{R}^{p \times q}$ ou une valeur approchée pour un λ' proche [1]), il est possible d'obtenir un point dual faisable Θ_0 [6].

À l'aide de la paire primale-duale (B_0, Θ_0), il est possible de calculer d_j pour tous les prédicteurs et de sélectionner ceux qui formeront le premier WS, $\mathcal{W}_1$. Ensuite, un *solveur interne* est utilisée pour résoudre de manière approchée le problème restreint à $\mathcal{W}_1$. L'itération pour cette procédure est la suivante : le solveur interne retourne une paire primale-duale $(\tilde{B}_t, \xi_t) \in \mathbb{R}^{p_t \times q} \times \mathbb{R}^{n \times q}$, où p_t est le cardinal de $\mathcal{W}_t$, à partir de laquelle on obtient une paire (B_t, ξ_t) définie par $(\mathrm{B}_t)_{\mathcal{W}_t,:} = \tilde{\mathrm{B}}_t$ and $(\mathrm{B}_t)_{\bar{\mathcal{W}}_t,:} = 0$. Θ_t est ensuite calculé à partir de ξ_t et de Θ_{t-1} de la même manière que dans [8].

Nous expliquons à présent comment utiliser les $d_j(\Theta)$ pour construire $\mathcal{W}_t$, le WS à l'itération t. Une première stratégie est d'introduire un paramètre r et d'inclure dans $\mathcal{W}_t$ les prédicteurs satisfaisant (5). Cependant, cette stratégie ne permet pas de contrôler la taille de $\mathcal{W}_t$. Nous utilisons donc plutôt une variante, qui consiste à limiter la taille de $\mathcal{W}_t$. En imposant que le cardinal de $\mathcal{W}_t$ soit au plus le double de celui de $\mathcal{S}^r_{B_{t-1}}$, nous incluons dans $\mathcal{W}_t$ les blocs ayant leur indice dans $\mathcal{S}^r_{B_{t-1}}$ et ajoutons ceux de $\bar{\mathcal{S}}^r_{B_{t-1}}$ qui ont le plus petit $d_j(\Theta_t)$.

Cette stratégie est résumée dans l'Algorithme 1.

Maintenant que nous avons défini la stratégie de construction des WS, nous pouvons nous pencher sur le choix du solveur interne utilisé pour minimiser (1) restreint aux prédicteurs de $\mathcal{W}_t$.

3 Descente par bloc de coordonnées

Nous nous intéressons à présent au choix de l'algorithme de résolution des sous-problèmes, une fois les WS définis. La fonction à minimiser est $\mathcal{P}^{(\lambda)}(\mathrm{B}) = f(\mathrm{B}) + \lambda \sum_{j=1}^p \|\mathrm{B}_j\|$, avec $f(\cdot) = \|Y - X\cdot\|_F^2 / 2$. Pour cette section, $\mathrm{B}_j \in \mathbb{R}^{1 \times q}$ désigne

1. Nous nous limitons au cas $\lambda \leq \lambda_{\max} = \left\|X^\top Y\right\|_{2,1}$, car 0 est une solution triviale de (1) autrement.

TABLE 1 – Coût de calcul (première ligne) pour la mise à jour d'un bloc avec différentes stratégies de BCD (cyclique, GS-r avec et sans calcul de la matrice de Gram). Coût en mémoire (seconde ligne) pour des mises à jours efficaces. Ces coûts correspondent au stockage des coefficients (pq), des constantes L_j (p), le stockage éventuel de la matrice de Gram (p^2) et la valeur des gradients. (pq) ou des résidus (nq).

Stratégie BCD	Cyclic	GS-r	Cyclique (Gram)	GS-r (Gram)	GS-rB (Gram)
Calcul	nq ou $2nq$	npq ou $npq + nq$	q ou $q + pq$	$2pq$	$(p + B)q$
Stockage	$nq + p + pq$	$nq + p + pq$	$2pq + p + p^2$	$2pq + p + p^2$	$2pq + p + p^2$

Algorithm 1 AGRESSIVE SCREENING W. WORKING SET

input : X, Y, λ
param: $p_0 = 100, \xi_0 = Y/\lambda, \Theta_0 = 0_{n,q}, \mathrm{B}_0 = 0_{p,q}$, $\bar{\epsilon} = 10^{-6}, \underline{\epsilon} = 0.3$
for $t = 1, \dots, T$ **do**
  $\alpha_t = \max\{\alpha \in [0,1] : (1-\alpha)\Theta_{t-1} + \alpha\xi_{t-1} \in \Delta_X\}$
  $\Theta_t = (1-\alpha_t)\Theta_{t-1} + \alpha_t\xi_{t-1}$
  $g_t = \mathcal{G}^{(X,\lambda)}(\mathrm{B}_{t-1}, \Theta_t)$ `// global gap`
  if $g_t \le \bar{\epsilon}$ **then**
    Break
  for $j = 1, \dots, p$ **do**
    Compute $d_j^t = (1 - \|X_{:,j}^\top \Theta_t\|)/\|X_{:,j}\|$
  Set $(d^t)_{\mathcal{S}^r_{\mathrm{B}_{t-1}}} = 0$ `// keep active features`
  $p_t = \max(p_0, \min(2\,\|\mathrm{B}_{t-1}\|_{2,0}, p))$ `// clipping`
  $\mathcal{W}_t = \{j \in [p] : d_j^t$ among p_t smallest values of $d^t\}$
  `// Solveur interne:`
  Get $\tilde{\mathrm{B}}_t, \xi_t \in \mathbb{R}^{p_t \times q} \times \Delta_{X_{:,\mathcal{W}_t}}$ s.t. $\mathcal{G}^{(X_{:,\mathcal{W}_t},\lambda)}(\tilde{\mathrm{B}}_t, \xi_t) \le \underline{\epsilon} g_t$
  Set $\mathrm{B}_t \in \mathbb{R}^{p \times q}$ s.t. $(\mathrm{B}_t)_{\mathcal{W}_t,:} = \tilde{\mathrm{B}}_t$ and $(\mathrm{B}_t)_{\bar{\mathcal{W}}_t,:} = 0$.
return B_t

la ligne $\mathrm{B}_{j,:}$. Dans un schéma classique de BCD, un bloc (ligne) j_k de B est choisi selon une règle, puis mis à jour suivant :

$$\mathrm{B}^k_{j_k} = \mathcal{T}_j(B) = \mathrm{prox}_{\frac{\lambda}{L_j}\cdot\|\cdot\|}\left(\mathrm{B}_j - \frac{1}{L_j}\nabla_j f(\mathrm{B})\right) \quad , \tag{6}$$

avec $L_j = \|X_{:,j}\|^2$ et, pour $z \in \mathbb{R}^q$ et $\mu > 0$,

$$\mathrm{prox}_{\mu\cdot\|\cdot\|}(z) = \mathrm{BST}(z,\mu) := \left(1 - \frac{\mu}{\|z\|}\right)_+ z \quad . \tag{7}$$

où pour tout réel a, $(a)_+ = \max(0, a)$.

3.1 Stratégies de Gauss-Southwell

En suivant la terminologie de [11], nous présentons une variante de la règle de Gauss-Southwell (GS).

Contrairement aux règles statiques comme la règle cyclique [2, 3] (où $j_k = k \pmod p$) et la règle aléatoire [10] (où j_k est tiré uniformément dans $[p]$), la règle de GS est dynamique. À chaque mise à jour, elle cherche à identifier le meilleur bloc.

La variante GS-r choisit le bloc qui maximise l'amplitude de la mise à jour :

$$j_k \in \arg\max_{j \in [p]} \left\| \mathcal{T}_j(\mathrm{B}^{k-1}) - \mathrm{B}^{k-1}_j \right\| \quad . \tag{8}$$

Pour réduire le coût de cette règle, nous utilisons une variante, GS-rB, qui cherche le meilleur prédicteur au sein de batches de taille B choisis cycliquement.

3.2 Précalcul de la matrice de Gram

La règle GS-r, puisqu'elle sélectionne le meilleur bloc à chaque itération, réduit le nombre de mises à jour nécessaire pour atteindre une précision donnée. Cependant, le calcul du choix du bloc est plus lourd, ce qui peut contrebalancer cet effet en termes de temps de calcul.

Cependant, lorsque la matrice de Gram des prédicteurs du sous-problème, $Q = X^\top X$, est précalculée (ce qui est possible puisque les sous-problèmes sont de petite taille), il devient possible de stocker à chaque itération les gradients $H^k = X^\top(X\mathrm{B}^k - Y) \in \mathbb{R}^{p \times q}$. L'étape de la mise à jour de BCD s'écrit alors :

$$\begin{cases} \delta\mathrm{B}_j & \leftarrow \mathrm{BST}\left(\mathrm{B}_j^{k-1} - \frac{1}{L_j}H_j^{k-1}, \frac{\lambda}{L_j}\right) - \mathrm{B}_j^{k-1} \\ \mathrm{B}_j^k & \leftarrow \mathrm{B}_j^{k-1} + \delta\mathrm{B}_j \quad \text{if} \quad \delta\mathrm{B}_j \neq 0 \\ H^k & \leftarrow H^{k-1} + Q_j\delta\mathrm{B}_j \quad \text{if} \quad \delta\mathrm{B}_j \neq 0 \end{cases} . \tag{9}$$

Si la mise à jour est nulle, le seul calcul requis est celui de la première ligne, qui coûte $O(q)$ puisque les gradients sont accessibles. Si la valeur de B_j est modifié, les coûts supplémentaires sont la mise à jour de B_j et et une mise à jour de rang 1 des gradients (troisième ligne). Ce faible coût rend rentable l'application de la règle GS-rB pour accélérer la résolution des sous-problèmes.

Le coût de calcul (sélection du bloc et mise à jour) pour les différentes stratégies est résumé Table 1.

4 Expériences

Dans cette section, nous présentons nos expériences sur deux jeux de données réels, pour le Lasso ($q = 1$) et le Lasso multi-tâches.

Tout d'abord, nous nous intéressons au Lasso, ce qui nous permet de nous comparer à l'état de l'art (BLITZ [8], implémenté en C++). Le saut de dualité en fonction du temps sur le jeu de données Leukemia est présenté Figure 1. Notre implémentation atteint une performance légèrement meilleure que celle de BLITZ, qui est elle-même largement supérieure à l'implémentation de scikit-learn, qui n'utilise pas de WS.

La Figure 2 montre les résultats pour le Lasso multi-tâches, appliqué à l'imagerie cérébrale magnéto-électroencéphalographique (M/EEG) [7], pour lequel Y et B sont des séries temporelles multivariées.

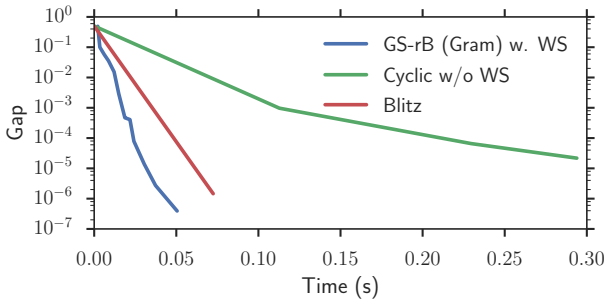

FIGURE 1 – Saut de dualité en fonction du temps pour le Lasso sur le jeu de données Leukemia ($n = 72, p = 7129$), avec $\lambda = 0.01 \left\| X^\top Y \right\|_{2,\infty}$. Les méthodes comparées sont la BCD cyclique de scikit-learn (Cyclic w/o WS), BLITZ et notre approche. Les deux approches utilisant des WS donnent des performances du même niveau, tout en surpassant largement l'algorithme de BCD cyclique.

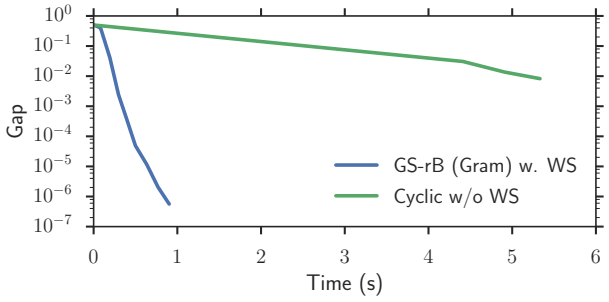

FIGURE 2 – Saut de dualité en fonction du temps pour le Lasso multi-tâches sur données M/EEG ($n = 302, p = 7498, q = 181$), avec $\lambda = 0.1 \left\| X^\top Y \right\|_{2,\infty}$. L'approche WS proposée est comparée à la BCD cyclique de scikit-learn, et la surpasse clairement.

5 Conclusion et travaux futurs

Dans cet article, nous avons proposé un lien entre les règles de Gap Safe screening et les stratégies de working set (WS) telles que BLITZ, pour résoudre des problèmes parcimonieux tels que la régression multi-tâches avec régularisation $\ell_{2,1}$. Nous avons montré que, dans le contexte de petits sous-problèmes, le calcul de la matrice de Gram permet à la règle de Gauss-Southwell (GS) d'atteindre des performances du même ordre que celles de la règle cyclique, non seulement en termes d'époques mais aussi en terme de temps de calcul. En particulier, la règle que nous baptisons GS-rB, qui restreint la recherche de la meilleure mise à jour à de petits baths cycliques a représenté le meilleur compromis. Enfin, la combinaison des stratégies de WS avec la règle de Gauss-Southwell permet d'atteindre des performances supérieures aux implémentations open source actuelles. Nous pensons qu'il est encore possible d'améliorer notre algorithme, notamment en construisant les batches de manière plus sophistiquée, ou en améliorant l'implémentation du critère d'arrêt pour le solveur interne. Enfin, l'impact de la croissance de la taille du WS pourra être étudié plus en détails.

Références

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

[2] Beck, A. and Tetruashvili, L. On the convergence of block coordinate type methods. *SIAM J. Imaging Sci.*, 23(4) : 651–694, 2013.

[3] Beck, A., Pauwels, E., and Sabach, S. The cyclic block conditional gradient method for convex optimization problems. *SIAM J. Optim.*, 25(4) :2024–2049, 2015.

[4] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37 (4) :1705–1732, 2009.

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

[6] Fercoq, O., Gramfort, A., and Salmon, J. Mind the duality gap : safer rules for the lasso. In *ICML*, pp. 333–342, 2015.

[7] Gramfort, A., Kowalski, M., and Hämäläinen, M. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Phys. Med. Biol.*, 57(7) : 1937–1961, 2012.

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

[9] Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. Gap safe screening rules for sparse multi-task and multi-class models. In *NIPS*, pp. 811–819, 2015.

[10] Nesterov, Y. Efficiency of CD methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2) :341–362, 2012.

[11] Nutini, J., Schmidt, M. W., Laradji, I. H., Friedlander, M. P., and Koepke, H. A. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *ICML*, pp. 1632–1641, 2015.

[12] Osborne, M. R., Presnell, B., and Turlach, B. A. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3) :389–403, 2000.

[13] Southwell, R. V. Relaxation methods in engineering science - a treatise on approximate computation. *The Mathematical Gazette*, 25(265) :180–182, 1941.

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

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