Fast solvers for sparse multi-task problems

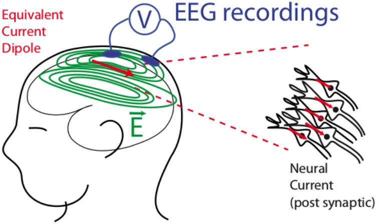
Joseph Salmon
http://josephsalmon.eu
LTCI, Télécom ParisTech, Université Paris-Saclay

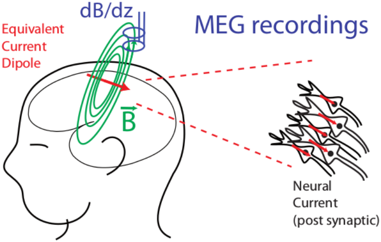
Joint work with:
Mathurin Massias (INRIA - Parietal team)
Alexandre Gramfort (INRIA - Parietal team)
+ contributions from
Eugene Ndiaye (Télécom ParisTech)
Olivier Fercoq (Télécom ParisTech)

Table of Contents

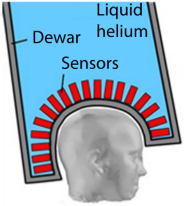
"One" motivation: M/EEG inverse problem

- sensors: magneto- and electro-encephalogram measurements during a cognitive experiment (*e.g.*, sensory or memory)
- sources: brain locations



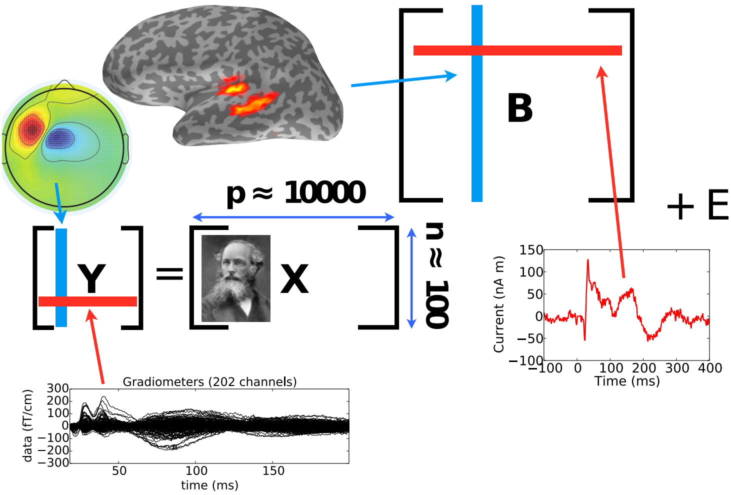


First EEG recordings in 1929 by H. Berger



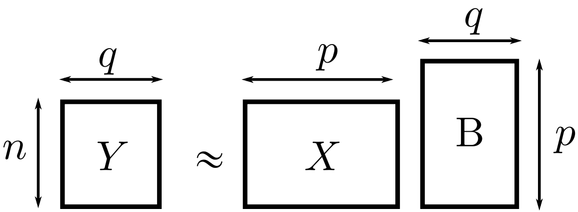
Hôpital La Timone
Marseille, France

The M/EEG inverse problem



The M/EEG inverse problem

Model: $\boxed{Y = X\mathrm{B} + E}$



- $Y \in \mathbb{R}^{n \times q}$: matrix of measurements
- $X \in \mathbb{R}^{n \times p}$: matrix describing the physics of the problem (maps sources activity to sensors through Maxwell equations)
- $\mathrm{B} \in \mathbb{R}^{p \times q}$: source activity matrix
- $E \in \mathbb{R}^{n \times q}$ additive white noise.

The M/EEG inverse problem

This problem is ill-posed ; typical values are:

- $n \approx 300$ sensors
- $p \approx 7,000$ sources
- $q \approx 200$ time instants.

$+$ biological prior: only a **few sources active simultaneously**

Vocabulary/Notation:

- $X_{:,j}$: referred to as a feature
- for simplicity: we write $\mathrm{B}_{j,:} = \mathrm{B}_j$ (j^{th} row of B)

The M/EEG inverse problem: regularization

How to infer B from $Y = X\mathrm{B} + E$?

Possible solution: penalized log-likelihood estimation

$$\hat{\mathrm{B}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \underbrace{\frac{1}{2} \left\| Y - X\mathrm{B} \right\|_F^2}_{:= f(\mathrm{B})}$$

where $f(\mathrm{B})$ is a data fitting term

$\rightarrow$ infinite number of solutions + sensitivity to noise

The M/EEG inverse problem: regularization

How to infer B from $Y = X\mathrm{B} + E$?

Possible solution: penalized log-likelihood estimation

$$\hat{\mathrm{B}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \underbrace{\frac{1}{2} \|Y - X\mathrm{B}\|_F^2}_{:= f(\mathrm{B})} + \lambda \|\mathrm{B}\|_F^2$$

where $f(\mathrm{B})$ is a data fitting term

$\rightarrow$ $\lambda > 0$ controls data fitting / regularization trade-off
$\rightarrow$ activity is spread over all sources

The M/EEG inverse problem: regularization

How to infer B from $Y = X\mathrm{B} + E$?

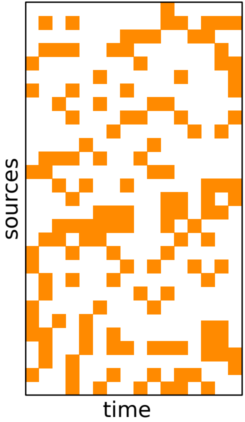
Possible solution: penalized log-likelihood estimation

$$\hat{\mathrm{B}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \underbrace{\frac{1}{2} \|Y - X\mathrm{B}\|_F^2}_{:= f(\mathrm{B})} + \underbrace{\lambda \Omega(\mathrm{B})}_{\textbf{sparsity}}$$

where $f(\mathrm{B})$ is a data fitting term

$\rightarrow$ $\lambda > 0$ controls data fitting / regularization trade-off
$\rightarrow$ activity is sparse

Choice of convex sparsity-inducing penalty Ω



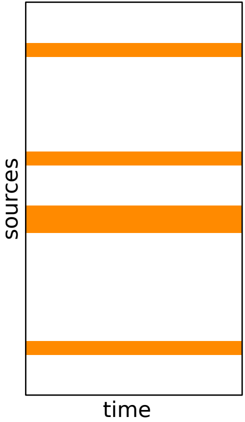
Source activity
$\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$

Penalty: ℓ_1, (Tibshirani, 1996)

$$\text{Lasso:} \quad \|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{q} |\mathrm{B}_{j,k}|$$

$\rightarrow$ weakness: activity scattered between all sources over time

Choice of convex sparsity-inducing penalty Ω



Source activity
$\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$

Penalty: $\ell_{2,1}$, (Yuan and Lin, 2006)

Group-Lasso: $\|\mathrm{B}\|_{2,1} = \sum_{j=1}^{p} \|\mathrm{B}_j\|_2$

(reminder: B_j, j-th row of B)
$\rightarrow$ we solve:

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

- Multiple Measurement Vector (MMV) in signal processing
- Multi-Task Lasso in ML

Table of Contents

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \|\mathrm{B}_j\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), ...

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \left\|\mathrm{B}_j\right\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \ldots, K$ **do**

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), ...

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \left\|\mathrm{B}_j\right\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \ldots, K$ **do**

$\quad \mathrm{B}_1^{(k)} \leftarrow \operatorname*{arg\,min}_{\mathrm{B}_1 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \mathrm{B}_2^{(k-1)}, \mathrm{B}_3^{(k-1)}, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), ...

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\|Y - X\mathrm{B}\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \|\mathrm{B}_j\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \dots, K$ **do**

$$\mathrm{B}_1^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_1 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \mathrm{B}_2^{(k-1)}, \mathrm{B}_3^{(k-1)}, \dots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_2^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_2 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2, \mathrm{B}_3^{(k-1)}, \dots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), . . .

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \left\|\mathrm{B}_j\right\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \ldots, K$ **do**

$$\mathrm{B}_1^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_1 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \mathrm{B}_2^{(k-1)}, \mathrm{B}_3^{(k-1)}, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_2^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_2 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2, \mathrm{B}_3^{(k-1)}, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_3^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_3 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2^{(k)}, \mathrm{B}_3, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), ...

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\left\|Y - X\mathrm{B}\right\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \left\|\mathrm{B}_j\right\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \ldots, K$ **do**

$$\mathrm{B}_1^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_1 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \mathrm{B}_2^{(k-1)}, \mathrm{B}_3^{(k-1)}, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_2^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_2 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2, \mathrm{B}_3^{(k-1)}, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_3^{(k)} \leftarrow \mathop{\arg\min}_{\mathrm{B}_3 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2^{(k)}, \mathrm{B}_3, \ldots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\vdots$$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), ...

(cyclic) Block Coordinate Descent

Minimize convex function: $\mathcal{P}(\mathrm{B}) = \underbrace{\frac{1}{2}\|Y - X\mathrm{B}\|_F^2}_{f(\mathrm{B})} + \lambda \sum_{j=1}^{p} \|\mathrm{B}_j\|_2$:

Algorithm: BCD

Initialization: $\mathrm{B}^{(0)} = 0 \in \mathbb{R}^{p \times q}$

for $k = 1, \dots, K$ **do**

$$\mathrm{B}_1^{(k)} \leftarrow \operatorname*{arg\,min}_{\mathrm{B}_1 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \mathrm{B}_2^{(k-1)}, \mathrm{B}_3^{(k-1)}, \dots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_2^{(k)} \leftarrow \operatorname*{arg\,min}_{\mathrm{B}_2 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2, \mathrm{B}_3^{(k-1)}, \dots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\mathrm{B}_3^{(k)} \leftarrow \operatorname*{arg\,min}_{\mathrm{B}_3 \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2^{(k)}, \mathrm{B}_3, \dots, \mathrm{B}_{p-1}^{(k-1)}, \mathrm{B}_p^{(k-1)})$$

$$\vdots$$

$$\mathrm{B}_p^{(k)} \leftarrow \operatorname*{arg\,min}_{\mathrm{B}_p \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1^{(k)}, \mathrm{B}_2^{(k)}, \mathrm{B}_3^{(k)}, \dots, \mathrm{B}_{p-1}^{(k)}, \mathrm{B}_p)$$

---

cf. (Tseng, 2001), (Friedman *et al.* , 2007), (Wu *et al.* , 2008), Nesterov (2012), (Beck *et al.* ,2013), . . .

Block updates

One does not need the exact solution of
$\arg\min_{z \in \mathbb{R}^q} \mathcal{P}(\mathrm{B}_1, \dots, \mathrm{B}_{j-1}, z, \mathrm{B}_{j+1}, \dots, \mathrm{B}_p)$, "descent" step enough

The update rule is (Kowalski, 2009) or (Parikh & Boyd, 2013)

$$\mathrm{B}_j \leftarrow \mathrm{BST}\left(\mathrm{B}_j - \frac{\nabla_j f(\mathrm{B})}{\|X_{:,j}\|^2}, \frac{\lambda}{\|X_{:,j}\|^2}\right) \ ,$$

where $\nabla_j f(\mathrm{B})$ is the gradient of the data fitting term *w.r.t.* B_j:

$$\nabla_j f(\mathrm{B}) = X_{:,j}^\top (X\mathrm{B} - Y) \ ,$$

and BST is the **block soft-thresholding** operator:

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

Rem: $(\cdot)_+ = \max(0, \cdot)$ makes the iterates block-sparse

Table of Contents

Possible Speed-ups for BCD solvers

Reminder: we expect that $\|\hat{\mathrm{B}}\|_{0,2}$ is small (few rows activated)

Idea: Can we ignore blocks which will be 0 at convergence, thus reducing the size of the problem?

- **Safe rules** (El Ghaoui *et al.* , 2012): screen out variables B_j guaranteed to be zero in $\hat{\mathrm{B}}_j$ (prior to any computation or thanks to solutions obtained for λ' close to λ)

- **Strong rules** (Tibshirani *et al.* , 2012): relaxed heuristics to start computing the $\hat{\mathrm{B}}_j$ for only a few j's

- **Working / active Set** (Joachims 1998) (Roth *et al.* , 2008), (Kim & Park, 2010), (Kowalski *et al.* , 2011), (Jonhson and Guestrin, 2015,16): solve problems with growing sizes

Dual problem detour (Kim *et al.* , 2007)

Primal function : $$\mathcal{P}(\mathrm{B}) = \frac{1}{2}\|Y - X\mathrm{B}\|^2 + \lambda\|\mathrm{B}\|_{2,1}$$

Dual problem : $$\hat{\Theta} = \underset{\Theta \in \Delta_X}{\arg\max} \underbrace{\frac{1}{2}\|Y\|^2 - \frac{\lambda^2}{2}\left\|\Theta - \frac{Y}{\lambda}\right\|^2}_{:=\mathcal{D}(\Theta)}$$

Dual feasible set : $$\Delta_X = \left\{\Theta \in \mathbb{R}^{n \times q} \ : \ \left\|X_{:,j}^\top \Theta\right\|_2 \leq 1, \forall j \in [p]\right\}$$

- $\Delta_X = \left\{\Theta \in \mathbb{R}^{n \times q} : \left\|X^\top \Theta\right\|_{2,\infty} \leq 1\right\}$ is a convex set
- The (unique) dual solution is the **projection** of Y/λ over Δ_X:

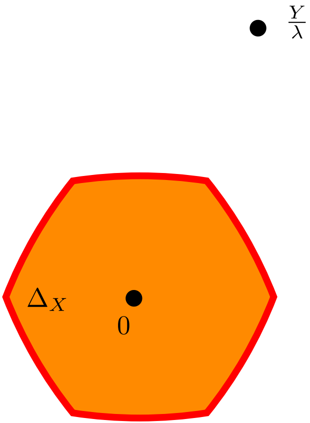
$$\hat{\Theta} = \underset{\Theta \in \Delta_X}{\arg\min} \left\|\frac{Y}{\lambda} - \Theta\right\|^2 := \Pi_{\Delta_X}\left(\frac{Y}{\lambda}\right)$$

Geometric interpretation

The dual optimal solution is the projection of Y/λ over the dual feasible set $\Delta_X = \left\{ \Theta \in \mathbb{R}^{n \times q} : \left\| X^\top \Theta \right\|_{2,\infty} \leq 1 \right\} : \hat{\Theta} = \Pi_{\Delta_X}\left(\frac{Y}{\lambda}\right)$

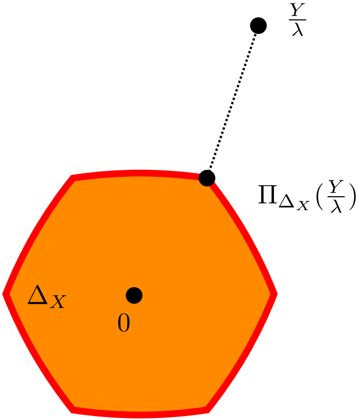
Geometric interpretation

The dual optimal solution is the projection of Y/λ over the dual feasible set $\Delta_X = \left\{ \Theta \in \mathbb{R}^{n \times q} : \left\| X^\top \Theta \right\|_{2,\infty} \leq 1 \right\} : \hat{\Theta} = \Pi_{\Delta_X}\left(\frac{Y}{\lambda}\right)$



Geometric interpretation

The dual optimal solution is the projection of Y/λ over the dual feasible set $\Delta_X = \left\{\Theta \in \mathbb{R}^{n\times q} : \left\|X^\top \Theta\right\|_{2,\infty} \leq 1\right\} : \hat{\Theta} = \Pi_{\Delta_X}\left(\frac{Y}{\lambda}\right)$



Fermat rule / KKT conditions

- **Primal solution :** $\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$
- **Dual solution :** $\hat{\Theta} \in \Delta_X \subset \mathbb{R}^{n \times q}$

Primal/Dual link: $\boxed{Y = X\hat{\mathrm{B}} + \lambda\hat{\Theta}}$

Necessary and sufficient optimality conditions:

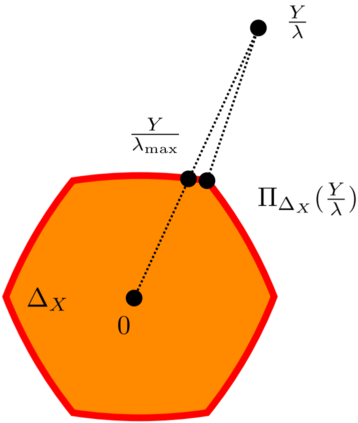
KKT/Fermat: $$\boxed{\forall j \in [p],\ X_{:,j}^\top \hat{\Theta} \in \begin{cases} \{\frac{\hat{\mathrm{B}}_j}{\|\mathrm{B}_j\|_2}\} & \text{if} \quad \hat{\mathrm{B}}_j \neq 0, \\ \mathcal{B}(0,1) & \text{if} \quad \hat{\mathrm{B}}_j = 0. \end{cases}}$$

<u>Mother of safe rules</u>:
Fermat's rule implies that if $\lambda \geq \lambda_{\max} = \|X^\top Y\|_{2,\infty}$, then $\hat{\mathrm{B}} = 0$ is a primal solution ($\hat{\Theta} = Y/\lambda$ is the associated dual)

<u>Rem</u>: $\mathcal{B}(0,1)$ is the standard Euclidean unit ball

Geometric interpretation (II)

Simple dual (feasible) point: $\frac{Y}{\lambda_{\max}} \in \Delta_X$, with $\lambda_{\max} = \|X^\top Y\|_{2,\infty}$



Rem: $(Y - X \cdot 0)/\lambda \in \Delta_X$ if $\lambda \geq \lambda_{\max}$, hence $\hat{\mathrm{B}} = Y/\lambda, \hat{\mathrm{B}} = 0$

Safe Screening rules: (El Ghaoui *et al.* , 2012)

Screening thanks to Fermat's Rule: $\boxed{\text{If } \left\| X_{:,j}^\top \hat{\Theta} \right\|_2 < 1 \text{ then, } \hat{\mathrm{B}}_j = 0}$

Safe Screening rules: (El Ghaoui *et al.* , 2012)

Screening thanks to Fermat's Rule: $\boxed{\text{If } \left\|X_{:,j}^\top \hat{\Theta}\right\|_2 < 1 \text{ then, } \hat{\mathrm{B}}_j = 0}$

Beware: $\hat{\Theta}$ is **unknown**; this is not practical !!!

Safe Screening rules: (El Ghaoui *et al.* , 2012)

Screening thanks to Fermat's Rule: $\boxed{\text{If } \left\| X_{:,j}^\top \hat{\Theta} \right\|_2 < 1 \text{ then, } \hat{\mathrm{B}}_j = 0}$

<u>Beware:</u> $\hat{\Theta}$ is **unknown**; this is not practical !!!

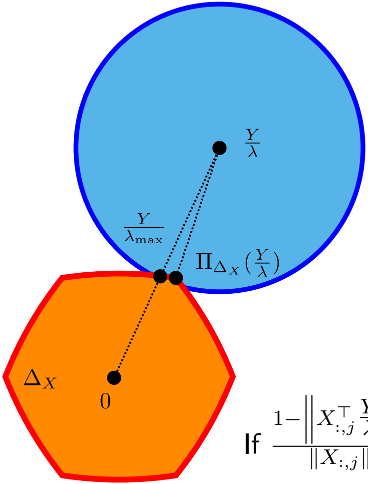
Yet, if one knows a **safe** ball $\mathcal{C}$ containing $\hat{\Theta}$, *i.e.*, a set s.t. $\hat{\Theta} \in \mathcal{C}$:

$$\textbf{safe rule}: \quad \boxed{\text{If } \sup_{\Theta \in \mathcal{C}} \left\| X_{:,j}^\top \Theta \right\|_2 < 1 \text{ then } \hat{\mathrm{B}}_j = 0} \qquad (\star)$$

<u>Rem</u>: simple bound available for the $\sup$ if $\mathcal{C}$ is a ball

<u>Rem</u>: if $(\star)$ is satisfied, you can remove $X_{:,j}$ from X and solving the smaller associated problem provide solution for the original one

Static safe rules: (El Ghaoui *et al.* , 2012)



Choose as a safe ball:

$$\mathcal{C} = \mathcal{B}\left(\frac{Y}{\lambda}, \left\|\frac{Y}{\lambda_{\max}} - \frac{Y}{\lambda}\right\|_2\right)$$

The screening test becomes:

If $\frac{1-\left\|X_{:,j}^\top \frac{Y}{\lambda}\right\|_2}{\|X_{:,j}\|} > \|Y\|_2 \left|\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right|$ then $\hat{\mathrm{B}}_j = 0$

Sequential safe rules - Strong rules

Often need $\hat{\mathrm{B}}^{(\lambda)}$ for $\lambda = \lambda_1, \lambda_2, \dots$, *e.g.*, for CV. Advice: use warm start (= start solver with closest $\hat{\mathrm{B}}^{(\lambda')}$ available)

Sequential safe rules[1] (Wang *et al.* , 2013): perform safe screening rule with the safe ball $\mathcal{C} = \mathcal{B}\left(\hat{\Theta}^{(\lambda')}, \left\|\hat{\Theta}^{(\lambda')}\right\|_2 \frac{|\lambda'-\lambda|}{\lambda'}\right)$:

$$\text{If } \frac{1 - \left\|X_{:,j}^\top \hat{\Theta}^{(\lambda')}\right\|_2}{\|X_{:,j}\|_2} > \left\|\hat{\Theta}^{(\lambda')}\right\|_2 \frac{|\lambda' - \lambda|}{\lambda'} \text{ then } \hat{\mathrm{B}}_j = 0$$

Notation: For $\lambda' > \lambda$ (close) : $\begin{cases} \hat{\mathrm{B}}^{(\lambda')} : \text{ primal optimal for } \lambda' \\ \hat{\Theta}^{(\lambda')} : \text{ dual optimal for } \lambda' \end{cases}$

[1] impossible when $\Theta^{(\lambda')}$ only approximated, Fercoq *et al.* (2015), Remark 8

Sequential safe rules - Strong rules

Often need $\hat{\mathrm{B}}^{(\lambda)}$ for $\lambda = \lambda_1, \lambda_2, \dots$, *e.g.*, for CV. Advice: use warm start (= start solver with closest $\hat{\mathrm{B}}^{(\lambda')}$ available)

Sequential safe rules[1] (Wang *et al.*, 2013): perform safe screening rule with the safe ball $\mathcal{C} = \mathcal{B}\left(\hat{\Theta}^{(\lambda')}, \left\|\hat{\Theta}^{(\lambda')}\right\|_2 \frac{|\lambda'-\lambda|}{\lambda'}\right)$:

$$\text{If } \frac{1 - \left\|X_{:,j}^\top \hat{\Theta}^{(\lambda')}\right\|_2}{\|X_{:,j}\|_2} > \left\|\hat{\Theta}^{(\lambda')}\right\|_2 \frac{|\lambda'-\lambda|}{\lambda'} \text{ then } \hat{\mathrm{B}}_j = 0$$

Strong rules (Tibshirani *et al.*, 2012): relax test to

$$\text{If } \frac{1 - \left\|X_{:,j}^\top \hat{\Theta}^{(\lambda')}\right\|_2}{\|X_{:,j}\|_2} > \frac{2}{\|X_{:,j}\|_2} \frac{|\lambda'-\lambda|}{\lambda'} \text{ then } \hat{\mathrm{B}}_j = 0$$

Notation: For $\lambda' > \lambda$ (close) : $\begin{cases} \hat{\mathrm{B}}^{(\lambda')} : \text{ primal optimal for } \lambda' \\ \hat{\Theta}^{(\lambda')} : \text{ dual optimal for } \lambda' \end{cases}$

[1] impossible when $\Theta^{(\lambda')}$ only approximated, Fercoq *et al.* (2015), Remark 8

Dynamic / Gap safe rules

Dynamic safe screening Bonnefoy *et al.* (2014, 15): perform screening at k^{th} step of a solver: periodically, screen-out variables based on Θ^k, current estimate of $\hat{\Theta}$

Gap Safe screening (Fercoq *et al.* , 2015), (Ndiaye *et al.* , 2015): same approach but rely on duality gap criterion (converging rule)

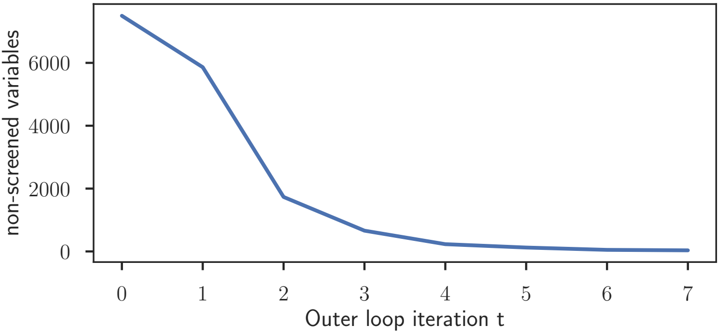
Theorem

$$d_j(\Theta^k) := \frac{1 - \left\| X_{:,j}^\top \Theta^k \right\|_2}{\|X_{:,j}\|_2} > \sqrt{\frac{2}{\lambda^2} \mathcal{G}(\mathrm{B}^k, \Theta^k)} \Rightarrow \hat{\mathrm{B}}_j = 0$$

for any primal point B^k and dual feasible point Θ^k where $\mathcal{G}(\mathrm{B}^k, \Theta^k) = \mathcal{P}(\mathrm{B}^k) - \mathcal{D}(\Theta^k)$ is the duality gap

Rem: get Θ^k by residual scaling, *i.e.*, rescale $Y - X\mathrm{B}^k$ to be in Δ_X

Can we speed up solvers more?



(Leukemia $n = 72, p = 7129, q = 1$, Lasso case)

$\rightarrow$ useless features are still included in the beginning !

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive Gap: include feature j if $d_j(\Theta^k) < r\sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\Theta)$

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive Gap: include feature j if $d_j(\Theta^k) < r\sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$ for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\Theta)$
- inner loop: solve subproblem keeping only these features (fast)

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive Gap: include feature j if $d_j(\Theta^k) < r\sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$ for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\Theta)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive Gap: include feature j if $d_j(\Theta^k) < r\sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\Theta)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Aggressive screening and working sets (WS)

Idea: To reduce the number of features, **drop the safety**.

Gap Safe: exclude source j if $d_j(\Theta^k) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$

Aggressive Gap: include feature j if $d_j(\Theta^k) < r\sqrt{\frac{2}{\lambda^2}\mathcal{G}(\mathrm{B}^k, \Theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\Theta)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Previous working/active set techniques for Lasso-type problems:
(Roth *et al.* , 2008), (Kim & Park, 2010), (Kowalski *et al.* , 2011),

AGGressive Gap Greedy with Gram (A5G)

Algorithm: A5G

input : X, Y, λ

param: $\mathrm{B}_0 = 0_{p,q}, \bar{\epsilon} = 10^{-6}, r \in]0, 1[$

```
// Outer loop:
```

for $k = 1, \dots, K$ **do**

  Compute dual point Θ^k and dual gap g^k

  if $g^k \leq \bar{\epsilon}$ **then**

    Break

  for $j = 1, \dots, p$ **do**

    Compute $d_j^k = (1 - \|X_{:,j}^\top \Theta^k\|)/\|X_{:,j}\|$

  $\mathcal{W}^k = \{j \in [p] : d_j^k < r\sqrt{2g^k}/\lambda \cup \{j : \mathrm{B}_j^{k-1} \neq 0\}$

```
  // Inner loop:
```

  Solve problem restricted to $\mathcal{W}^k$ approximately and get B^k

return B^k

<u>Rem</u>: easy to add gap safe screening once g^k and d_j^k computed

AGGressive Gap Greedy with Gram (A5G)

Algorithm: A5G

input : X, Y, λ

param: $\mathrm{B}_0 = 0_{p,q}, \bar{\epsilon} = 10^{-6}, \cancel{r \in]0,1[}, p_0 = 100$ (or other guess)

```
// Outer loop:
```

for $k = 1, \dots, K$ **do**

  Compute dual point Θ^k and dual gap g^k

  if $g^k \leq \bar{\epsilon}$ **then**

    Break

  for $j = 1, \dots, p$ **do**

    Compute $d_j^k = (1 - \|X_{:,j}^\top \Theta^k\|)/\|X_{:,j}\|$

  $p^k = \max(p_0, \min(2 \left\|\mathrm{B}^{k-1}\right\|_{2,0}, p))$ `// clipping`

  $\mathcal{W}^k = \{j \in [p] : d_j^k$ among $p^k/2$ smallest ones$\} \cup \{j : \mathrm{B}_j^{k-1} \neq 0\}$

  ```
  // Inner loop:
  ```

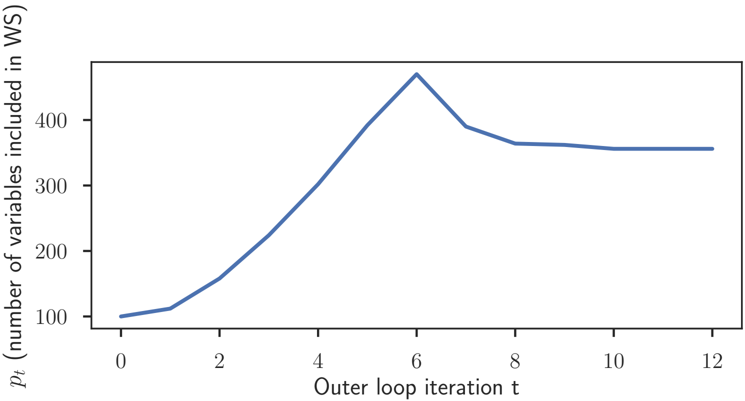
  Solve problem restricted to $\mathcal{W}^k$ approximately and get B^k

return B^k

Rem: easy to add gap safe screening once g^k and d_j^k computed

Table of Contents

Further speed improvement



(Leukemia $n = 72, p = 7129, q = 1$, Lasso case)

Smaller subproblems solved $\rightarrow$ Gram matrix $X_{\mathcal{W}^k}^\top X_{\mathcal{W}^k}$ fits in!

Fast update with pre-computed Gram

When one can pre-compute and store the Gram matrix

$$Q = X^\top X = [Q_1, \ldots, Q_p] \in \mathbb{R}^{p \times p}$$

maintain the gradients

$$H^k = X^\top (X\mathrm{B}^k - Y) \in \mathbb{R}^{p \times q}$$

rather than the residuals, and use

$$\textbf{BCD update} \;:\; \begin{cases} \delta \mathrm{B}_j & \leftarrow \mathrm{BST}\left(\mathrm{B}_j^{k-1} - \frac{H_j^{k-1}}{\|X_{:,j}\|^2}, \frac{\lambda}{\|X_{:,j}\|^2}\right) - \mathrm{B}_j^{k-1} \\ \mathrm{B}_j^k & \leftarrow \mathrm{B}_j^{k-1} + \delta \mathrm{B}_j \qquad \text{if} \quad \delta \mathrm{B}_j \neq 0 \\ H^k & \leftarrow H^{k-1} + Q_j \delta \mathrm{B}_j \quad \text{if} \quad \delta \mathrm{B}_j \neq 0 \end{cases}$$

Rem: For "non 0-updates", the main cost is a $p \times q$ matrix rank one update (the third line); "0-updates" almost free

Gauss-Southwell (GS) selection rule

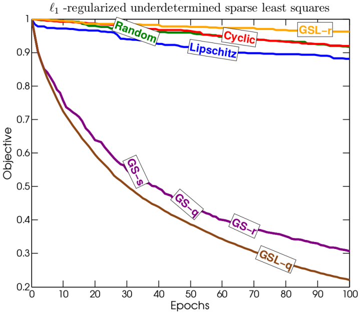


Figure: courtesy of Nutini *et al.* (2015), various BCD strategies

Greedy selection rules for BCD (Soutwhell, 1941), (Tseng & Yun, 2009) instead of cyclic:

- larger decrease in objective for each update
- but costly to compute

Gauss-Southwell (GS) selection rule

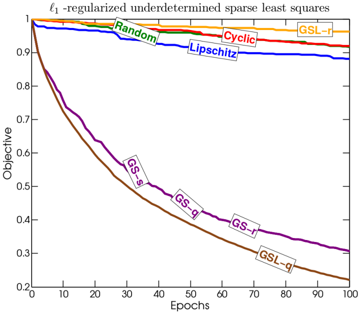
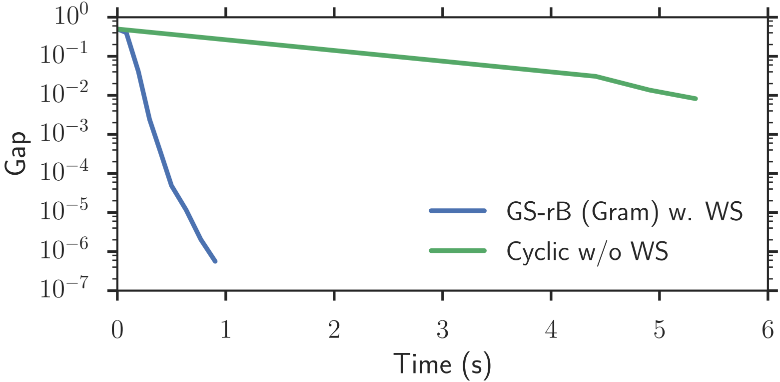


Figure: courtesy of Nutini *et al.* (2015), various BCD strategies

Greedy selection rules for BCD (Soutwhell, 1941), (Tseng & Yun, 2009) instead of cyclic:

- larger decrease in objective for each update
- but costly to compute
- way **cheaper when the Gram matrix is available !**

Results on MEG data



(MEG: $n = 302, p = 7498, q = 181$)

about $10\times$ speed-up w.r.t. state-of-the-art multi-task Lasso solver from `scikit-learn` (Pedregosa *et al.* , 2011)

Take home message

- Safe screening rules can help (**sequential,dynamic**, . . .)
- Relaxing safety / aggressive screening : provide **growth strategy** for working set (WS)

Take home message

- Safe screening rules can help (**sequential,dynamic**, ...)
- Relaxing safety / aggressive screening : provide **growth strategy** for working set (WS)
- WS with small size : **Gram precomputation** possible leads to **greedy/GS** BCD competitive in terms of **time**

Take home message

- Safe screening rules can help (**sequential,dynamic**, . . .)
- Relaxing safety / aggressive screening : provide **growth strategy** for working set (WS)
- WS with small size : **Gram precomputation** possible leads to **greedy/GS** BCD competitive in terms of **time**
- **Flexible** : can handle more cases (Sparse Group Lasso, Sparse logistic regression, etc.)

Take home message

- Safe screening rules can help (**sequential,dynamic**, ...)
- Relaxing safety / aggressive screening : provide **growth strategy** for working set (WS)
- WS with small size : **Gram precomputation** possible leads to **greedy/GS** BCD competitive in terms of **time**
- **Flexible** : can handle more cases (Sparse Group Lasso, Sparse logistic regression, etc.)

More info : Papers / Code

Papers:

- Gap Safe Rules: ICML 2015 (Lasso case), NIPS 2015 (General loss + multi-task), long version ArXiV 1611.05780

A5G:

- ArXiV 1703.07285

Codes:

- Python Code on-line: `https://github.com/EugeneNdiaye`
- pull requests (#5075) (#7853) on `sklearn`
- A5G code: to be released

Powered with **MooseTeX**

References I

- A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval.
  A dynamic screening principle for the lasso.
  In *EUSIPCO*, 2014.
- A. Beck and L. Tetruashvili.
  On the convergence of block coordinate type methods.
  SIAM J. Imaging Sci., 23(4):651–694, 2013.
- L. El Ghaoui, V. Viallon, and T. Rabbani.
  Safe feature elimination in sparse supervised learning.
  J. Pacific Optim., 8(4):667–698, 2012.
- O. Fercoq, A. Gramfort, and J. Salmon.
  Mind the duality gap: safer rules for the lasso.
  In *ICML*, pages 333–342, 2015.
- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani.
  Pathwise coordinate optimization.
  Ann. Appl. Stat., 1(2):302–332, 2007.

References II

- T. B. Johnson and C. Guestrin.
  Blitz: A principled meta-algorithm for scaling sparse optimization.
  In *ICML*, pages 1171–1179, 2015.
- T. B. Johnson and C. Guestrin.
  Unified methods for exploiting piecewise linear structure in convex optimization.
  In *NIPS*, pages 4754–4762, 2016.
- T. Joachims.
  Text categorization with support vector machines: Learning with many relevant features.
  In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.
- S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky.
  An interior-point method for large-scale l_1-regularized least squares.
  IEEE J. Sel. Topics Signal Process., 1(4):606–617, 2007.

References III

- Matthieu Kowalski.
  Sparse regression using mixed norms.
  Applied and Computational Harmonic Analysis, 27(3):303 – 324, 2009.

- J. Kim and H. Park.
  Fast active-set-type algorithms for l1-regularized linear regression.
  In *AISTATS*, pages 397–404, 2010.

- M. Kowalski, P. Weiss, A. Gramfort, and S. Anthoine.
  Accelerating ISTA with an active set strategy.
  In *OPT 2011: 4th International Workshop on Optimization for Machine Learning*, page 7, 2011.

- Y. Nesterov.
  Efficiency of coordinate descent methods on huge-scale optimization problems.
  SIAM J. Optim., 22(2):341–362, 2012.

References IV

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

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

- N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein.
  Proximal algorithms.
  Foundations and Trends in Machine Learning, 1(3):1–108, 2013.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
  Scikit-learn: Machine learning in Python.
  J. Mach. Learn. Res., 12:2825–2830, 2011.

References V

- V. Roth and B. Fischer.
  The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms.
  In *ICML*, pages 848–855, 2008.
- R. V. Southwell.
  Relaxation methods in engineering science - a treatise on approximate computation.
  The Mathematical Gazette, 25(265):180–182, 1941.
- R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani.
  Strong rules for discarding predictors in lasso-type problems.
  J. Roy. Statist. Soc. Ser. B, 74(2):245–266, 2012.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1):267–288, 1996.

References VI

- P. Tseng.
  Convergence of a block coordinate descent method for nondifferentiable minimization.
  J. Optim. Theory Appl., 109(3):475–494, 2001.
- P. Tseng and S. Yun.
  Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization.
  J. Optim. Theory Appl., 140(3):513, 2009.
- T. T. Wu and K. Lange.
  Coordinate descent algorithms for lasso penalized regression.
  Ann. Appl. Stat., pages 224–244, 2008.
- J. Wang, J. Zhou, P. Wonka, and J. Ye.
  Lasso screening rules via dual polytope projection.
  In *NIPS*, pages 1070–1078, 2013.
- M. Yuan and Y. Lin.
  Model selection and estimation in regression with grouped variables.
  J. Roy. Statist. Soc. Ser. B, 68(1):49–67, 2006.