Exploiting structure in sparse GLMs for fast & safe support identification

Mathurin Massias (INRIA)

Joint work with:
Alexandre Gramfort (INRIA)
Joseph Salmon (Université de Montpellier)
Samuel Vaiter (CNRS, UMB)

Table of Contents

Motivation: sparse inverse problems

Pedagogical example: the Lasso

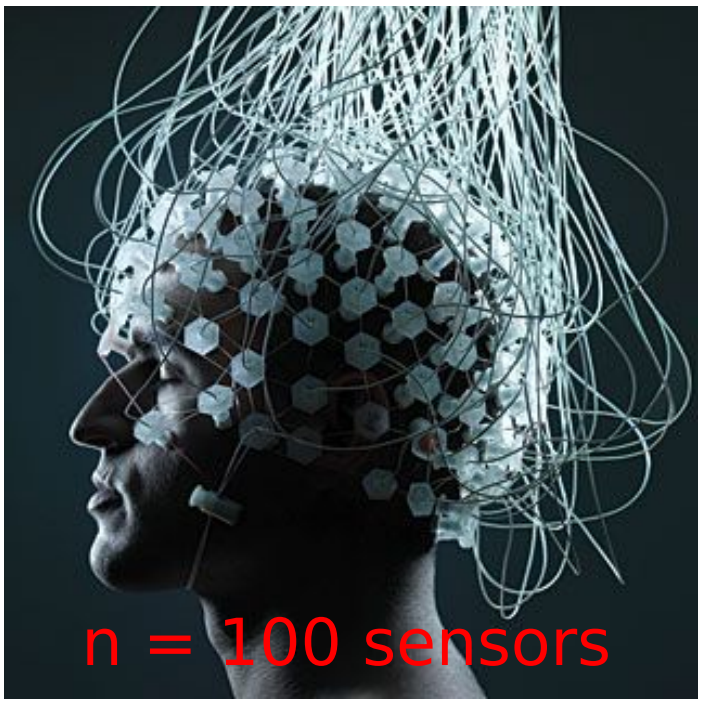
Exploiting regularity

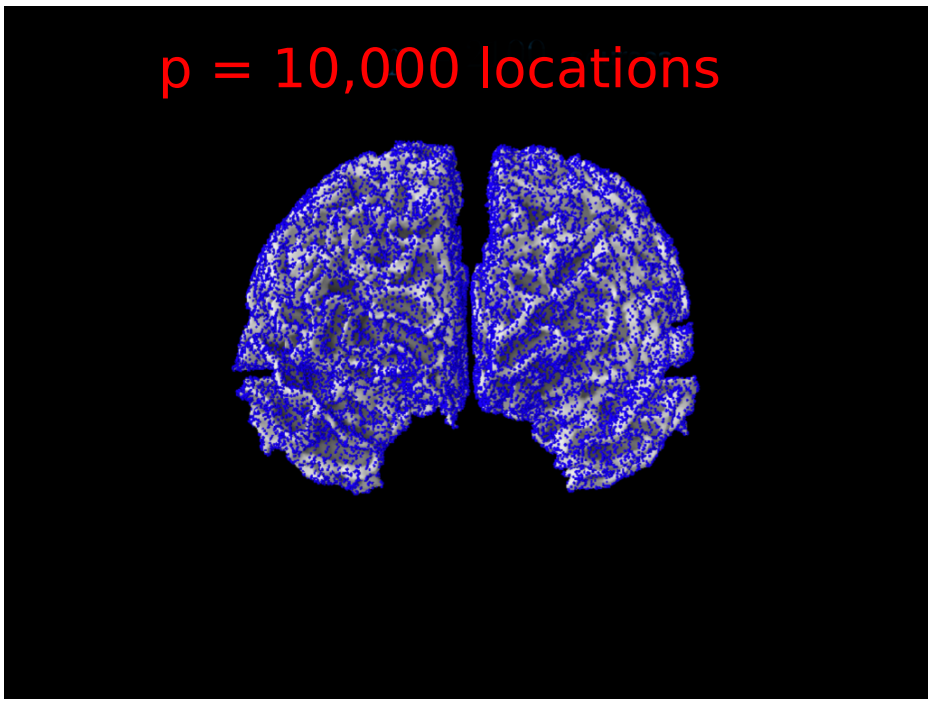
Sparse GLMs

More solvers speed-up

The M/EEG inverse problem

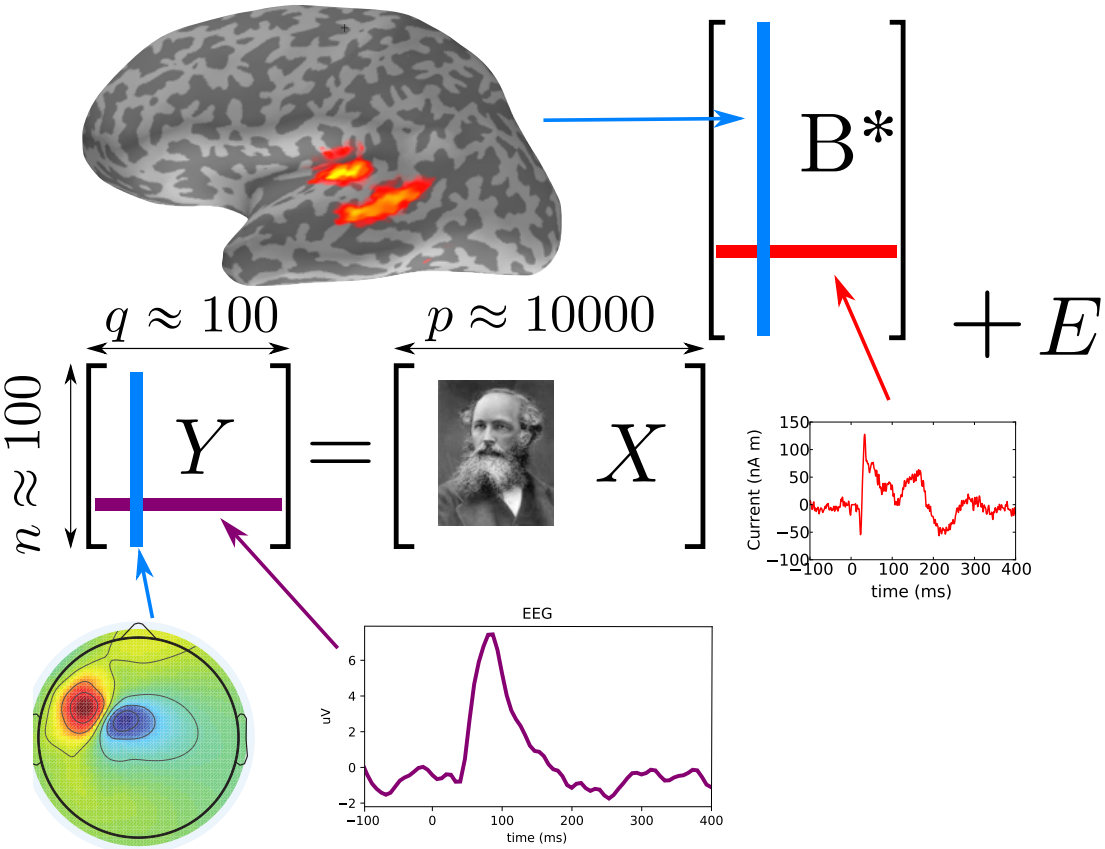
- observe magnetoelectric field outside the scalp (100 sensors)
- reconstruct cerebral activity inside the brain (10,000 locations)





Identifying the correct locations is critical (epilepsy surgery)

Mathematical model: multitask regression



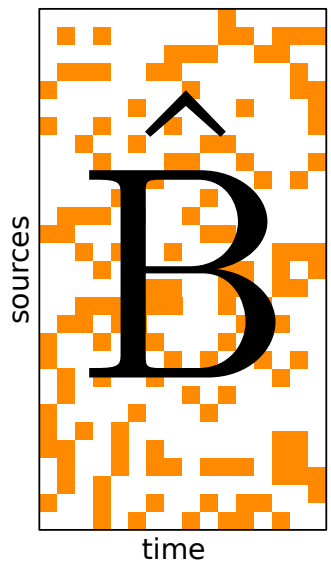
One way to estimate B^*:

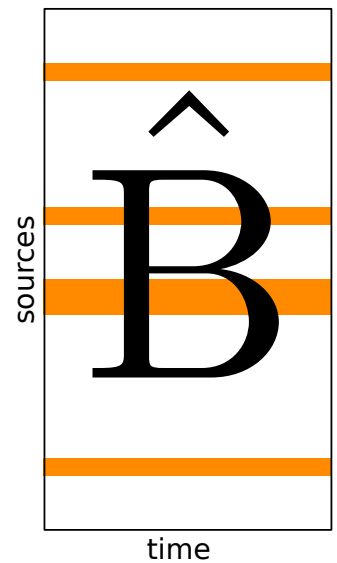
$$\hat{\mathrm{B}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \frac{1}{2} \|Y - X\mathrm{B}\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathrm{B}_{j:}\|_2$$

The $\ell_{2,1}$ penalty

$$\|\mathrm{B}\|_{1,1} = \sum_{j=1}^{p}\sum_{l=1}^{q}|\mathrm{B}_{jl}|$$

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





Our focus: identify the support of $\hat{\mathrm{B}}$ **with guarantees**

Table of Contents

Motivation: sparse inverse problems

Pedagogical example: the Lasso

Exploiting regularity

Sparse GLMs

More solvers speed-up

The Lasso[1,2]

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1}_{\mathcal{P}(\beta)}$$

- $y \in \mathbb{R}^n$: observations
- $X = [X_1 | \dots | X_p] \in \mathbb{R}^{n \times p}$: design matrix

[1] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

[2] S. S. Chen and D. L. Donoho. "Atomic decomposition by basis pursuit". In: *SPIE*. 1995.

Duality for the Lasso

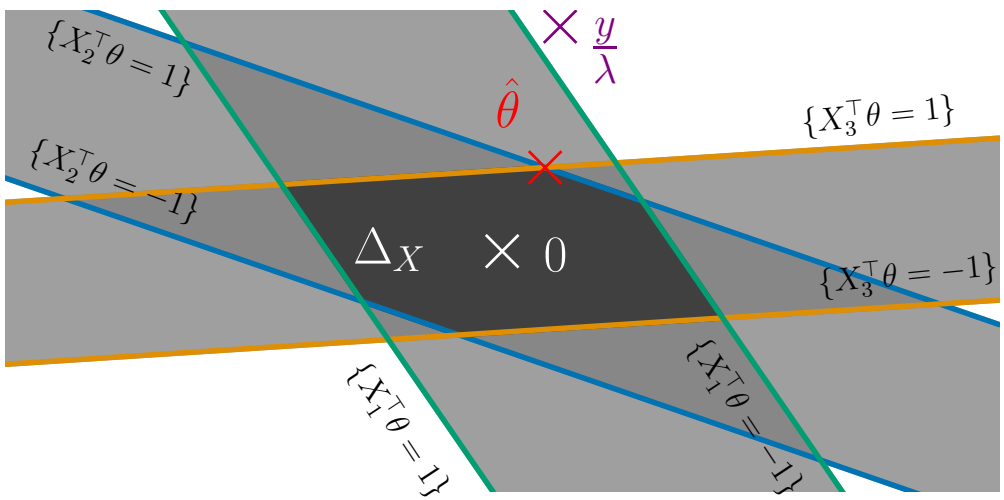
$$\hat{\theta} = \underset{\theta \in \Delta_X}{\arg\max} \underbrace{\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \|y/\lambda - \theta\|^2}_{\mathcal{D}(\theta)}$$

$\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \forall j \in [p], \ |X_j^\top \theta| \leq 1\right\}$: **dual feasible set**

Duality for the Lasso

$$\hat{\theta} = \underset{\theta \in \Delta_X}{\arg\max} \underbrace{\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \|y/\lambda - \theta\|^2}_{\mathcal{D}(\theta)}$$

$\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \forall j \in [p], \ |X_j^\top \theta| \leq 1\right\}$: **dual feasible set**

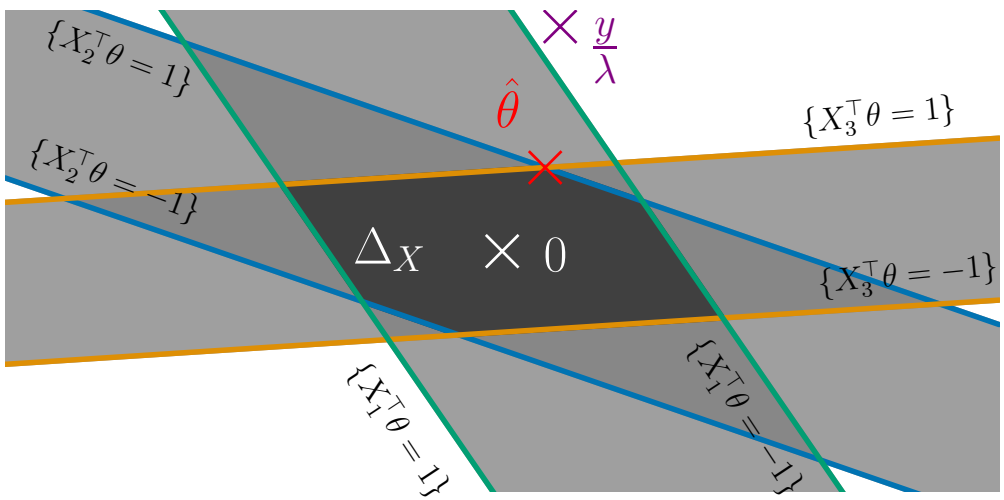


Toy visualization example: $n = 2, p = 3$

Duality for the Lasso

$$\hat{\theta} = \underset{\theta \in \Delta_X}{\arg\max} \underbrace{\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \|y/\lambda - \theta\|^2}_{\mathcal{D}(\theta)}$$

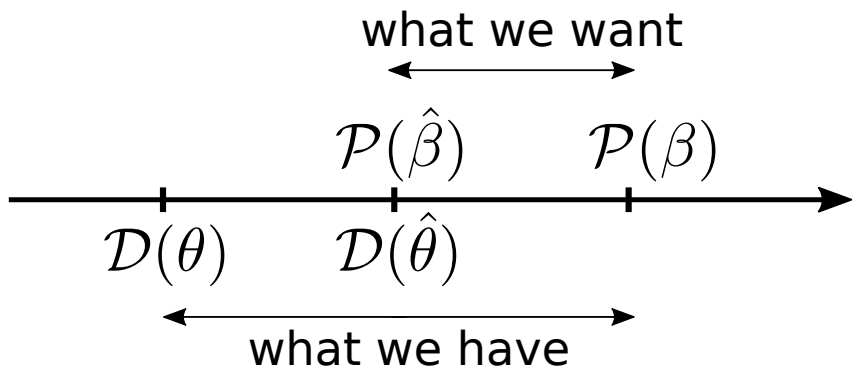
$\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \forall j \in [p], \ |X_j^\top \theta| \leq 1\right\}$: **dual feasible set**



Projection problem: $\hat{\theta} = \Pi_{\Delta_X}(y/\lambda)$

Duality gap as a stopping criterion

$$\mathcal{P}(\beta) \geq \mathcal{P}(\hat{\beta}) = \mathcal{D}(\hat{\theta}) \geq \mathcal{D}(\theta)$$



$$\forall \beta, (\exists \theta \in \Delta_X, \text{ dgap}(\beta, \theta) \leq \epsilon) \Rightarrow \mathcal{P}(\beta) - \mathcal{P}(\hat{\beta}) \leq \epsilon$$

β is an ϵ-solution whenever $\text{dgap}(\beta, \theta) \leq \epsilon$

Choice of dual point

Primal-dual link at optimum:

$$\boxed{\hat{\theta} = (y - X\hat{\beta})/\lambda}$$

Standard approach[3]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} := y - X\beta^{(t)}$, take

$$\theta = \theta^{(t)}_{\text{res}} := r^{(t)}/\lambda$$

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

Choice of dual point

Primal-dual link at optimum:

$$\boxed{\hat{\theta} = (y - X\hat{\beta})/\lambda}$$

Standard approach[3]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} := y - X\beta^{(t)}$, take

$$\theta = \theta_{\text{res}}^{(t)} := r^{(t)} / \max(\lambda, \|X^\top r^{(t)}\|_\infty)$$

residuals rescaling

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

Choice of dual point

Primal-dual link at optimum:

$$\boxed{\hat{\theta} = (y - X\hat{\beta})/\lambda}$$

Standard approach[3]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} := y - X\beta^{(t)}$, take

$$\theta = \theta_{\text{res}}^{(t)} := r^{(t)} / \max(\lambda, \|X^\top r^{(t)}\|_\infty)$$

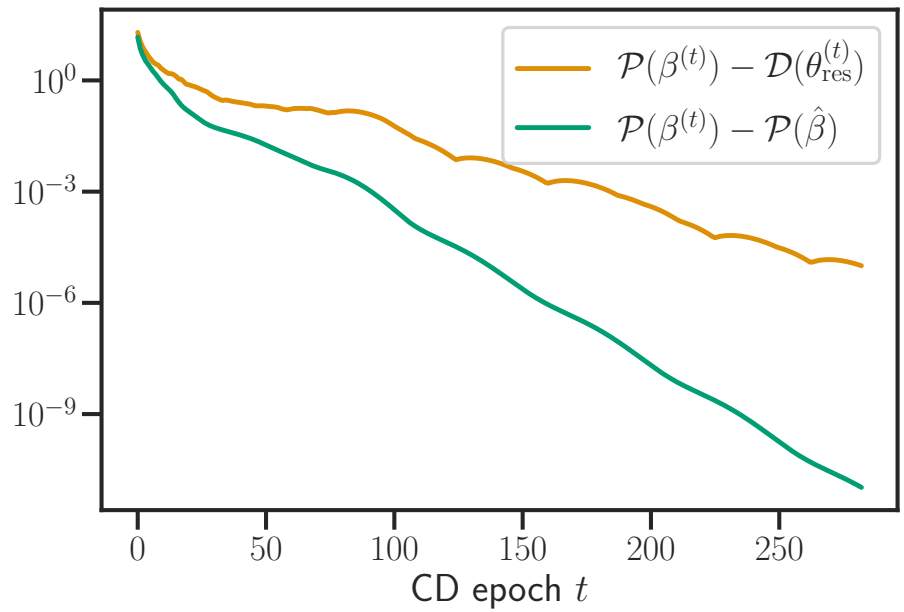
residuals rescaling

- converges to $\hat{\theta}$
- $\mathcal{O}(np)$ to compute ($=$ 1 epoch of CD)
  $\hookrightarrow$ rule of thumb: compute $\theta_{\text{res}}^{(t)}$ and dgap every 10 epochs

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

Issues with residuals rescaling

$$\theta_{\text{res}}^{(t)} = r^{(t)} / \max(\lambda, \|X^\top r^{(t)}\|_\infty)$$



Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

$\lambda_{\max} = \|X^\top y\|_\infty$ is the smallest λ giving $\hat{\beta} = 0$

Table of Contents

Motivation: sparse inverse problems

Pedagogical example: the Lasso

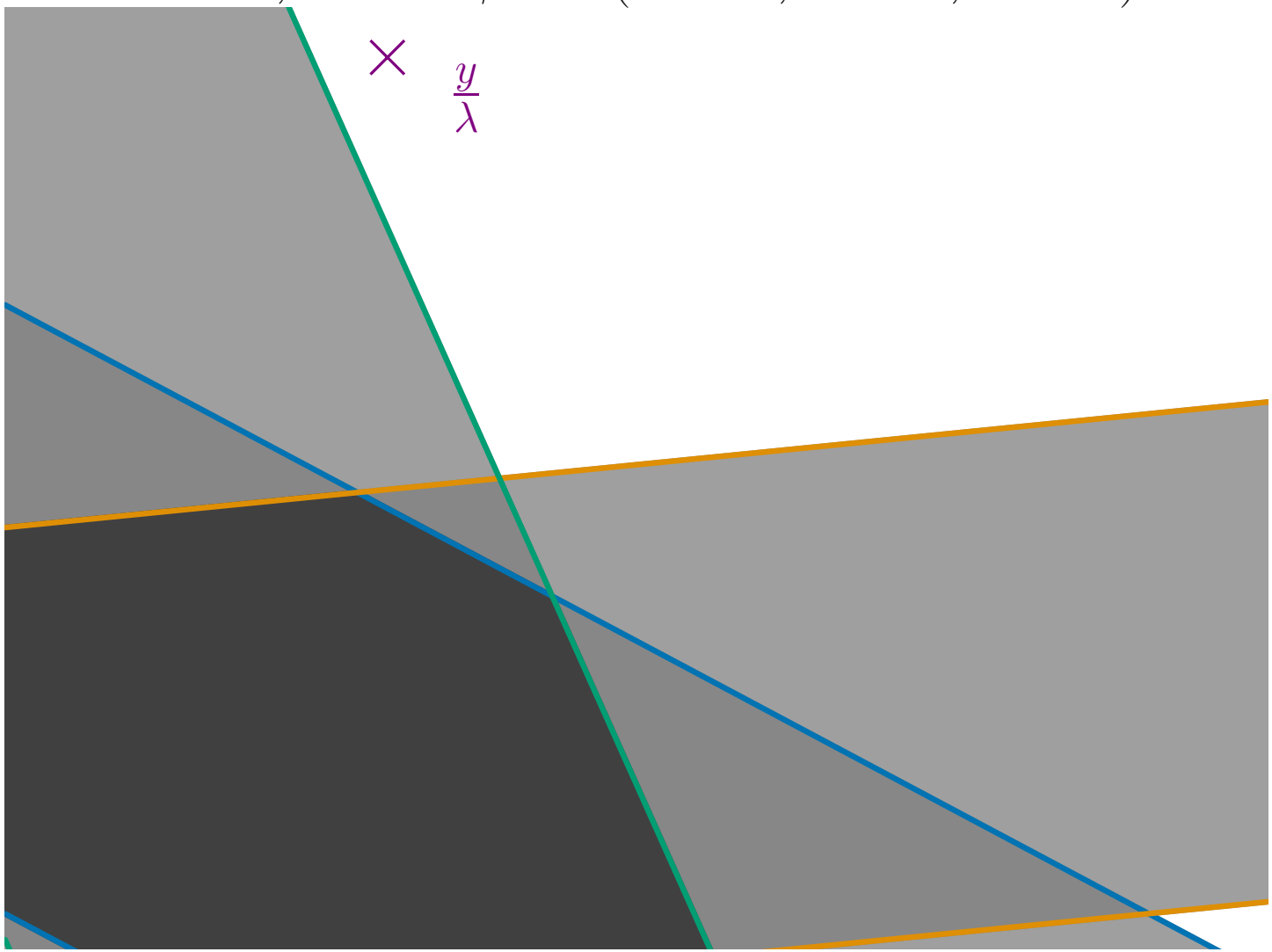
Exploiting regularity

Sparse GLMs

More solvers speed-up

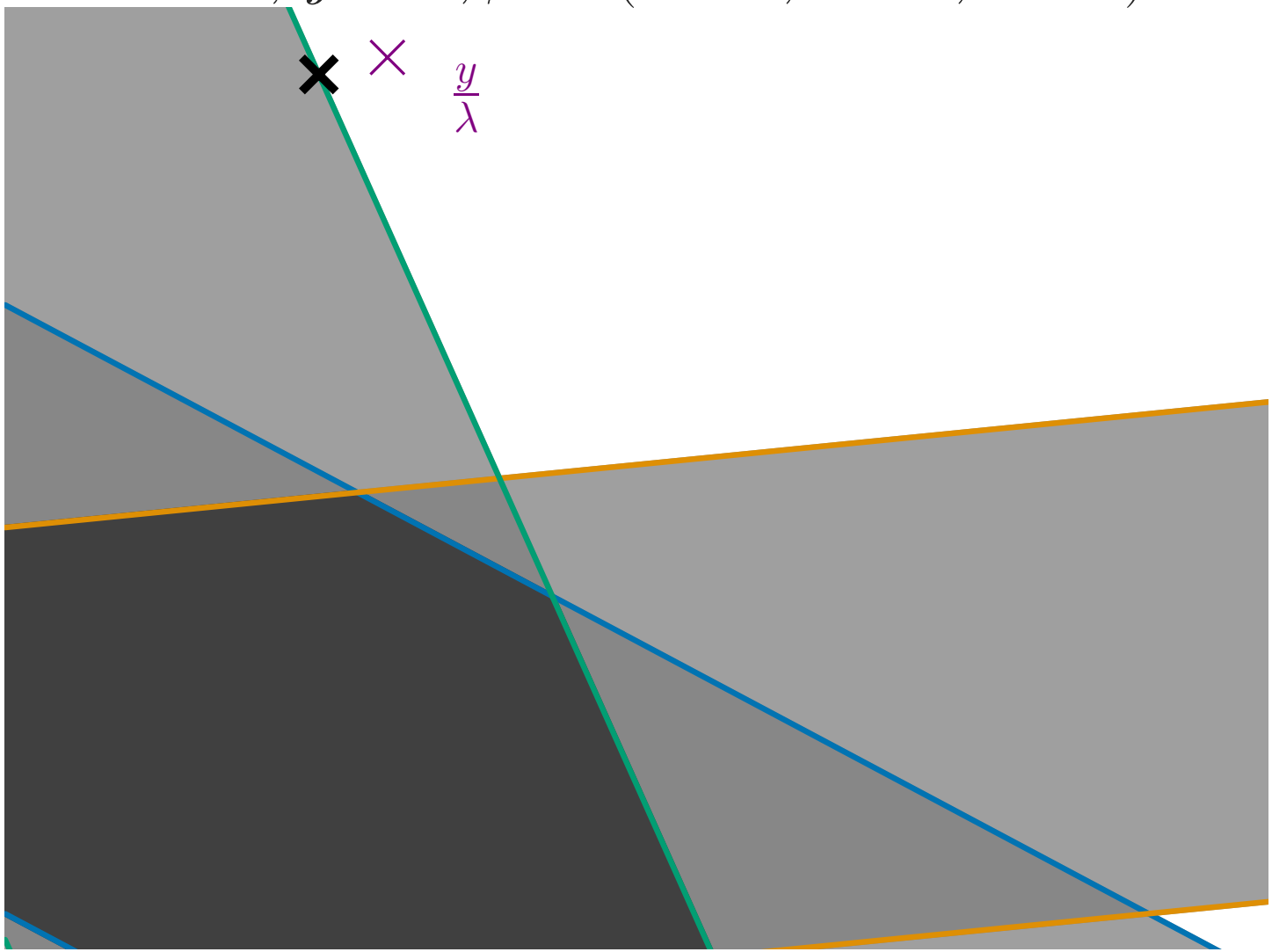
Regularity after sign identification

$t = 0, \qquad \beta = (0.000, 0.000, 0.000)$



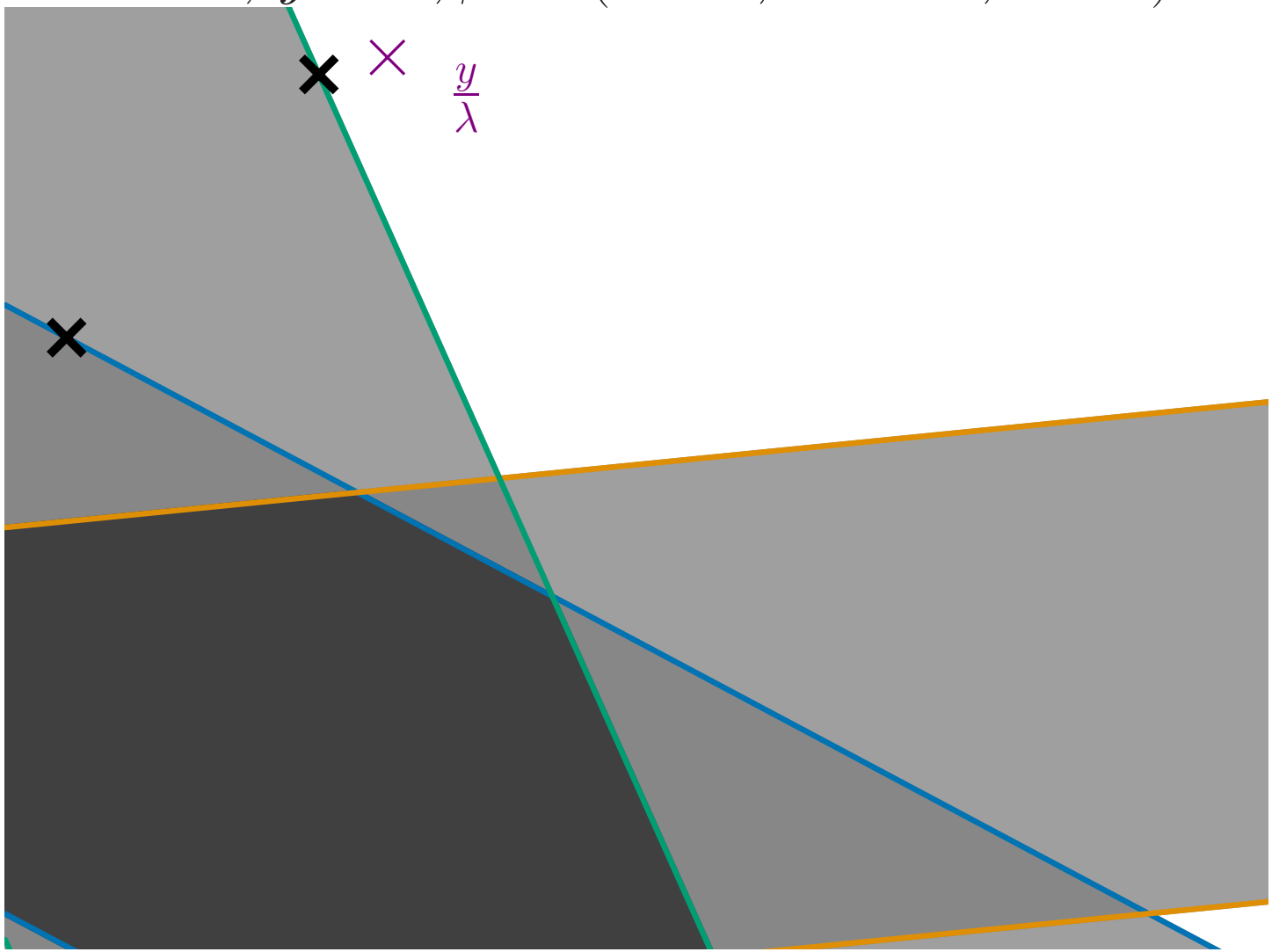
Regularity after sign identification

$$t = 1,\ j = 1, \beta = (0.217, 0.000, 0.000)$$



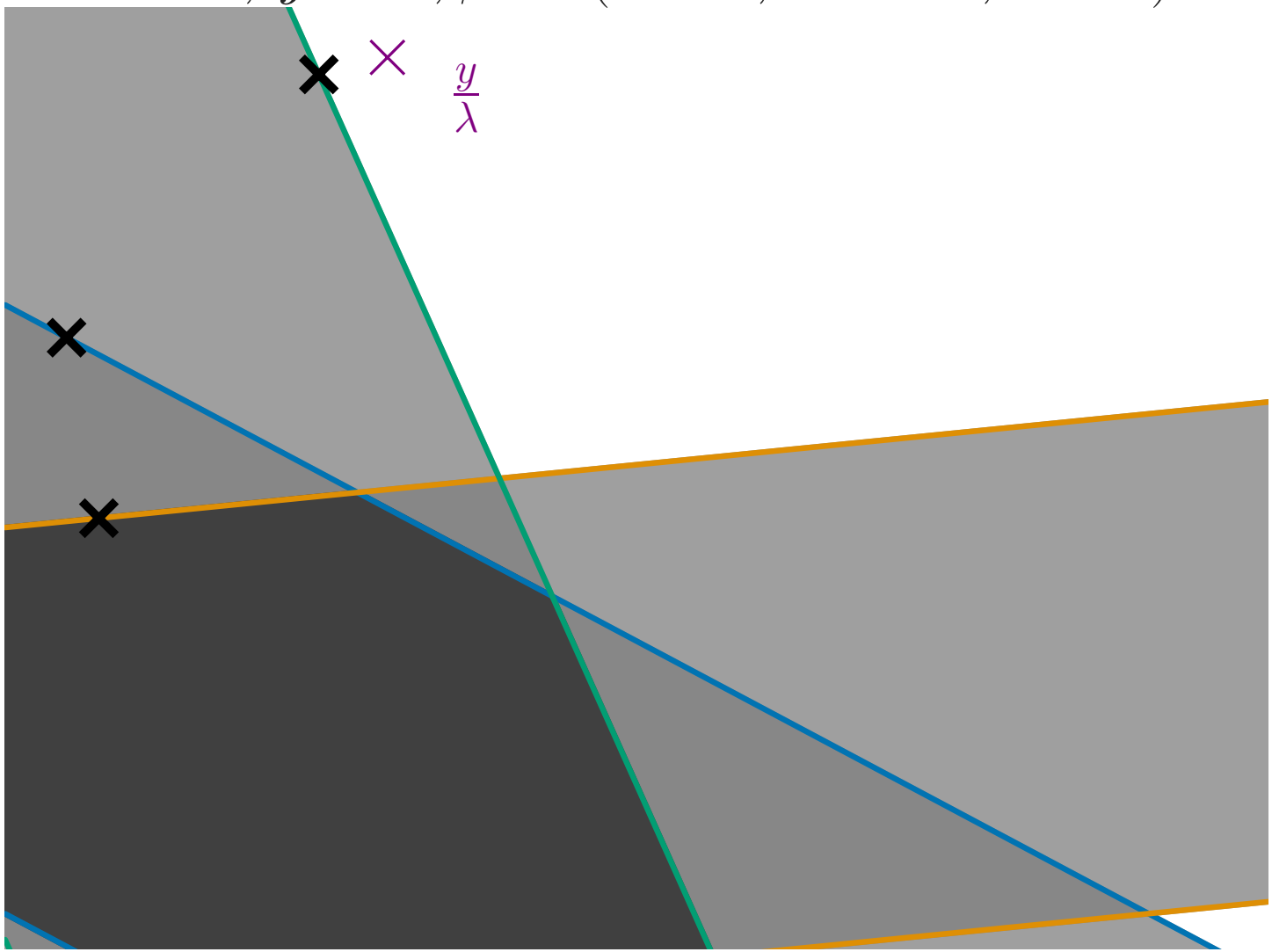
Regularity after sign identification

$$t = 1,\ j = 2, \beta = (0.217, -1.306, 0.000)$$



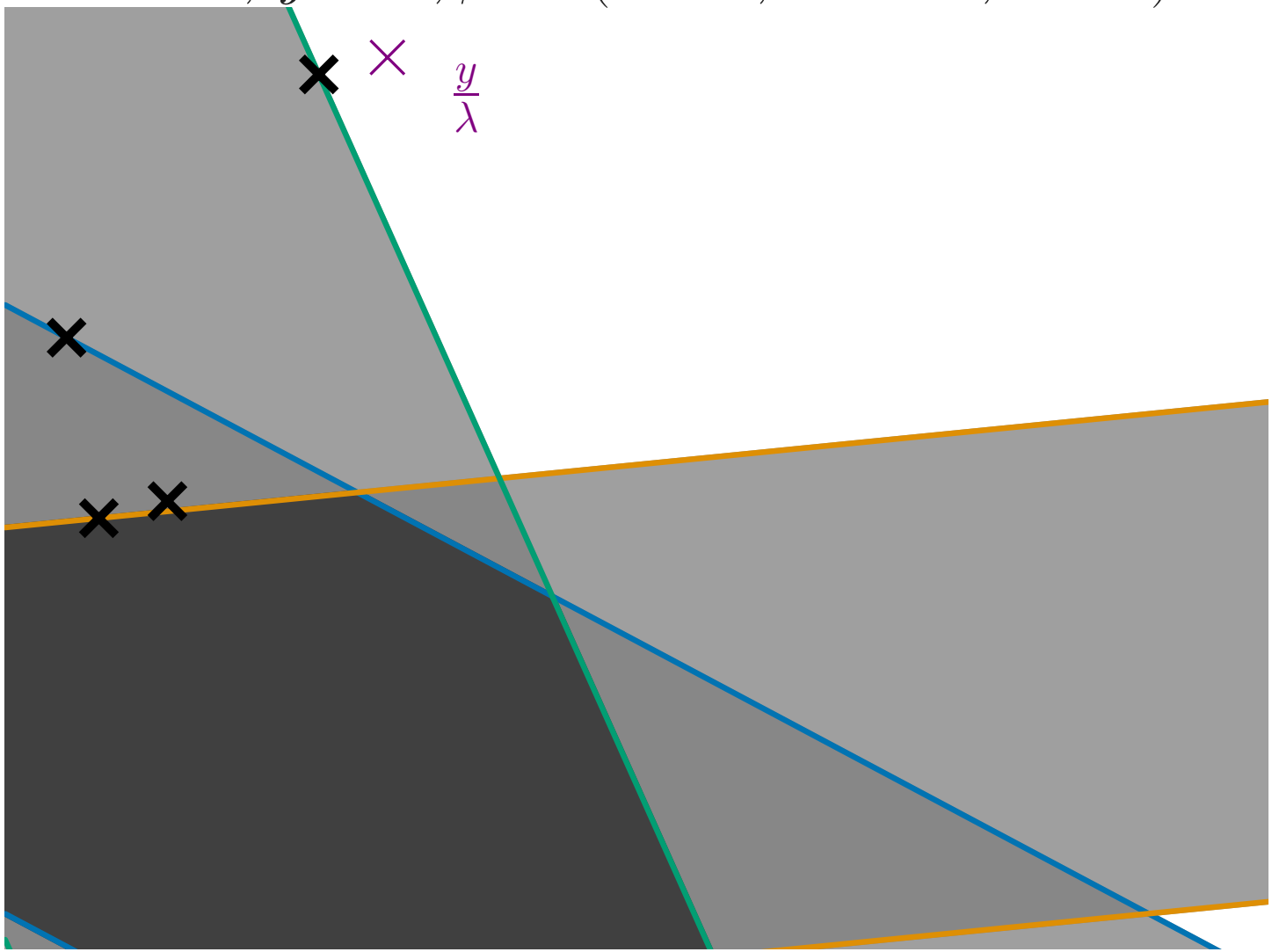
Regularity after sign identification

$$t = 1,\ j = 3, \beta = (0.217, -1.306, 0.735)$$



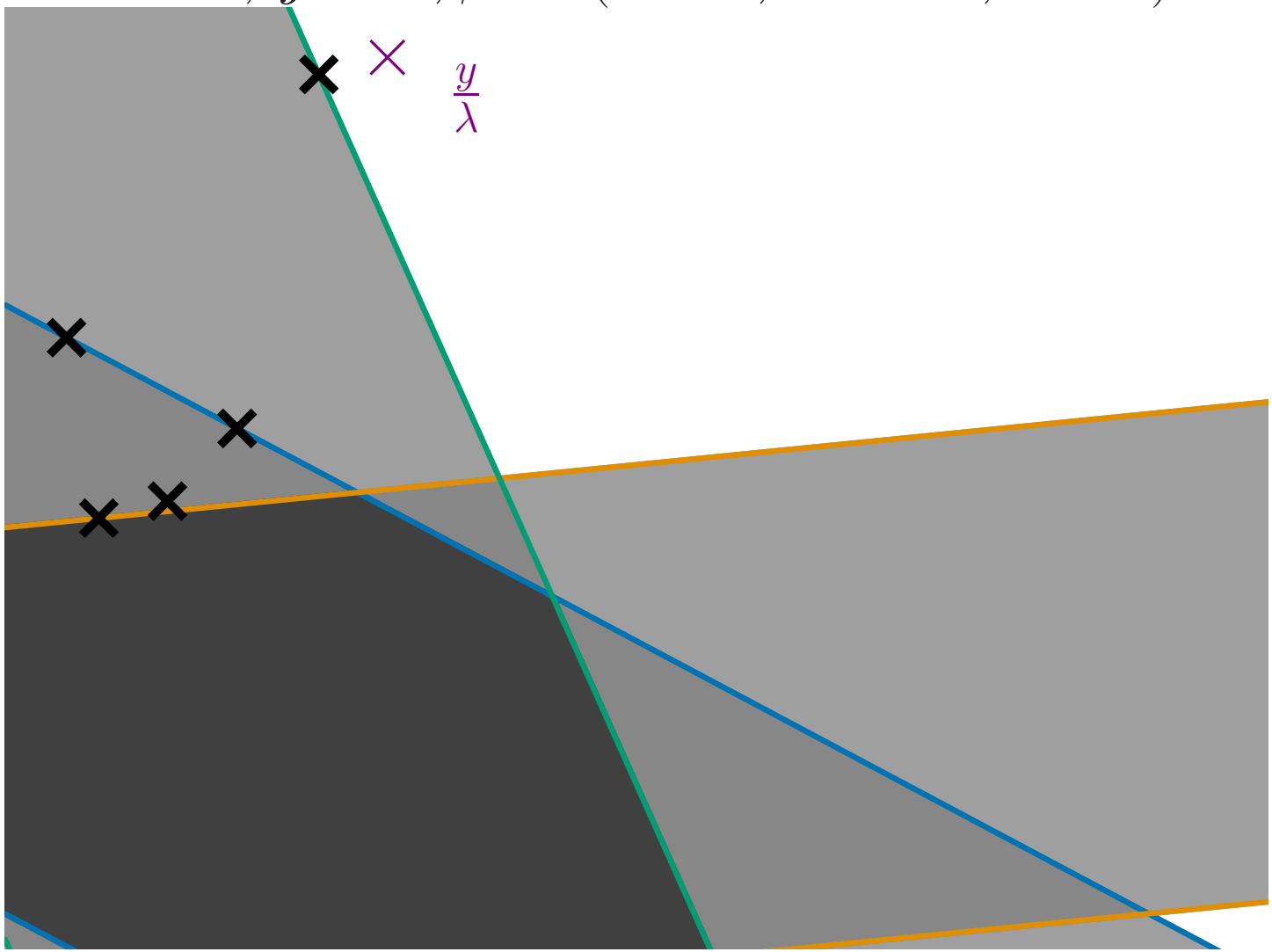
Regularity after sign identification

$$t = 2,\ j = 1, \beta = (0.000, -1.306, 0.735)$$



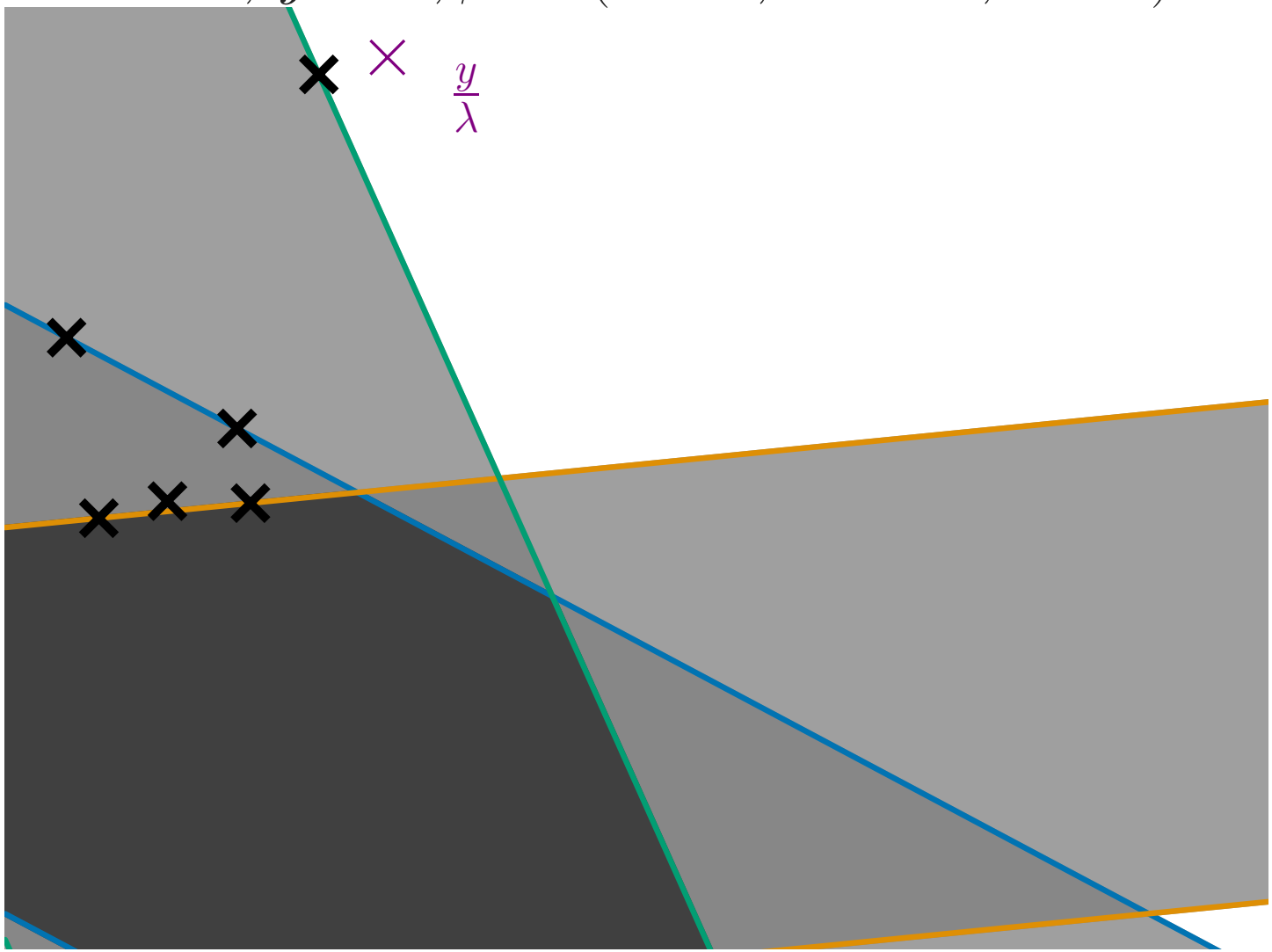
Regularity after sign identification

$$t = 2,\ j = 2, \beta = (0.000, -0.945, 0.735)$$



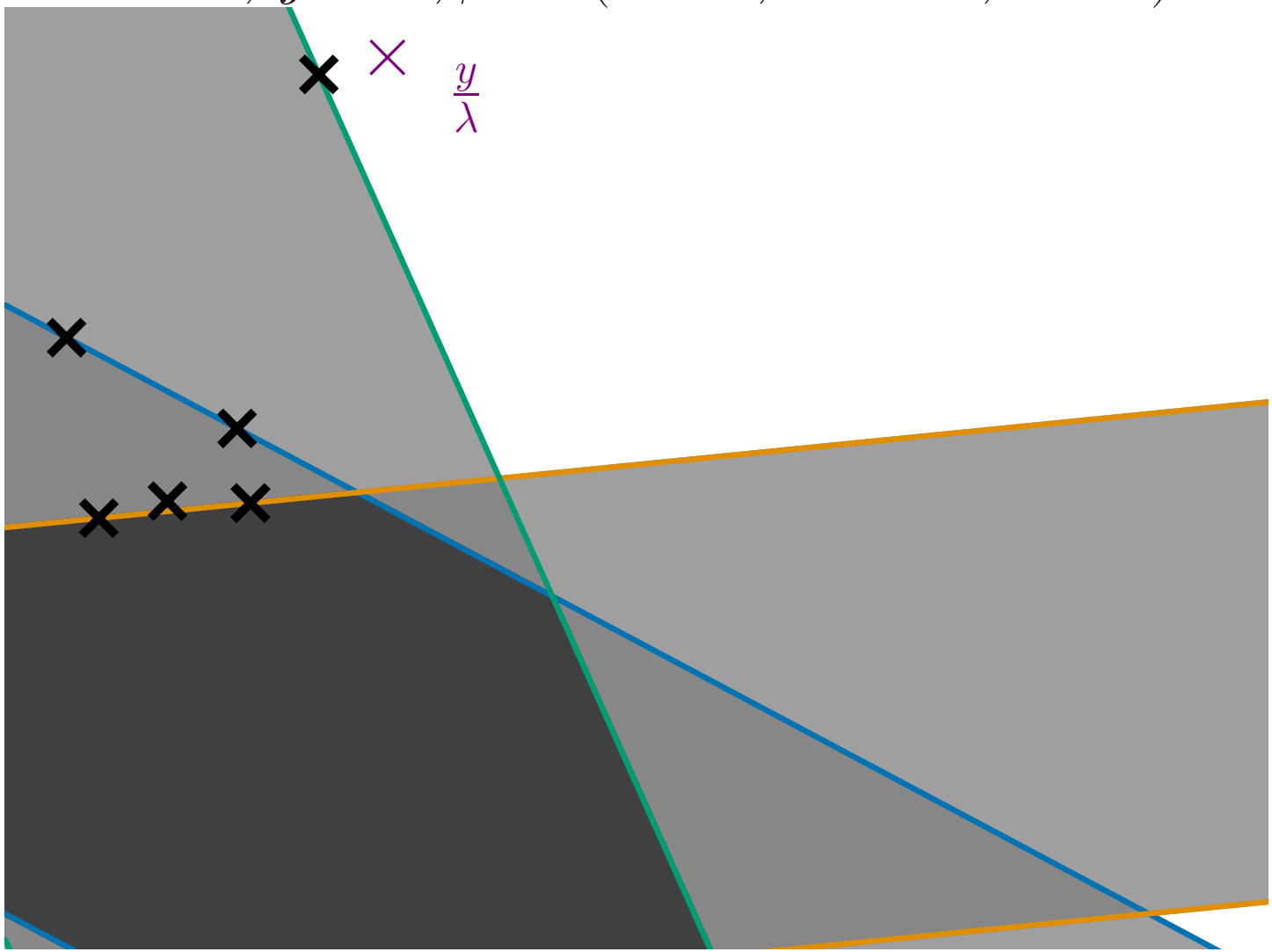
Regularity after sign identification

$$t = 2,\ j = 3, \beta = (0.000, -0.945, 1.039)$$



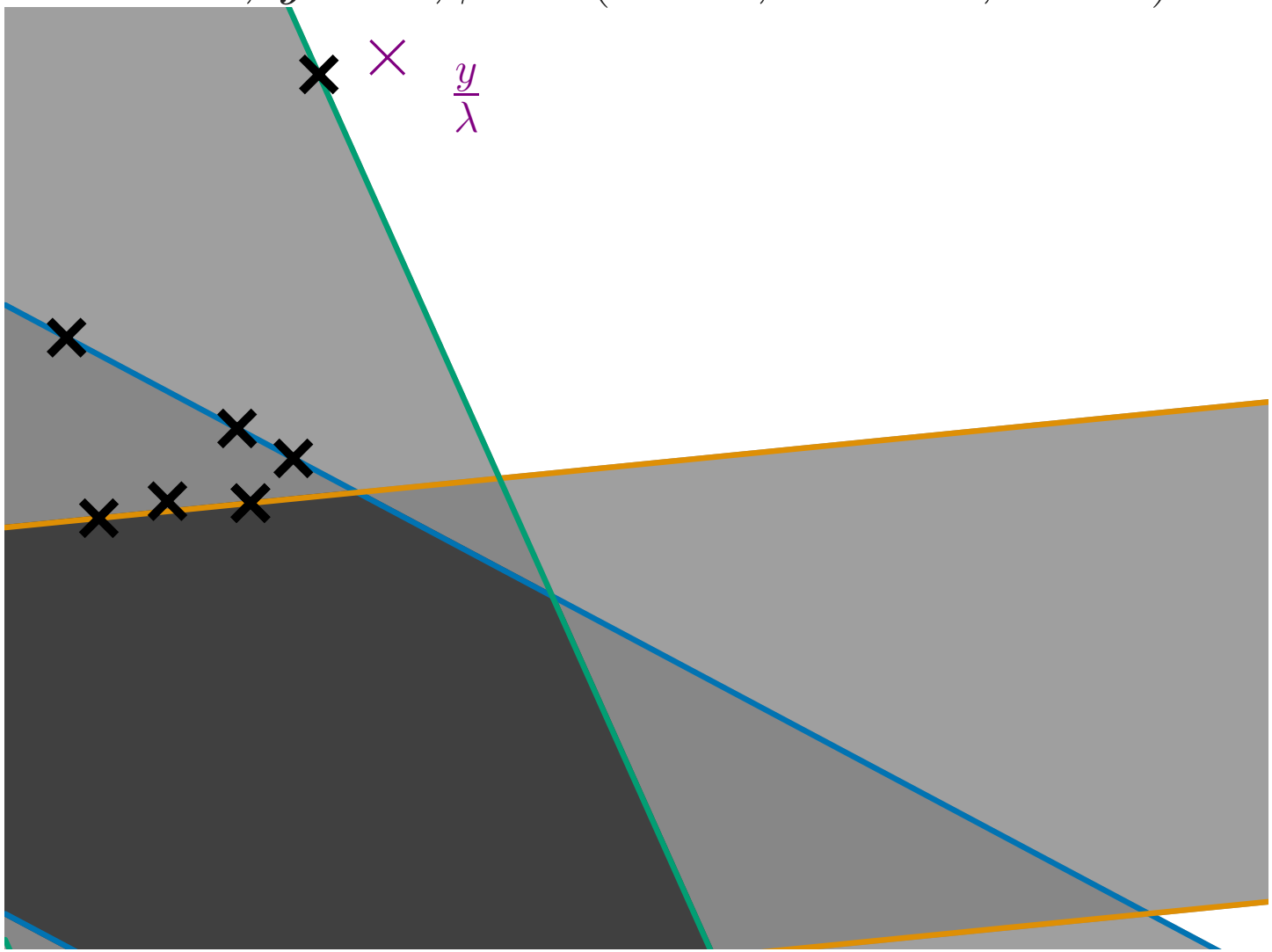
Regularity after sign identification

$$t = 3,\ j = 1, \beta = (0.000, -0.945, 1.039)$$



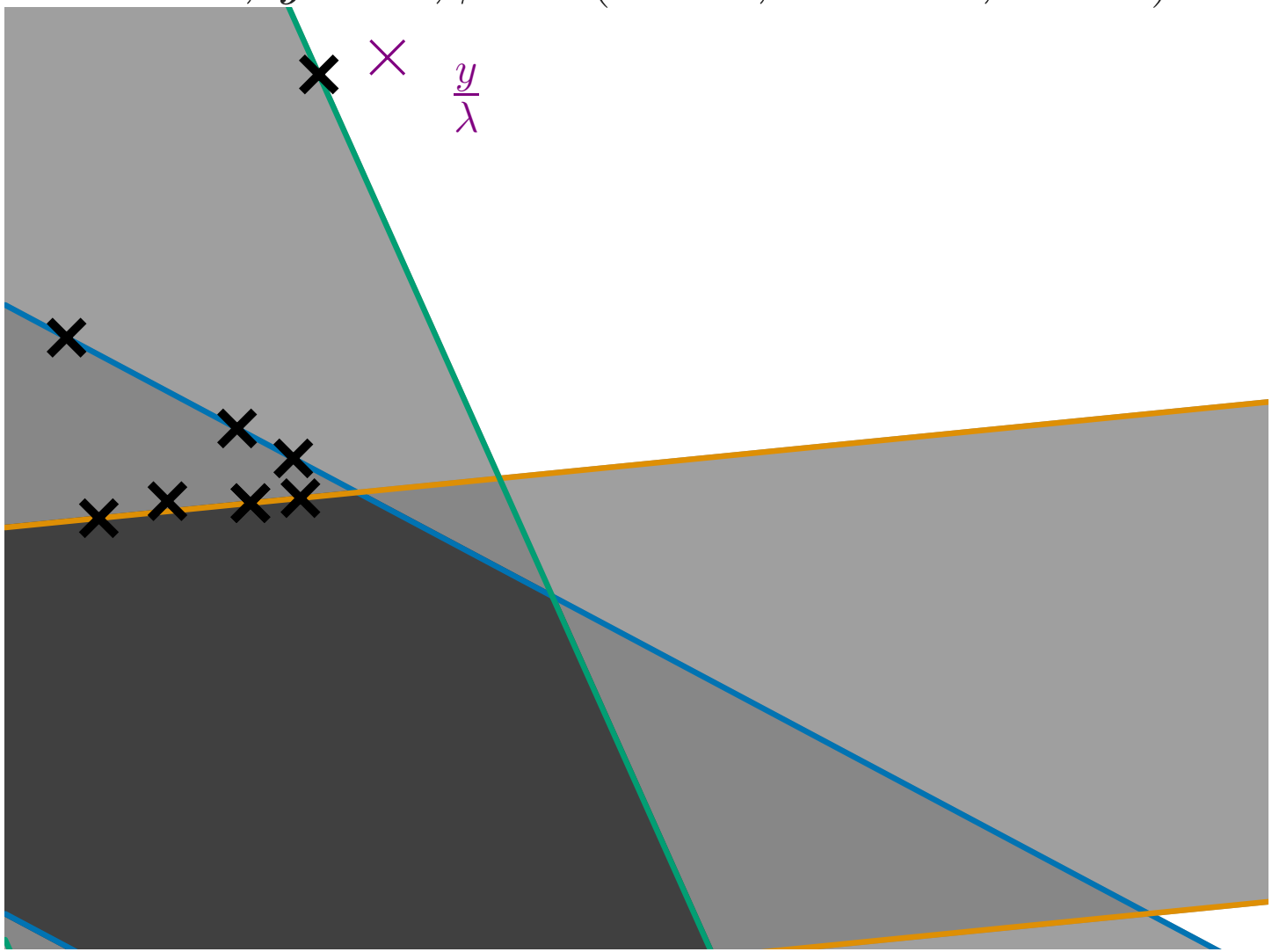
Regularity after sign identification

$$t = 3,\ j = 2, \beta = (0.000, -0.723, 1.039)$$



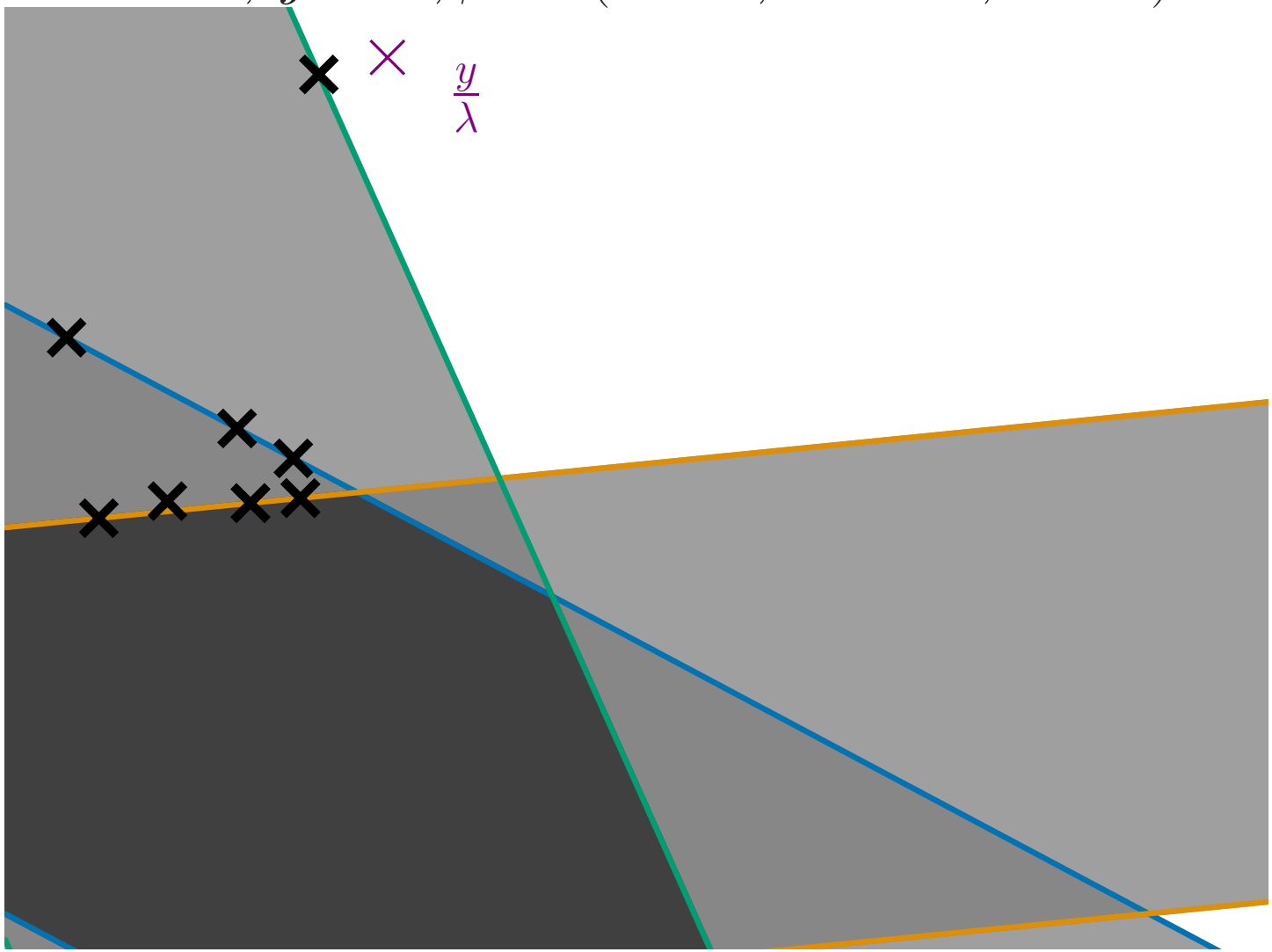
Regularity after sign identification

$$t = 3,\ j = 3, \beta = (0.000, -0.723, 1.201)$$



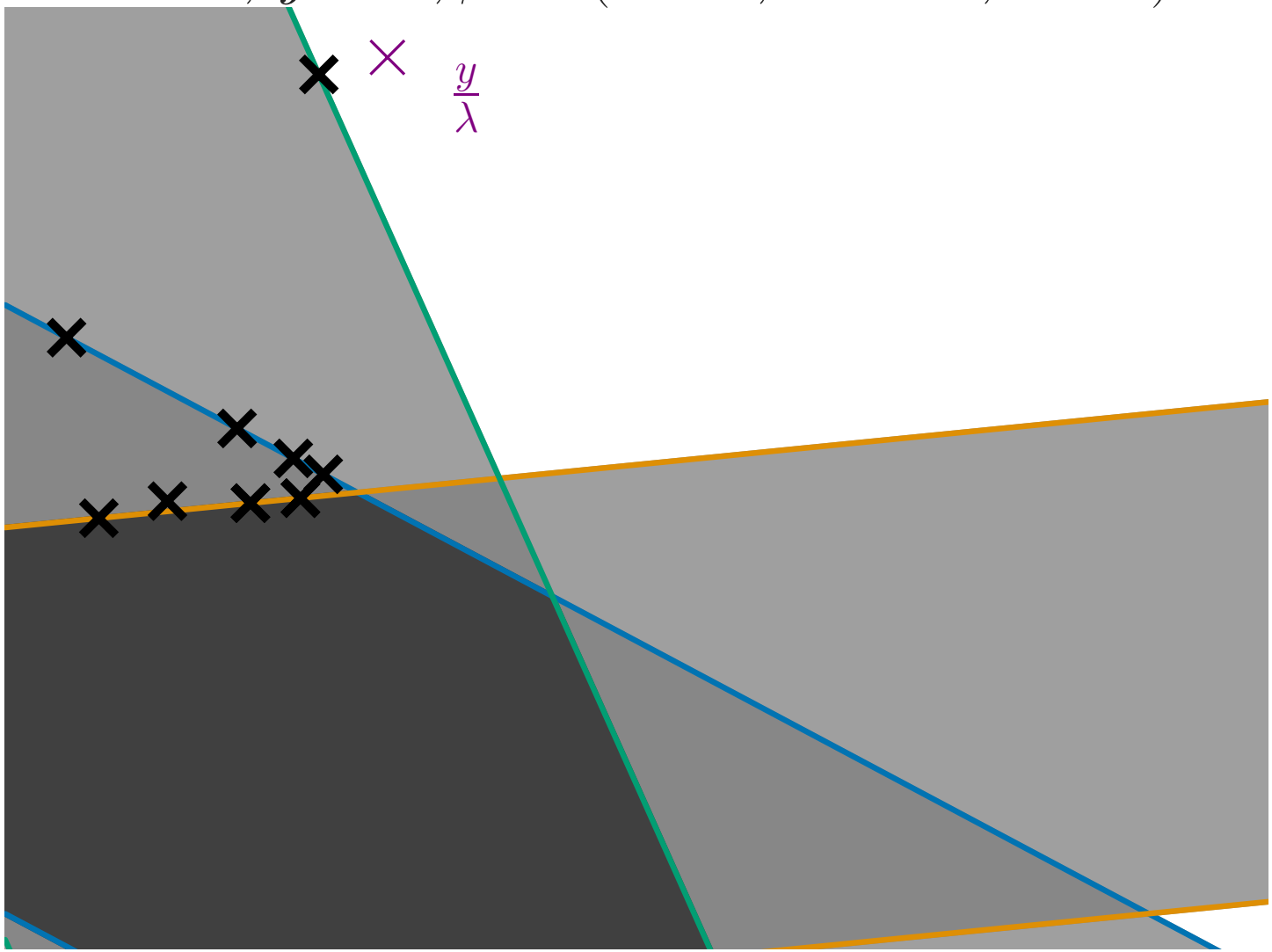
Regularity after sign identification

$$t = 4,\ j = 1, \beta = (0.000, -0.723, 1.201)$$



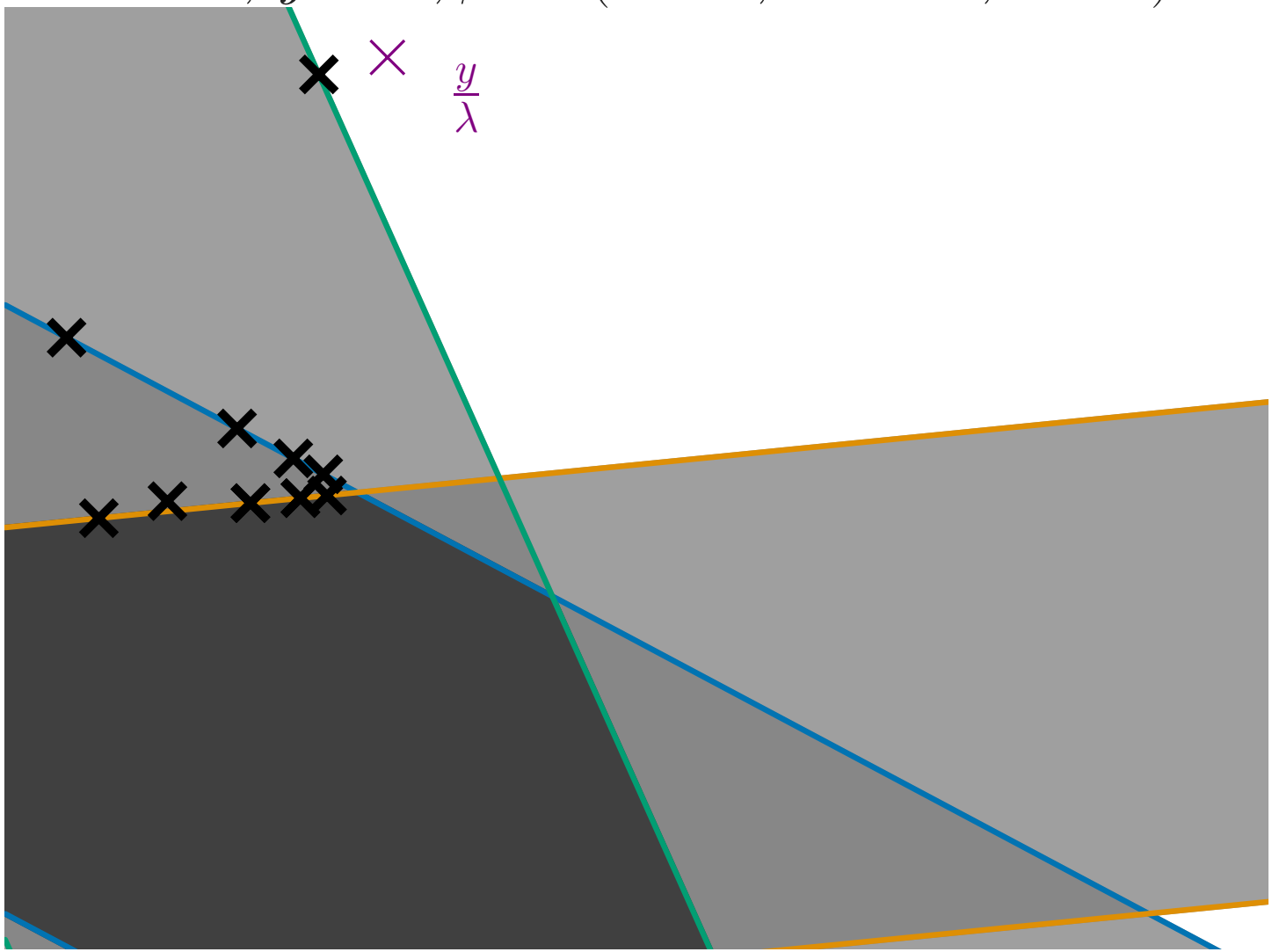
Regularity after sign identification

$$t = 4,\ j = 2, \beta = (0.000, -0.605, 1.201)$$



Regularity after sign identification

$$t = 4,\ j = 3, \beta = (0.000, -0.605, 1.287)$$



Regularity in residuals

Thm: CD achieves sign identification ($\operatorname{sign} \beta_j^{(t)} = \operatorname{sign} \hat{\beta}_j$)

Thm: Residuals from CD are Vector AutoRegressive[4] (**VAR**):

$$\boxed{r^{(t+1)} = Ar^{(t)} + b}$$

$\hookrightarrow$ we just need to fit a VAR to infer $\lim_{t\to\infty} r^{(t)} = \lambda\hat{\theta}$

[4] after sign ID

Regularity in residuals

Thm: CD achieves sign identification ($\operatorname{sign} \beta_j^{(t)} = \operatorname{sign} \hat{\beta}_j$)

Thm: Residuals from CD are Vector AutoRegressive[4] (**VAR**):

$$\boxed{r^{(t+1)} = Ar^{(t)} + b}$$

$\hookrightarrow$ we just need to fit a VAR to infer $\lim_{t\to\infty} r^{(t)} = \lambda\hat{\theta}$

It is costly (OLS) + we don't know when the sign is identified

Instead: **extrapolation**

[4] after sign ID

Acceleration through extrapolation[5]

What is the limit of $(0, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \ldots)$?

[5] D. Scieur, A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NIPS*. 2016, pp. 712–720.

Extrapolated dual point[6]

- Keep track of $K+1$ past residuals $r^{(t)}, \dots, r^{(t-K)}$
- Solve constrained least squares:

$$c^* = \underset{c^\top \mathbf{1}_K = 1}{\arg\min} \left\| \sum_{k=1}^{K} c_k (r^{(t-K+k)} - r^{(t-K-1+k)}) \right\|$$

[6] M. Massias, A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.

Extrapolated dual point[6]

- Keep track of $K+1$ past residuals $r^{(t)}, \dots, r^{(t-K)}$
- Solve constrained least squares:

$$c^* = \underset{c^\top \mathbf{1}_K = 1}{\arg\min} \left\| \sum_{k=1}^{K} c_k \left(r^{(t-K+k)} - r^{(t-K-1+k)} \right) \right\|$$

- Extrapolate:

$$r^t_{\text{accel}} = \sum_{k=1}^{K} c^*_k r^{(t+1-k)}, \quad \text{for } t > K$$

[6] M. Massias, A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.

Extrapolated dual point[6]

- Keep track of $K+1$ past residuals $r^{(t)}, \dots, r^{(t-K)}$
- Solve constrained least squares:

$$c^* = \underset{c^\top \mathbf{1}_K = 1}{\arg\min} \left\| \sum_{k=1}^{K} c_k \left(r^{(t-K+k)} - r^{(t-K-1+k)} \right) \right\|$$

- Extrapolate:

$$r^t_{\text{accel}} = \sum_{k=1}^{K} c^*_k r^{(t+1-k)}, \quad \text{for } t > K$$

- Get dual feasible point:

$$\theta^{(t)}_{\text{accel}} := r^{(t)}_{\text{accel}} / \max(\lambda, \|X^\top r^{(t)}_{\text{accel}}\|_\infty)$$

[6] M. Massias, A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.

Extrapolated dual point[6]

- Keep track of $K+1$ past residuals $r^{(t)}, \dots, r^{(t-K)}$
- Solve constrained least squares:

$$c^* = \underset{c^\top \mathbf{1}_K = 1}{\arg\min} \left\| \sum_{k=1}^{K} c_k \left(r^{(t-K+k)} - r^{(t-K-1+k)} \right) \right\|$$

- Extrapolate:

$$r^t_{\text{accel}} = \sum_{k=1}^{K} c^*_k r^{(t+1-k)}, \quad \text{for } t > K$$

- Get dual feasible point:

$$\boxed{\theta^{(t)}_{\text{accel}} := r^{(t)}_{\text{accel}} / \max(\lambda, \|X^\top r^{(t)}_{\text{accel}}\|_\infty)}$$

[6] M. Massias, A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.

Extrapolated dual point[6]

- Keep track of $K+1$ past residuals $r^{(t)}, \dots, r^{(t-K)}$
- Solve constrained least squares:

$$c^* = \underset{c^\top \mathbf{1}_K = 1}{\arg\min} \left\| \sum_{k=1}^{K} c_k \left(r^{(t-K+k)} - r^{(t-K-1+k)} \right) \right\|$$

- Extrapolate:

$$r^t_{\text{accel}} = \sum_{k=1}^{K} c^*_k r^{(t+1-k)}, \quad \text{for } t > K$$

- Get dual feasible point:

$$\boxed{\theta^{(t)}_{\text{accel}} := r^{(t)}_{\text{accel}} / \max(\lambda, \|X^\top r^{(t)}_{\text{accel}}\|_\infty)}$$

$K = 5$ is enough in practice!

---

[6] M. Massias, A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.

Why we have a VAR sequence

After **sign identification**: $\operatorname{sign} \beta_j^{(t)} = \operatorname{sign} \hat{\beta}_j$

Support of $\hat{\beta}$: $\{j_1, \ldots, j_S\}$ (other coordinates stay at 0)

Consider 1 epoch of CD:

$$\beta^{(t)} \to \beta^{(t+1)}$$

Decompose into non-zero coordinate updates

$$\beta^{(t)} = \tilde{\beta}^{(0)} \xrightarrow{j_1} \tilde{\beta}^{(1)} \xrightarrow{j_2} \ldots \xrightarrow{j_S} \tilde{\beta}^{(S)} = \beta^{(t+1)}$$

$\tilde{\beta}^{(s)} = \tilde{\beta}^{(s-1)}$ except at coordinate j_s:

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

Why we have a VAR sequence

$$X\tilde{\beta}^{(s)} = \underbrace{\left(\mathrm{Id}_n - \frac{1}{\|x_{j_s}\|^2} x_{j_s} x_{j_s}^\top\right)}_{A_s \in \mathbb{R}^{n\times n}} X\tilde{\beta}^{(s-1)} + \underbrace{\frac{x_{j_s}^\top y - \lambda \operatorname{sign}(\hat{\beta}_{j_s})}{\|x_{j_s}\|^2} x_{j_s}}_{b_s \in \mathbb{R}^n}$$

So for the full epoch $t \to t+1$:

$$\begin{aligned} X\tilde{\beta}^{(S)} &= A_S X\tilde{\beta}^{(S-1)} + b_S \\ &= A_S A_{S-1} X\tilde{\beta}^{(S-2)} + A_S b_{S-1} + b_S \\ &= \underbrace{A_S \dots A_1}_{A} X\tilde{\beta}^{(0)} + \underbrace{A_S \dots A_2 b_1 + \dots + A_S b_{S-1} + b_S}_{b} \end{aligned}$$

$$\boxed{X\beta^{(t+1)} = AX\beta^{(t)} + b}$$

Guarantees

- extrapolation works when sign is identified
- before that, $r^{(t)}$ follow VARs with different A's $\hookrightarrow$ stable behavior

Guarantees

- extrapolation works when sign is identified
- before that, $r^{(t)}$ follow VARs with different A's $\hookrightarrow$ stable behavior

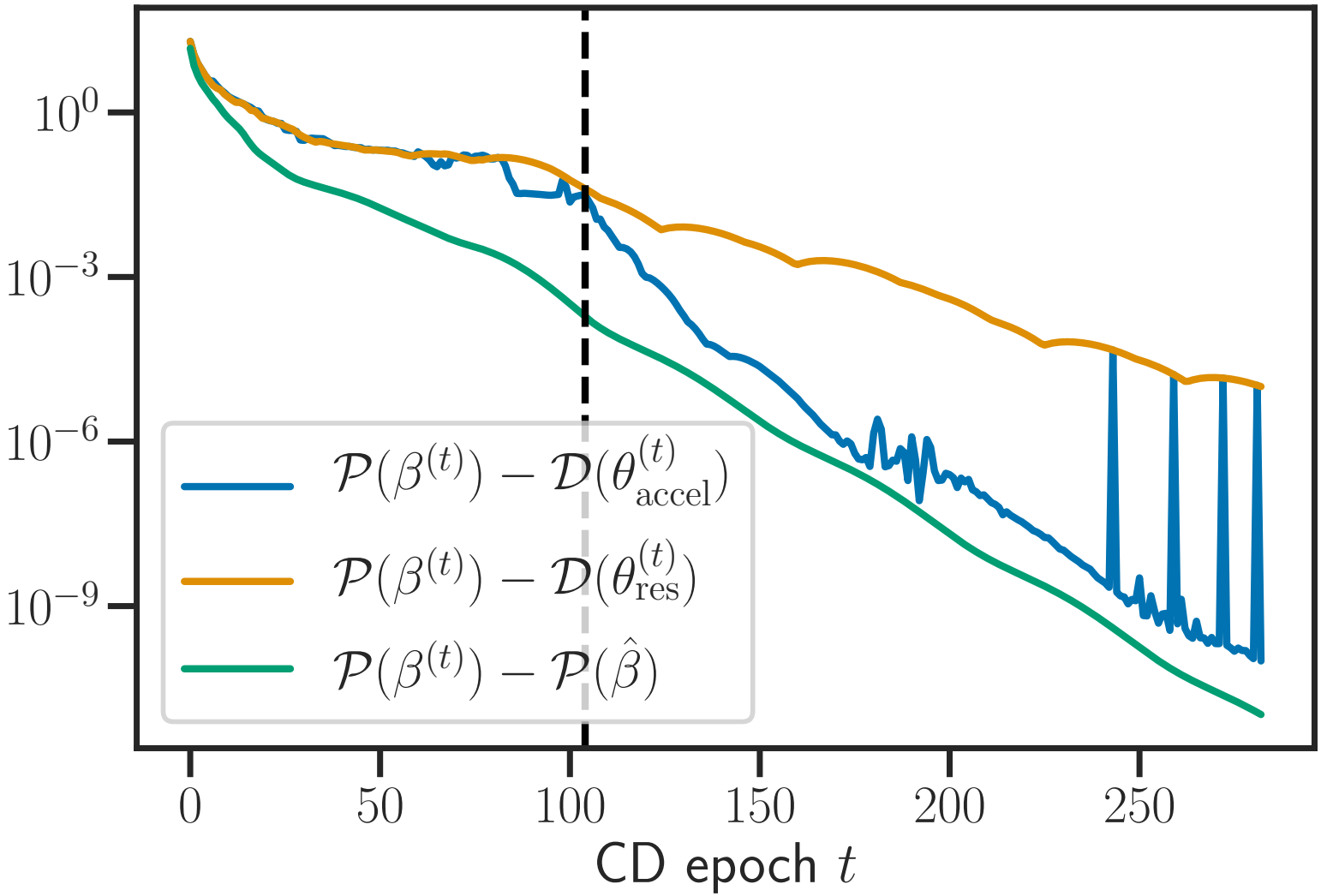
θ_{accel} is $\mathcal{O}(np + K^2 n + K^3)$ to compute, so compute θ_{res} as well

$$\text{use } \theta^{(t)} = \underset{\theta \in \{\theta^{(t)}_{\text{res}}, \theta^{(t)}_{\text{accel}}, \theta^{(t-1)}\}}{\arg\max} \mathcal{D}(\theta)$$

Cost (including stopping criterion evaluation):

- classical: evaluate 1 dual point every 10 CD epochs $\approx 11np$
- new: evaluate 2 dual points every 10 CD epochs $\approx 12np$

Lasso: in practice



Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

Table of Contents

Motivation: sparse inverse problems

Pedagogical example: the Lasso

Exploiting regularity

Sparse GLMs

More solvers speed-up

Structure for other GLMs

Sparse GLM:

$$\hat{\beta} \in \arg\min F(X\beta) + \lambda\|\beta\|_1$$
$$\hat{\theta} = -\nabla F(X\hat{\beta})/\lambda$$

1 CD update after sign ID:

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

∇F not linear in $X\beta$ if F is not a quadratic data-fitting term!

Lasso: $-\nabla F = y - X\beta$, hence: **C**onstraint **E**limination for the **L**asso with **E**xtrapolated **R**esiduals

Structure for GLMs

Solution: linearization of ∇F around optimum

$$\nabla F(X\beta) = \nabla F(X\hat{\beta}) + \underbrace{D}_{\in \mathbb{R}^{n\times n}} (X\beta - X\hat{\beta}) + o(X\beta - X\hat{\beta})$$

Leads to **asymptotic VAR sequence**:

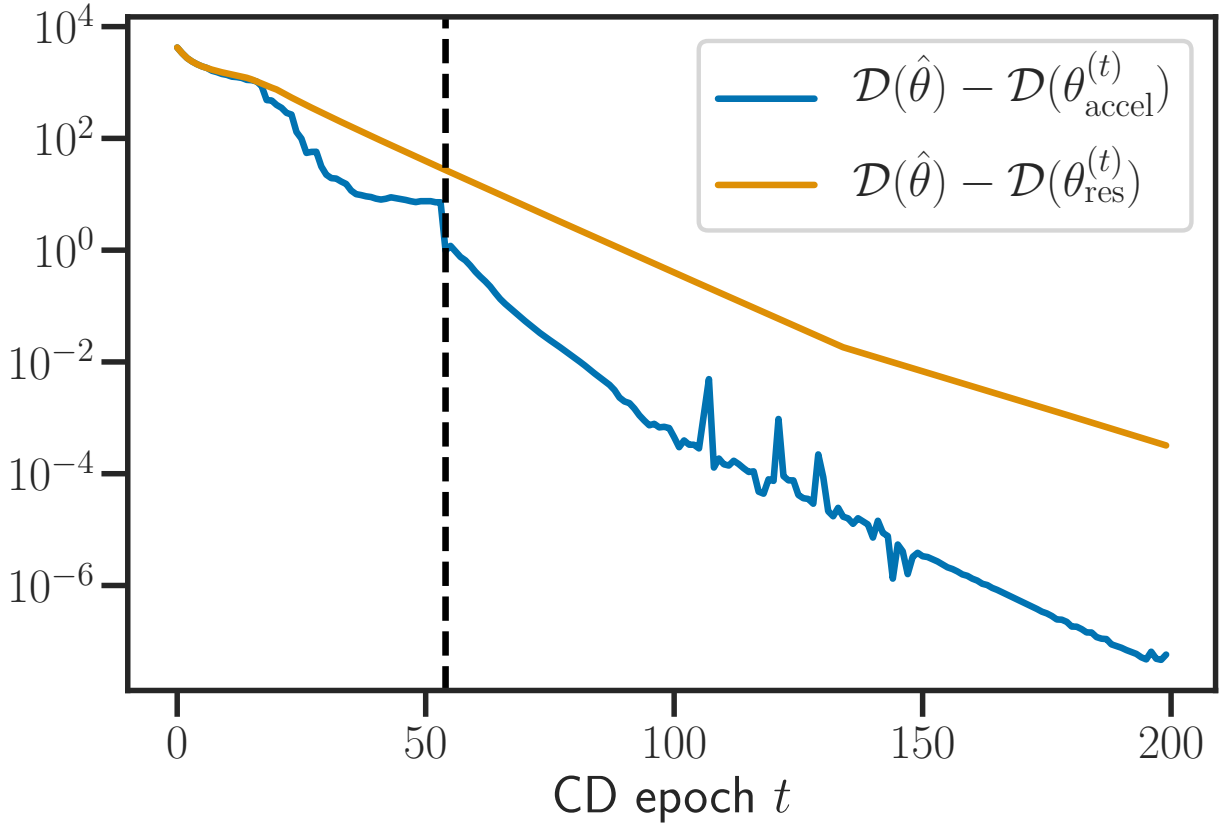
$$X\tilde{\beta}^{(s)} = \left(\mathrm{Id}_n - \tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s} x_{j_s}^\top D\right) X\tilde{\beta}^{(s-1)} + cst + o(\dots)$$

$$D^{1/2}X\tilde{\beta}^{(s)} = \underbrace{\left(\mathrm{Id}_n - \tfrac{\gamma}{\|x_{j_s}\|^2} D^{1/2} x_{j_s} x_{j_s}^\top D^{1/2}\right)}_{A_s} D^{1/2}X\tilde{\beta}^{(s-1)} + b_s + o(\dots)$$

$$D^{1/2}X\beta^{(t+1)} = A_S \dots A_1 D^{1/2} X\beta^{(t)} + b_S + \dots + A_S \dots A_2 b_1 + o(\dots)$$

$$\boxed{X\beta^{(t+1)} = AX\beta^{(t)} + b + o(X\beta - X\hat{\beta})}$$

Sparse logreg: it works



rcv1 dataset: $p = 20\text{k}, \quad n = 20\text{k}, \quad \lambda = \lambda_{\max}/20 \;\; (\|\hat{\beta}\|_0 = 395)$

Structure for group penalties

Multitask Lasso:

$$\hat{\mathrm{B}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min}\, F(X\mathrm{B}) + \lambda \sum_{j=1}^{q} \|\mathrm{B}_{j:}\|_1$$

1 CD update after sign ID:

$$\tilde{\mathrm{B}}^{(s)}_{j_s:} = \mathrm{BST}\Big(\tilde{\mathrm{B}}^{(s-1)}_{j_s:} - \tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}^\top \nabla F(X\tilde{\mathrm{B}}^{(s-1)}), \tfrac{\gamma}{\|x_{j_s}\|^2}\lambda\Big)$$

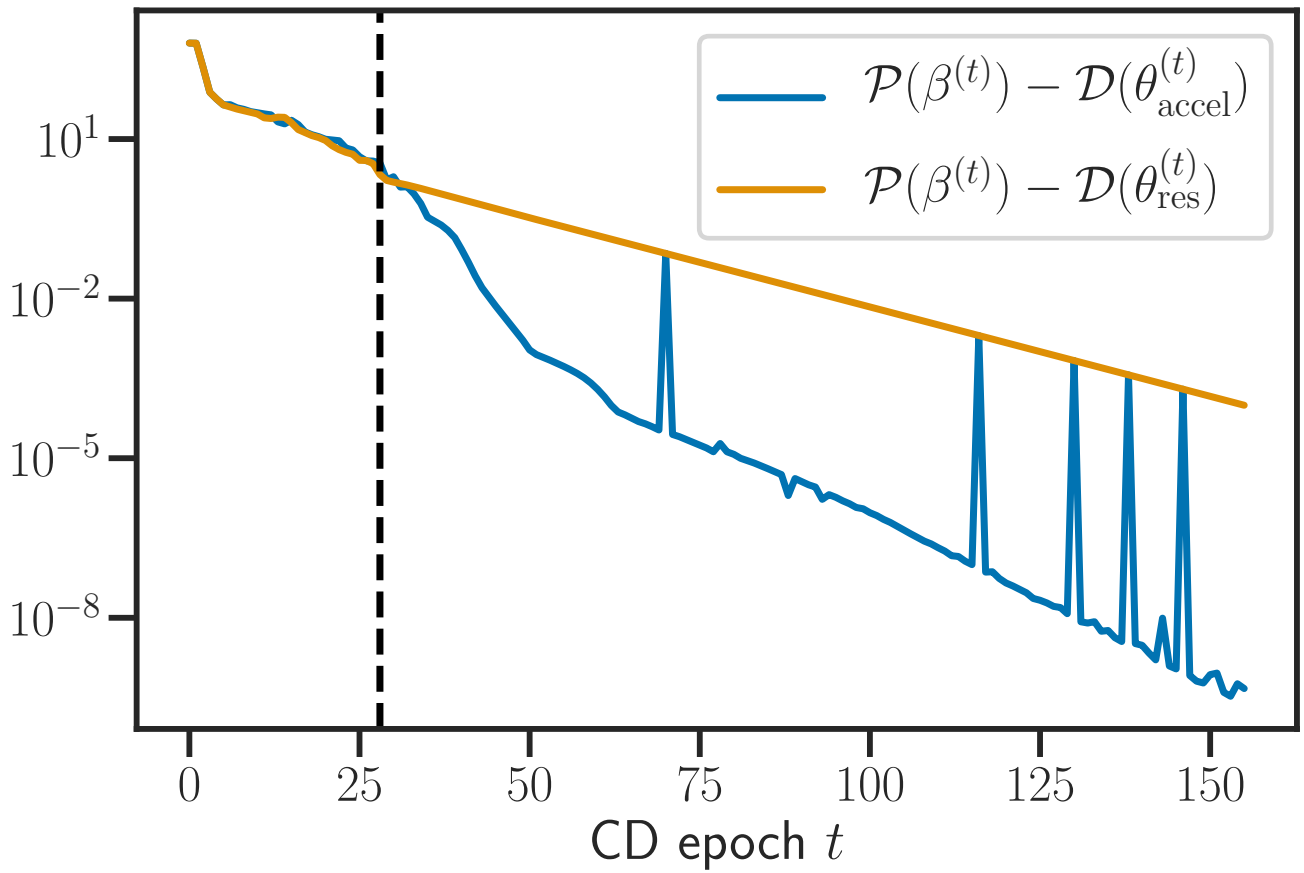
$$\neq \tilde{\mathrm{B}}^{(s-1)}_{j_s:} - \tfrac{\gamma}{\|x_{j_s}\|^2} x_{j_s}^\top \nabla F(X\tilde{\mathrm{B}}^{(s-1)}) - \tfrac{\gamma}{\|x_{j_s}\|^2}\lambda \operatorname{sign}(\hat{\mathrm{B}}_{j_s:})$$

$$\mathrm{BST}(x, \tau) = \left(1 - \frac{\tau}{\|x\|}\right)_+ x$$

BST no longer an additive bias when $q > 1$...

... but same linearization ideas work in practice

Multitask Lasso: it works



MEG data: $p = 7498, \quad n = 305, \quad \lambda = \lambda_{\max}/10 \;\; (\|\hat{\mathrm{B}}\|_{2,0} = 45)$

Intermediate summary

- Lasso: exact VAR sequence
- Other data-fitting term, group penalty: still works

Current questions:

- combination with accelarating β
- combination with accelerated CD/line search

Table of Contents

Motivation: sparse inverse problems

Pedagogical example: the Lasso

Exploiting regularity

Sparse GLMs

More solvers speed-up

Speeding-up solvers

Two approaches:

- safe screening[7,8] (**backward approach**): remove feature j when it is certified that $\hat{\beta}_j = 0$
- working set[9] (**forward approach**): focus on j's for which it is very likely that $\hat{\beta}_j \neq 0$.

---

[7] L. El Ghaoui, V. Viallon, and T. Rabbani. "Safe feature elimination in sparse supervised learning". In: *J. Pacific Optim.* 8.4 (2012), pp. 667–698.

[8] A. Bonnefoy et al. "A dynamic screening principle for the lasso". In: *EUSIPCO*. 2014.

[9] T. B. Johnson and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.

Identifying features

Equicorrelation set[10]

$$E := \left\{ j \in [p] \,:\, |X_j^\top \hat{\theta}| = 1 \right\} = \left\{ j \in [p] \,:\, \frac{|X_j^\top (y - X\hat{\beta})|}{\lambda} = 1 \right\}$$

- For any primal solution, $j \notin E \implies \hat{\beta}_j = 0$

[10] R. J. Tibshirani. "The lasso problem and uniqueness". In: *Electron. J. Stat.* 7 (2013), pp. 1456–1490.

Identifying features

Equicorrelation set[10]

$$E := \left\{ j \in [p] \,:\, |X_j^\top \hat{\theta}| = 1 \right\} = \left\{ j \in [p] \,:\, \frac{|X_j^\top (y - X\hat{\beta})|}{\lambda} = 1 \right\}$$

- For any primal solution, $j \notin E \implies \hat{\beta}_j = 0$

Idea for speed-up: identify E, solve only on E

Practical observation: generally $\#E \ll p$

[10] R. J. Tibshirani. "The lasso problem and uniqueness". In: *Electron. J. Stat.* 7 (2013), pp. 1456–1490.

Duality again: gap screening

Cannot identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

Good proxy: find a region $\mathcal{C} \subset \mathbb{R}^n$ containing $\hat{\theta}$

$$\sup_{\theta \in \mathcal{C}} |X_j^\top \theta| < 1 \Rightarrow |X_j^\top \hat{\theta}| < 1 \Rightarrow j \notin E \Rightarrow \hat{\beta}_j = 0$$

[11] E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.

Duality again: gap screening

Cannot identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

Good proxy: find a region $\mathcal{C} \subset \mathbb{R}^n$ containing $\hat{\theta}$

$$\sup_{\theta \in \mathcal{C}} |X_j^\top \theta| < 1 \Rightarrow |X_j^\top \hat{\theta}| < 1 \Rightarrow j \notin E \Rightarrow \hat{\beta}_j = 0$$

Gap Safe screening rule[11]: $\mathcal{C}$ is a ball of radius $\rho = \sqrt{\frac{2}{\lambda^2} \mathsf{dgap}(\beta, \theta)}$ centered at $\theta \in \Delta_X$

$$\boxed{\forall (\beta, \theta) \in \mathbb{R}^p \times \Delta_X, \quad |X_j^\top \theta| < 1 - \|X_j\| \rho \Rightarrow \hat{\beta}_j = 0}$$

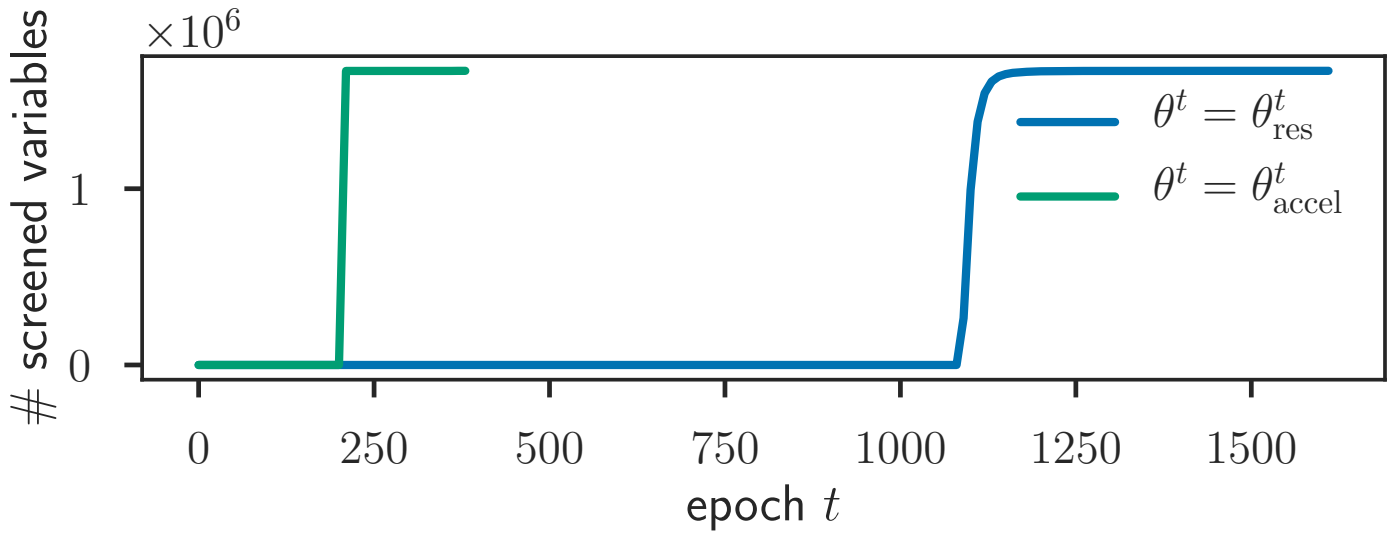
[11] E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.

Better Gap Safe screening[12]

Gap Safe screening rule:

$$\forall \theta \in \Delta_X, |X_j^\top \theta| < 1 - \|X_j\| \sqrt{\tfrac{2}{\lambda^2} \mathsf{dgap}(\beta, \theta)} \Rightarrow \hat{\beta}_j = 0$$

better dual point $\Rightarrow$ better safe screening



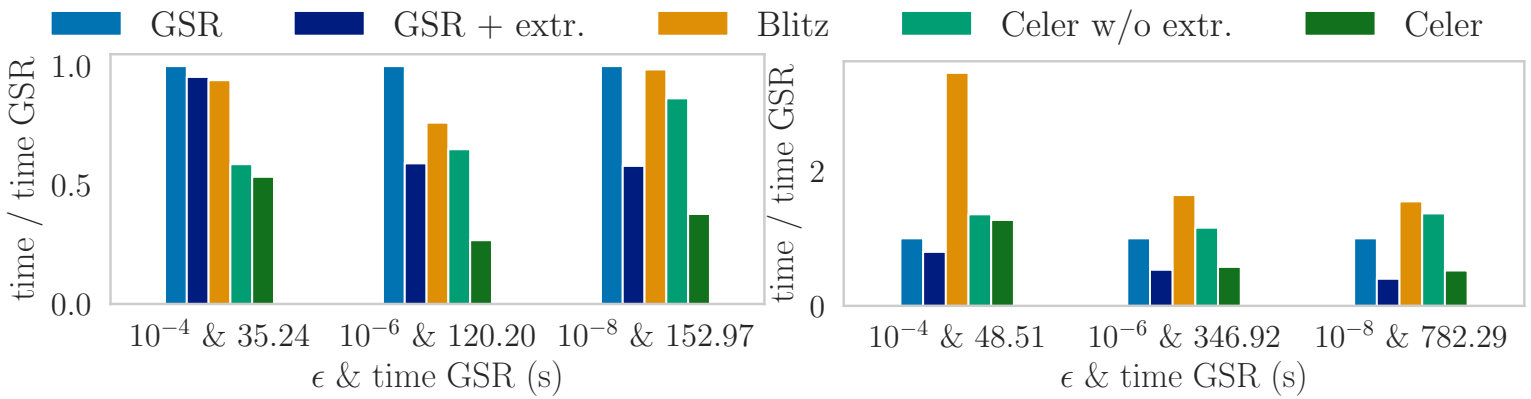
Finance dataset: $(p = 1.5 \times 10^6, n = 1.5 \times 10^4)$, $\lambda = \lambda_{\max}/5$

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

Working sets

State-of-the-art WS solver for sparse problems: Blitz[13]

Screening can be used aggressively to define WS, and a **better dual point also helps** in this case



news20 dataset, coarse and fine Lasso paths computation

[13] T. B. Johnson and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.

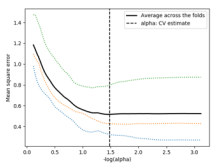
Online code

Fast & pip-installable Cython code, continuous integration, bug tracker, code coverage

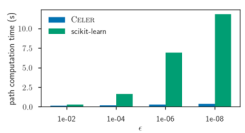
Figures & doc at https://mathurinm.github.io/celer

Examples Gallery¶

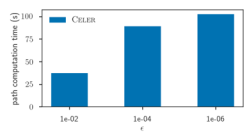
Fork me on GitHub



Run LassoCV for cross-validation on Leukemia



Lasso path computation on Leukemia dataset



Lasso path computation on Finance/log1p

```
from sklearn.linear_model import Lasso, LassoCV
from celer import Lasso, LassoCV
```

From 10,000 s to 50 s for cross-validation on Finance

Conclusion

Duality matters at several levels for sparse GLMs:

- stopping criterion
- safe feature identification (screening or working set)

Lasso: Exploiting the VAR structure of $y - X\beta^{(t)} \hookrightarrow$ better dual

Generalization

- any twice differentiable separable (samples) data-fitting term
- group penalties (multitask Lasso)

with proof of asymptotic VAR structure & extrapolation is useful

Code: https://github.com/mathurinm/celer
ICML: http://proceedings.mlr.press/v80/massias18a.html

References I

- Bonnefoy, A. et al. "A dynamic screening principle for the lasso". In: *EUSIPCO*. 2014.
- Chen, S. S. and D. L. Donoho. "Atomic decomposition by basis pursuit". In: *SPIE*. 1995.
- El Ghaoui, L., V. Viallon, and T. Rabbani. "Safe feature elimination in sparse supervised learning". In: *J. Pacific Optim.* 8.4 (2012), pp. 667–698.
- Fercoq, O., A. Gramfort, and J. Salmon. "Mind the duality gap: safer rules for the lasso". In: *ICML*. 2015, pp. 333–342.
- Johnson, T. B. and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.
- Mairal, J. "Sparse coding for machine learning, image processing and computer vision". PhD thesis. École normale supérieure de Cachan, 2010.

References II

- Massias, M., A. Gramfort, and J. Salmon. "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*. 2018.
- Ndiaye, E. et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.
- Scieur, D., A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NIPS*. 2016, pp. 712–720.
- Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.
- Tibshirani, R. J. "The lasso problem and uniqueness". In: *Electron. J. Stat.* 7 (2013), pp. 1456–1490.

Intuition for extrapolation

Take a VAR sequence $x^{(t)} \to x^*$:

$$x^{(t+1)} - x^* = A(x^{(t)} - x^*)$$

Cayley-Hamilton: find coefficients[14] s.t. $\sum_{k=0}^{n} a_k A^k = 0$

$$\sum_{k=0}^{n} a_k (x^{(t+k+1)} - x^*) = \sum_{k=0}^{n} a_k A^k (x^{(t)} - x^*) = 0$$

$\hookrightarrow x^* \in \mathrm{Span}\,(x^{(t)}, \dots, x^{(t+n+1)})$
$\hookrightarrow$ approximate x^* under the form $x_{\mathrm{acc}} = \sum_{k=1}^{K} c_k x^{(t+k)}$

minimizing $\|x_{\mathrm{acc}} - (Ax_{\mathrm{acc}} + b)\|$ leads to the previous formulas

[14] wlog, assume $\sum_0^n a_k = 1$ if $\|A\| < 1$

Aitken's rule

For a converging sequence $(r_n)_{n\in\mathbb{N}}$, Aitken's rule replaces r_{n+1} by

$$\Delta^2 = r_n + \frac{1}{\frac{1}{r_{n+1}-r_n} - \frac{1}{r_n - r_{n-1}}}$$

Proof of the dual formulation

$$\min_{\beta\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|y-X\beta\|^2}_{f(y-X\beta)} + \lambda \underbrace{\|\beta\|_1}_{\Omega(\beta)} \Leftrightarrow \min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \begin{cases} f(z)+\lambda\Omega(\beta) \\ \text{s.t.} \quad z = y - X\beta \end{cases}$$

Lagrangian : $$\mathcal{L}(z,\beta,\theta) := \frac{1}{2}\|z\|^2 + \lambda\Omega(\beta) + \lambda\theta^\top(y - X\beta - z).$$

It is equivalent to finding a saddle point $(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)})$ of the Lagrangian (Strong duality):

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \max_{\theta\in\mathbb{R}^n} \mathcal{L}(z,\beta,\theta) = \max_{\theta\in\mathbb{R}^n} \min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \mathcal{L}(z,\beta,\theta) =$$

$$\max_{\theta\in\mathbb{R}^n} \left\{ \min_{z\in\mathbb{R}^n} [f(z) - \lambda\theta^\top z] + \min_{\beta\in\mathbb{R}^p} [\lambda\Omega(\beta) - \lambda\theta^\top X\beta] + \lambda\theta^\top y \right\} =$$

$$\max_{\theta\in\mathbb{R}^n} \left\{ -f^*(\lambda\theta) - \lambda\Omega^*(X^\top\theta) + \lambda\theta^\top y \right\}$$

which is the formulation asserted (with conjugacy properties)

Conjugation

For any $f : \mathbb{R}^n \to \mathbb{R}$, the (Fenchel) conjugate f^* is defined as

$$f^*(z) = \sup_{x \in \mathbb{R}^n} x^\top z - f(z)$$

- If $f(\cdot) = \|\cdot\|^2/2$ then $f^*(\cdot) = f(\cdot)$
- If $f(\cdot) = \Omega(\cdot)$ is a norm, then $f^*(\cdot) = \iota_{\mathcal{B}_*(0,1)}(\cdot)$, *i.e.*, it is the indicator function of the dual norm unit ball, where the dual norm Ω^* is defined by:

$$\Omega^*(z) = \sup_{x:\Omega(x)\leq 1} x^\top z = \iota^*_{\mathcal{B}(0,1)}$$

and

$$\iota_{\mathcal{B}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{B} \\ +\infty & \text{otherwise} \end{cases}, \text{ where } \mathcal{B} = \{x \in \mathbb{R}^n : \Omega(x) \leq 1\}$$

KKT: Karush-Khun-Tucker (KKT) conditions

- **Primal solution :** $\hat{\beta}^{(\lambda)} \in \mathbb{R}^p$
- **Dual solution :** $\hat{\theta}^{(\lambda)} \in \mathcal{D} \subset \mathbb{R}^n$

Primal/Dual link: $\boxed{y = X\hat{\beta}^{(\lambda)} + \lambda\hat{\theta}^{(\lambda)}}$

Necessary and sufficient optimality conditions:

KKT/Fermat: $$\boxed{\forall j \in [p],\ X_j^\top \hat{\theta}^{(\lambda)} \in \begin{cases} \{\mathrm{sign}(\hat{\beta}_j^{(\lambda)})\} & \text{if} \quad \hat{\beta}_j^{(\lambda)} \neq 0, \\ [-1, 1] & \text{if} \quad \hat{\beta}_j^{(\lambda)} = 0. \end{cases}}$$

Mother of safe rules: the KKT implies that If $\lambda \geq \lambda_{\max} = \|X^\top y\|_\infty = \max_{j\in[p]} |X_j^\top \hat{\theta}^{(\lambda)}|$, then $0 \in \mathbb{R}^p$ is the (unique here) primal solution

Proof in next slide (if any interest)

Proof Fermat/KKT + primal/dual link

$$\text{Lagrangian}: \quad \mathcal{L}(z,\beta,\theta) := \underbrace{\frac{1}{2}\|z\|^2}_{f(z)} + \lambda \underbrace{\|\beta\|_1}_{\Omega(\beta)} + \lambda\theta^\top(y - X\beta - z).$$

A saddle point $(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)})$ of the Lagrangian satisfies:

$$\begin{cases} 0 &= \frac{\partial \mathcal{L}}{\partial z}(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)}) = \nabla f(z^\star) = z^\star - \lambda\hat{\theta}^{(\lambda)}, \\ 0 &\in \partial\mathcal{L}(z^\star, \cdot, \hat{\theta}^{(\lambda)})(\hat{\beta}^{(\lambda)}) = -\lambda X^\top\hat{\theta}^{(\lambda)} + \lambda\partial\Omega(\hat{\beta}^{(\lambda)}) \\ 0 &= \frac{\partial \mathcal{L}}{\partial \theta}(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)}) = y - X\hat{\beta}^{(\lambda)} - z^\star. \end{cases}$$

Hence, $y - X\hat{\beta}^{(\lambda)} = z^\star = \lambda\hat{\theta}^{(\lambda)}$ and $X^\top\hat{\theta}^{(\lambda)} \in \partial\Omega(\hat{\beta}^{(\lambda)})$ so

$$\forall j \in \{1, \dots, p\}, \quad X_j^\top\hat{\theta}^{(\lambda)} \in \partial\|\cdot\|_1(\hat{\beta}^{(\lambda)})$$