Exploiting structure in sparse GLMs for fast & safe support identification

Mathurin Massias (University of Genoa)

Joint work with:

Alexandre Gramfort (INRIA)
Joseph Salmon (Université de Montpellier)
Samuel Vaiter (CNRS, UMB)

Table of Contents

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
- sparsity: for λ large enough, $\|\hat{\beta}\|_0 \ll p$
- popular solver in ML: coordinate descent (used here)

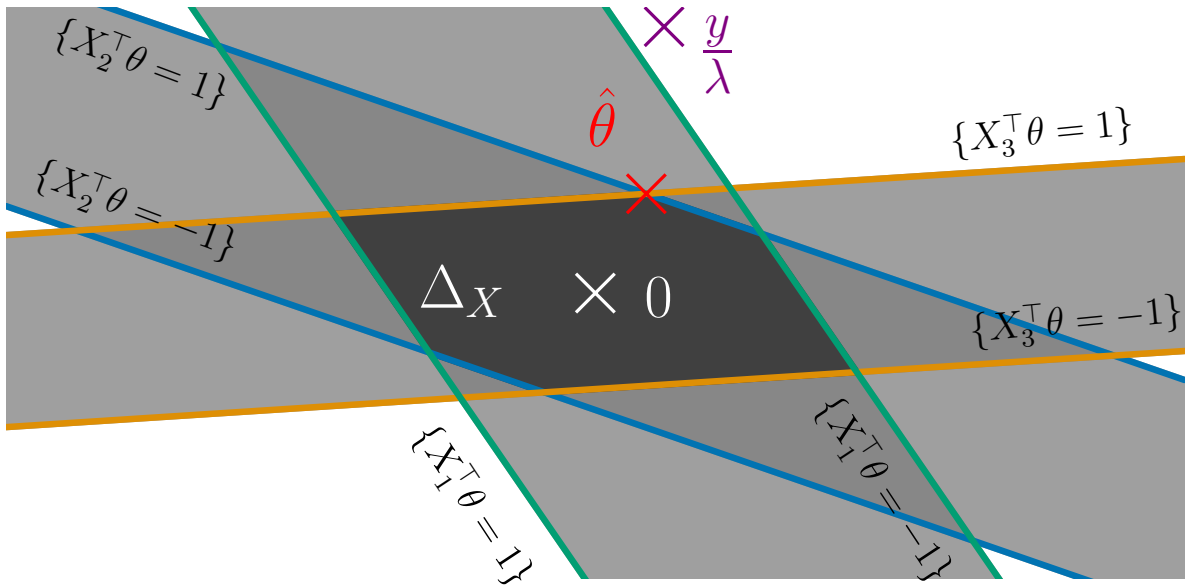
[1] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

[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**

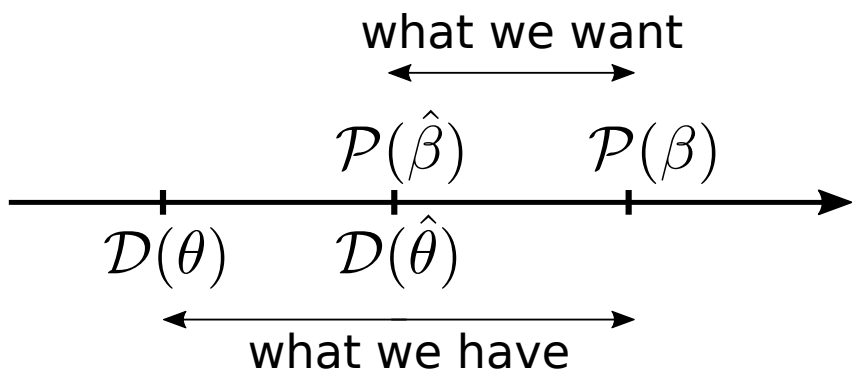


Example: $n = 2, p = 3$

Primal-dual links

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

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



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

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

Choice of dual point

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

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

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

[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

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

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

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

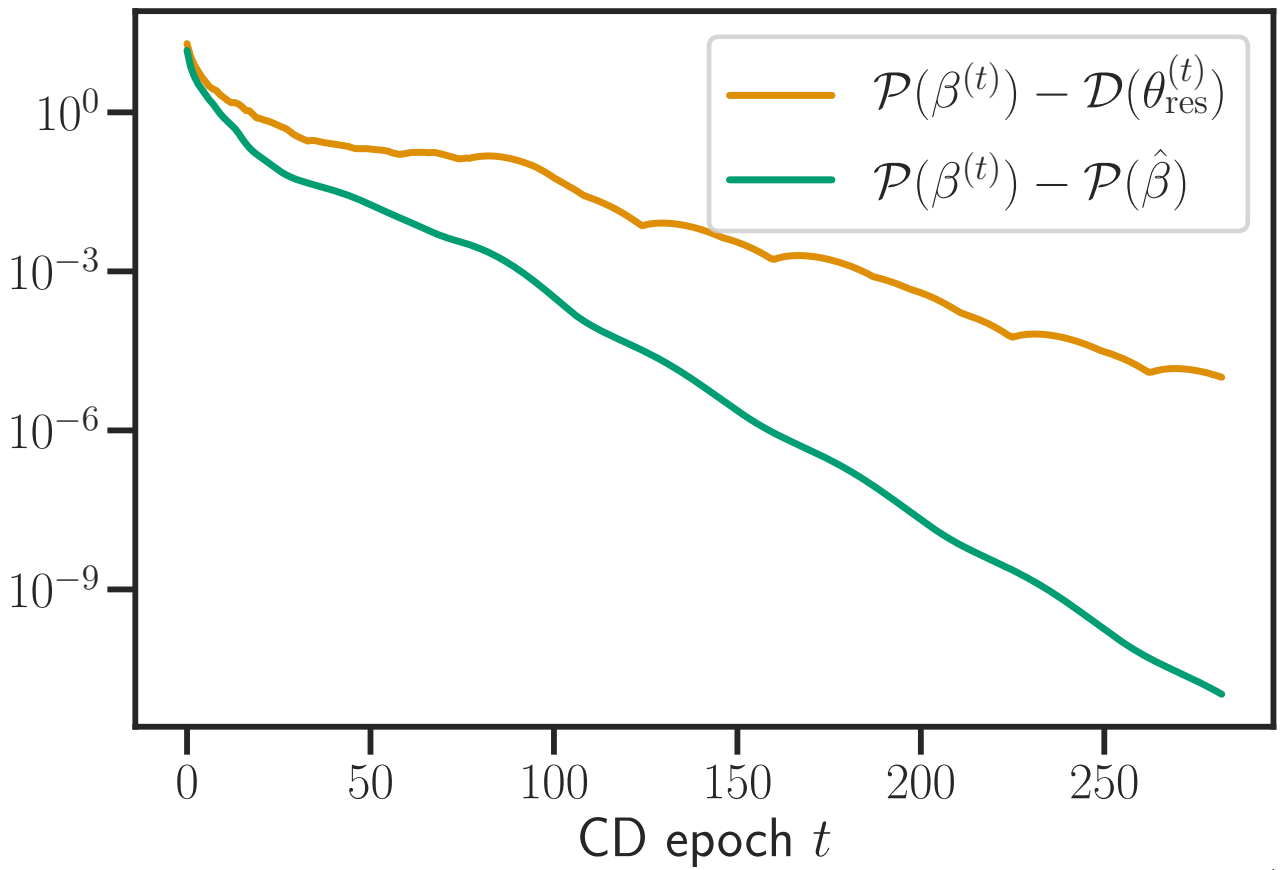
residuals rescaling

- 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.

Slower convergence of dual

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

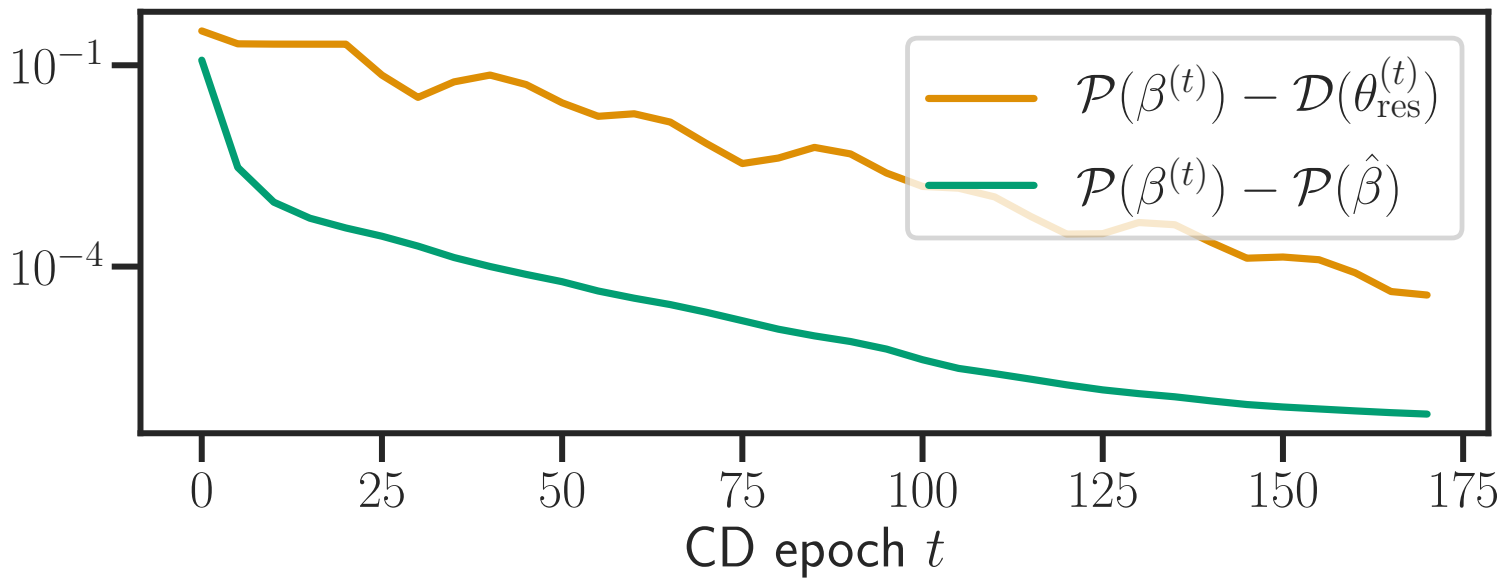


Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

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

Slower convergence of dual

Unconstrained dual of Elastic net:



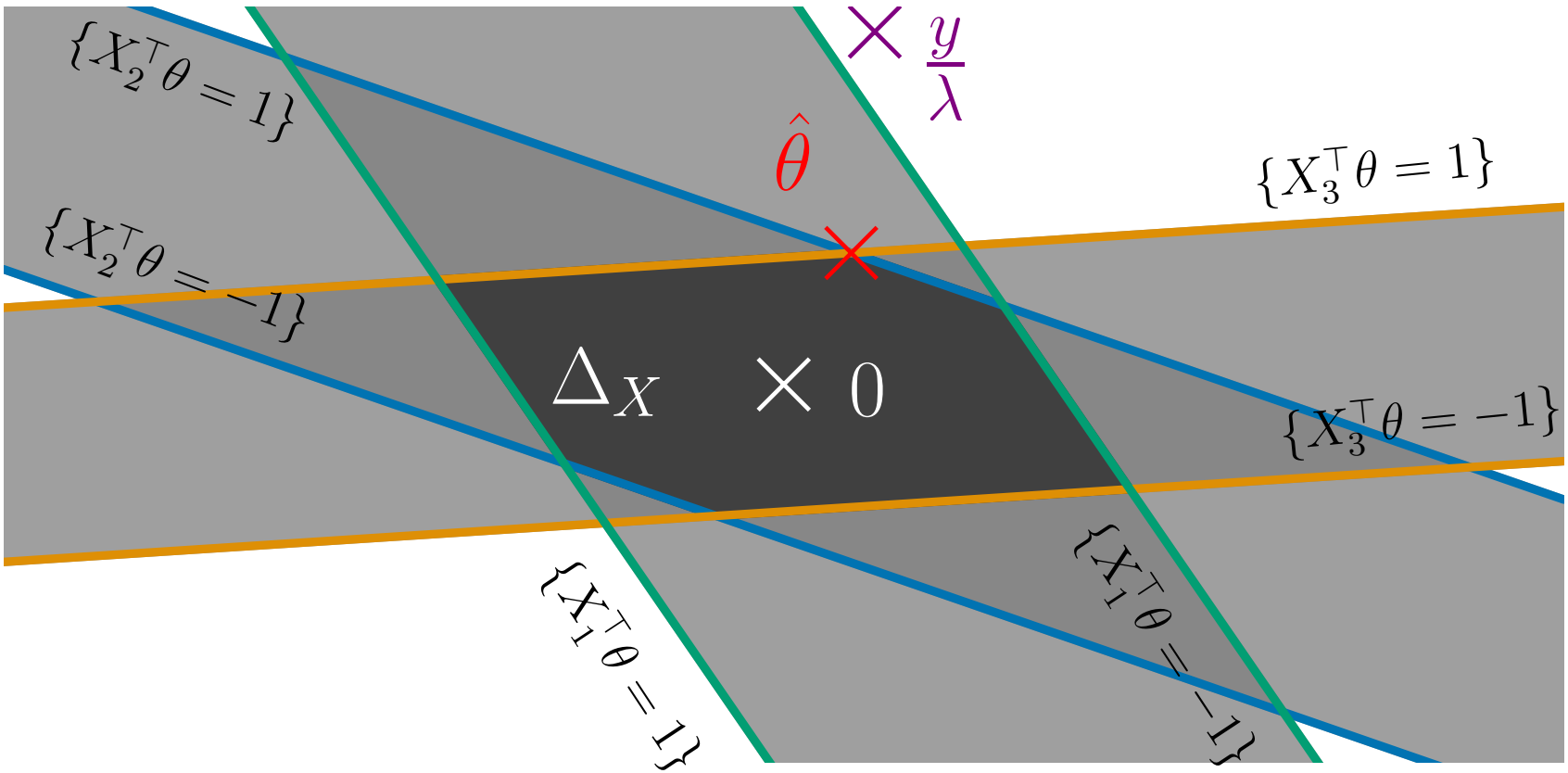
news20 dataset:

$$p = 632982, \quad n = 19996, \quad \lambda = \lambda_{\max}/20, \rho = 0.1$$

Table of Contents

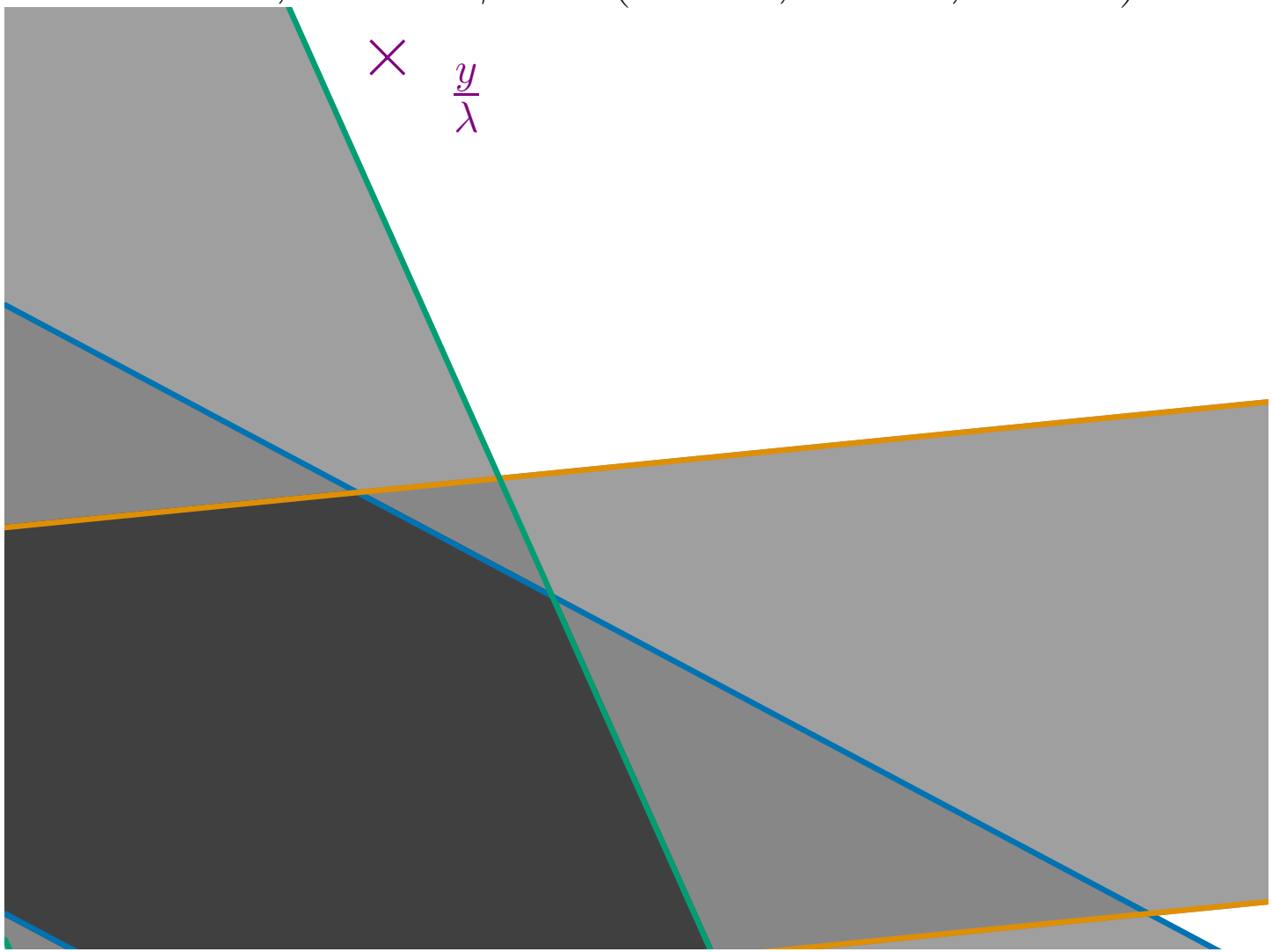
Residuals regularity, sign identification

$$n = 2, p = 3, \hat{\beta} = (0, -0.6, 1.3)$$



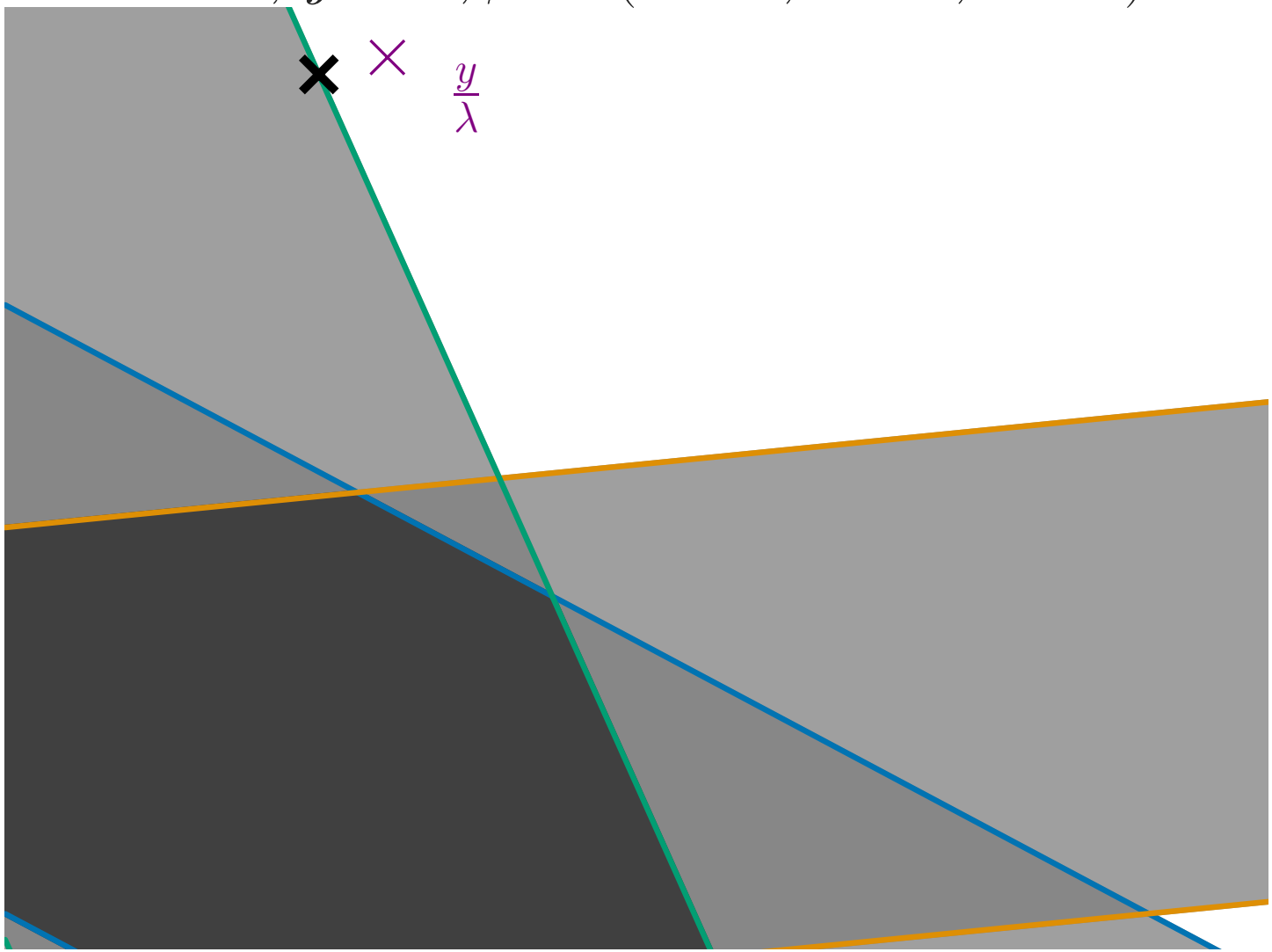
Residuals regularity, sign identification

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



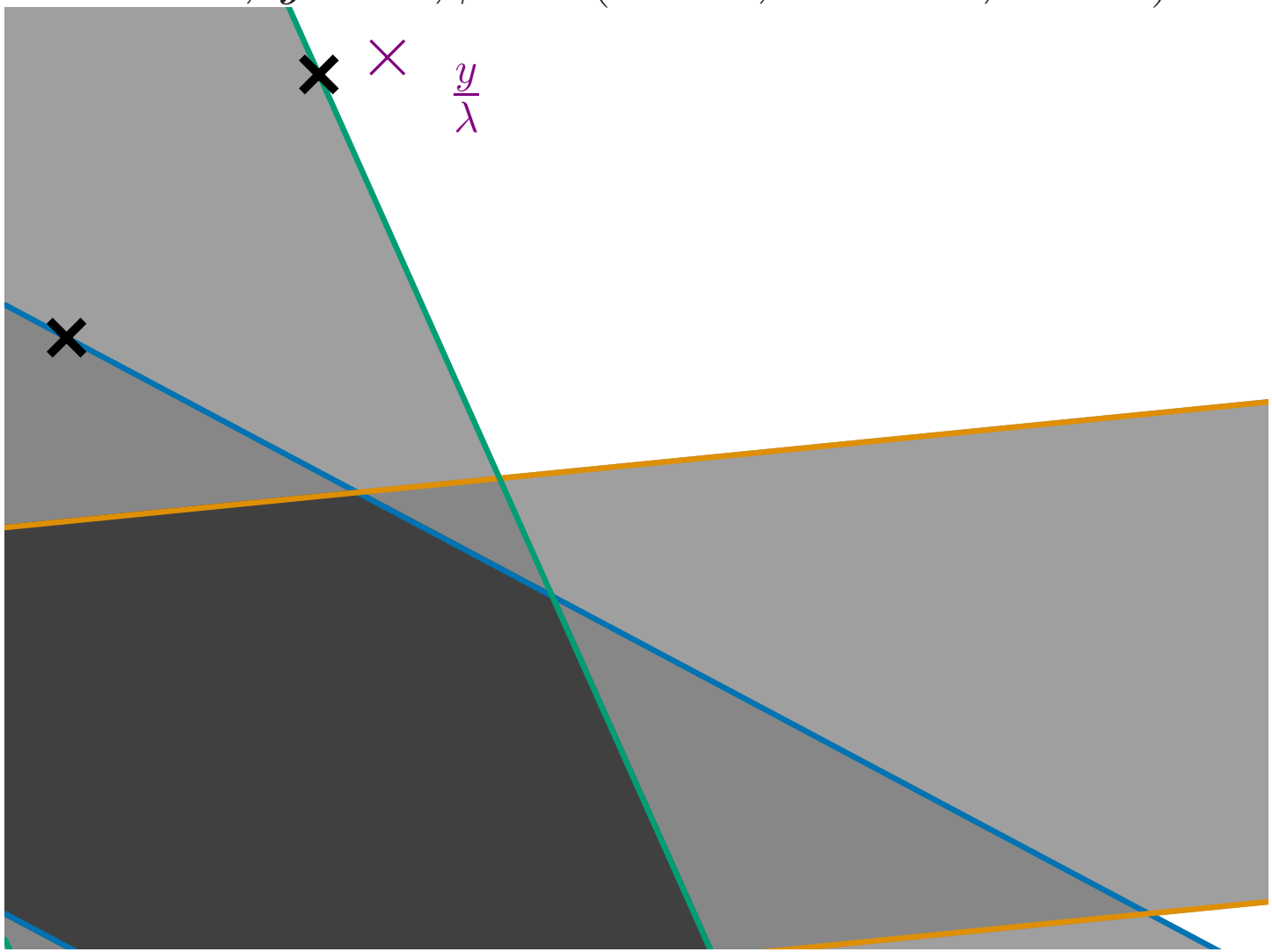
Residuals regularity, sign identification

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



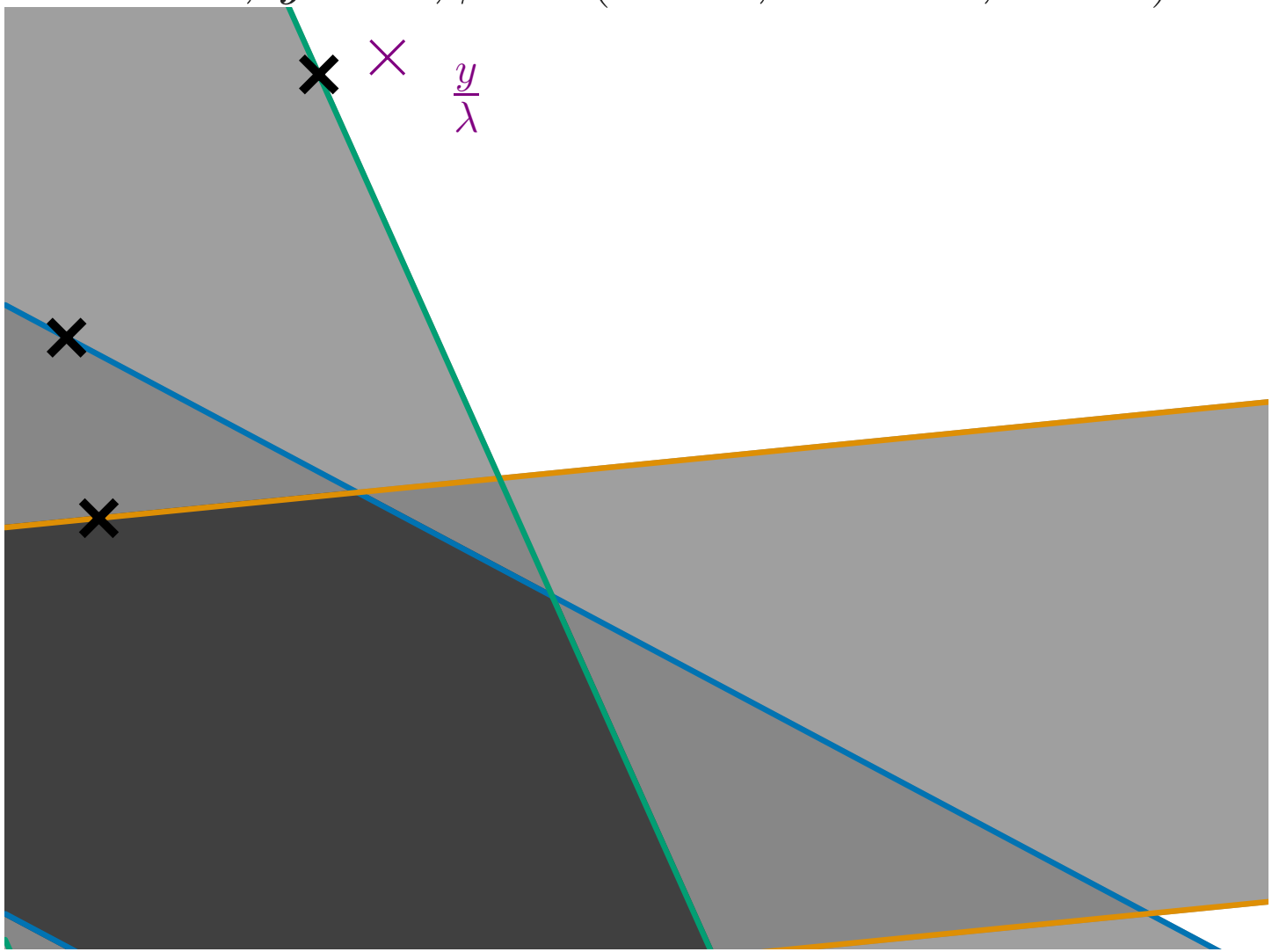
Residuals regularity, sign identification

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



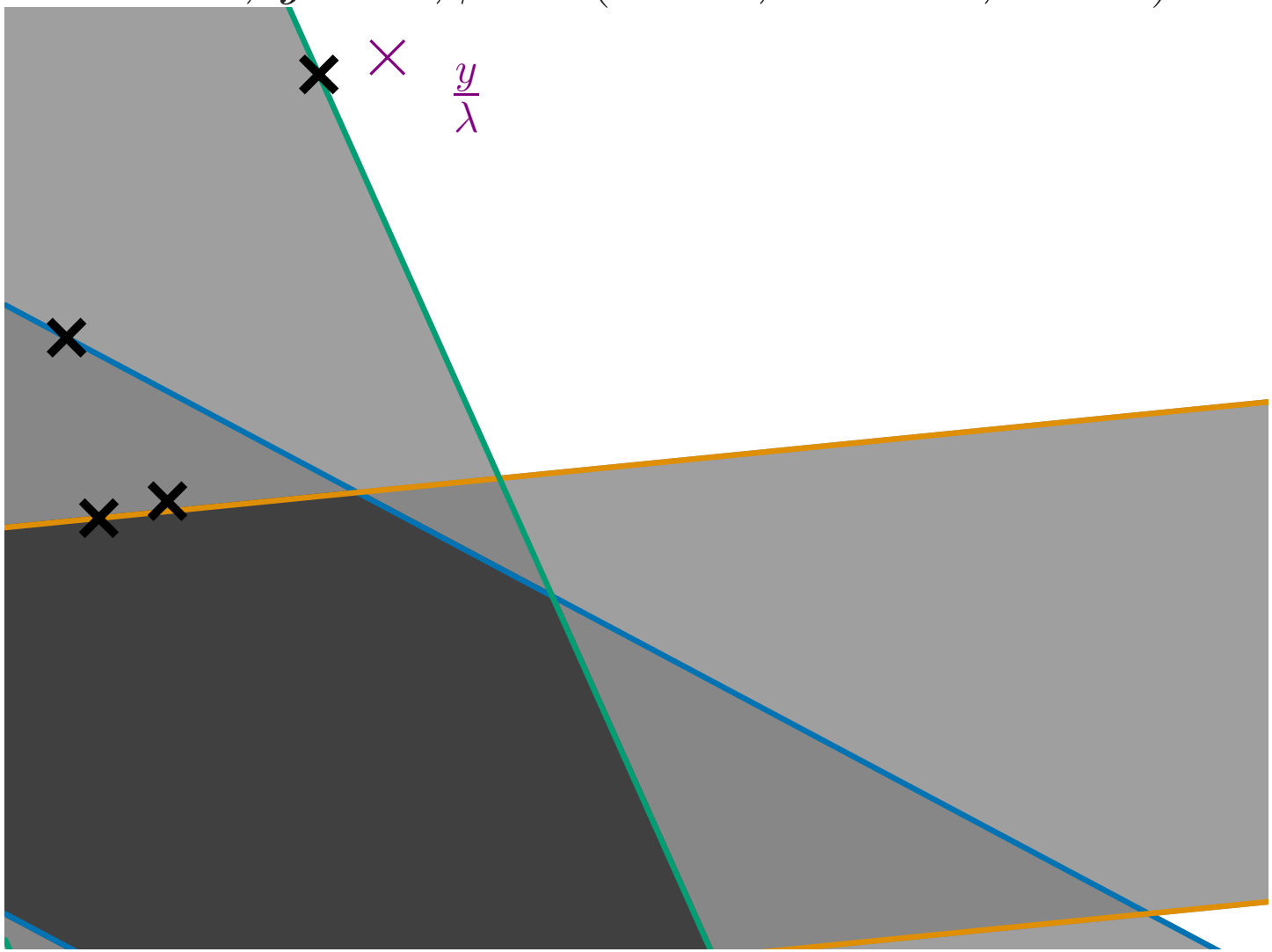
Residuals regularity, sign identification

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



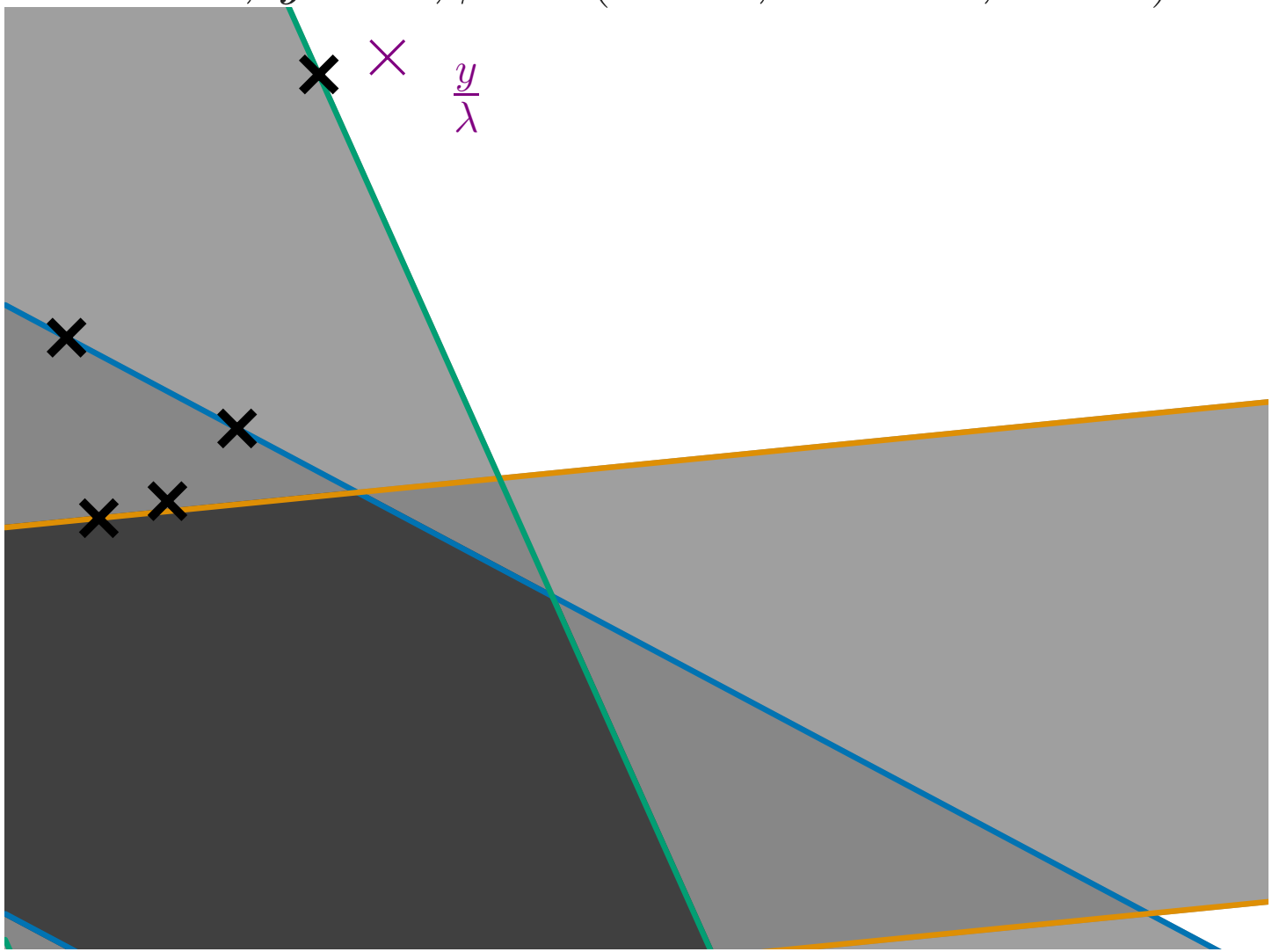
Residuals regularity, sign identification

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



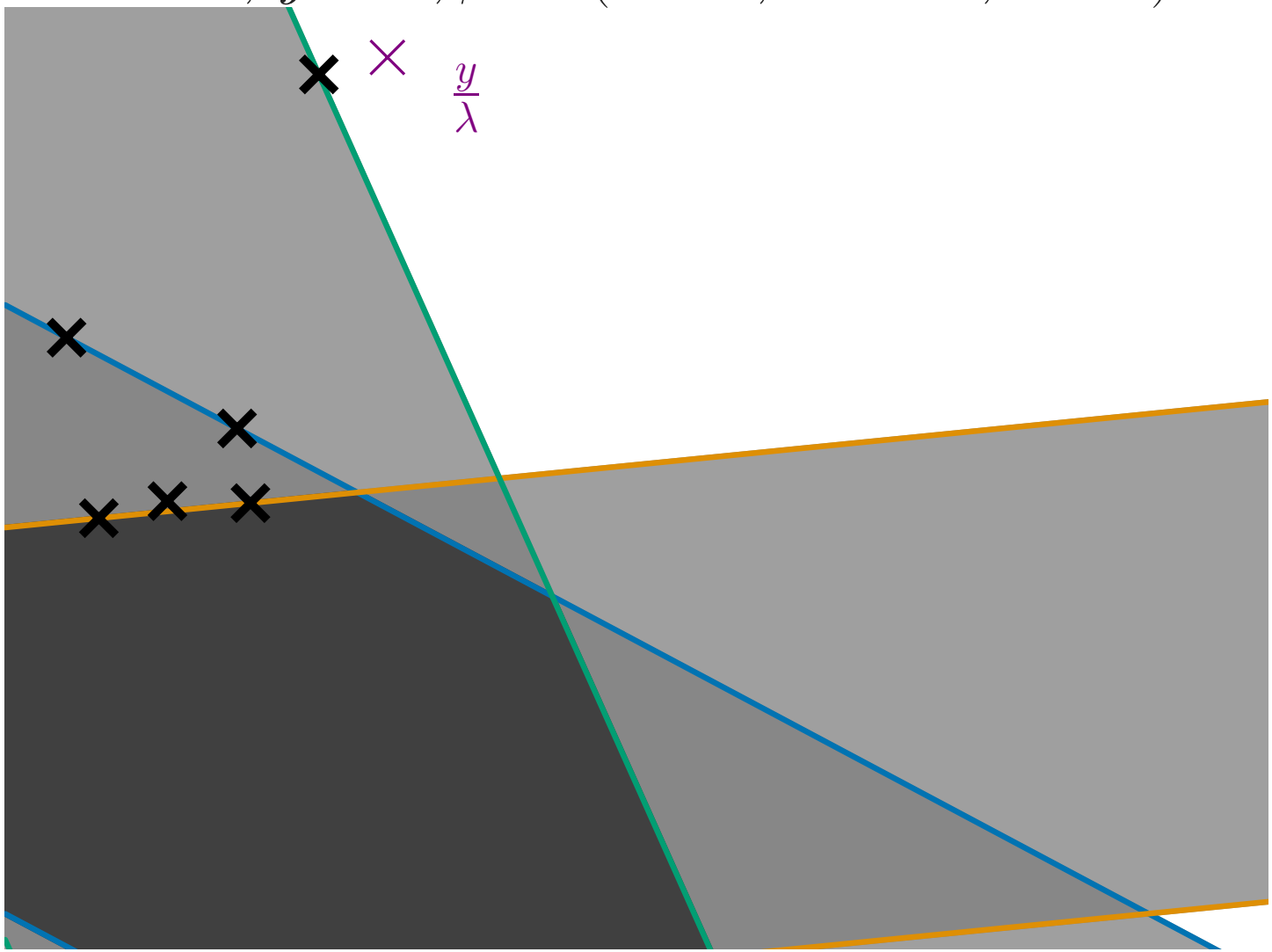
Residuals regularity, sign identification

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



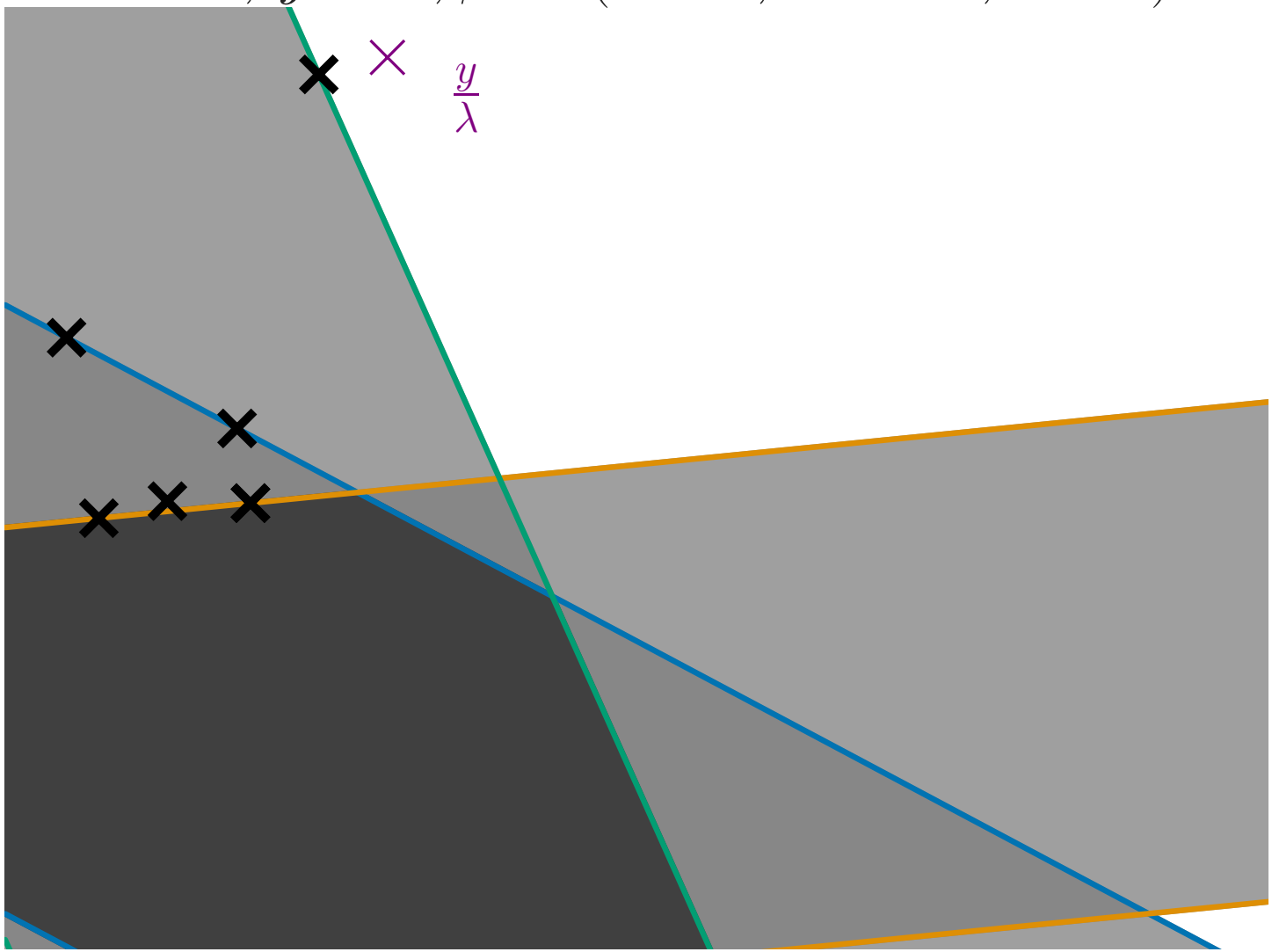
Residuals regularity, sign identification

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



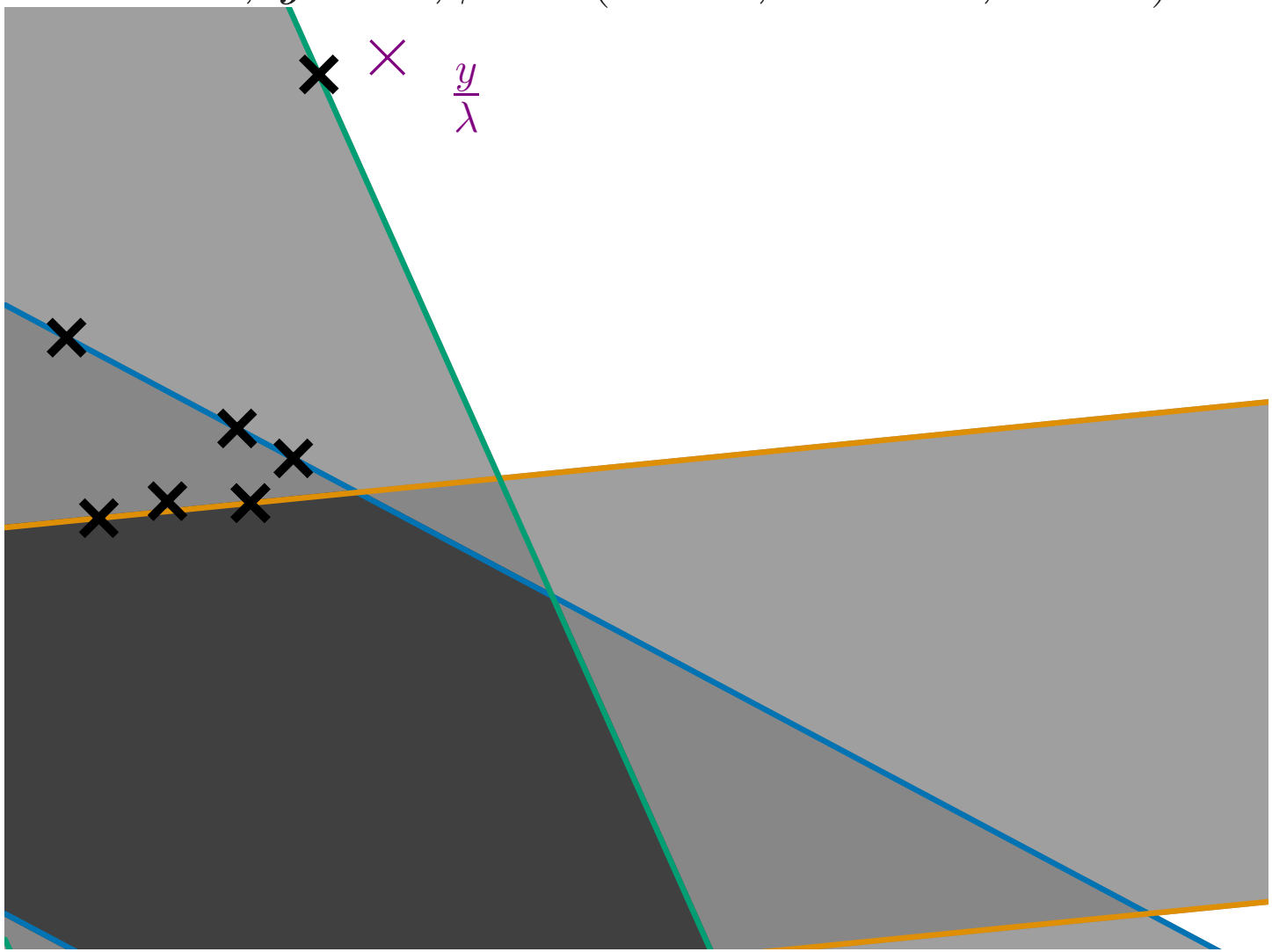
Residuals regularity, sign identification

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



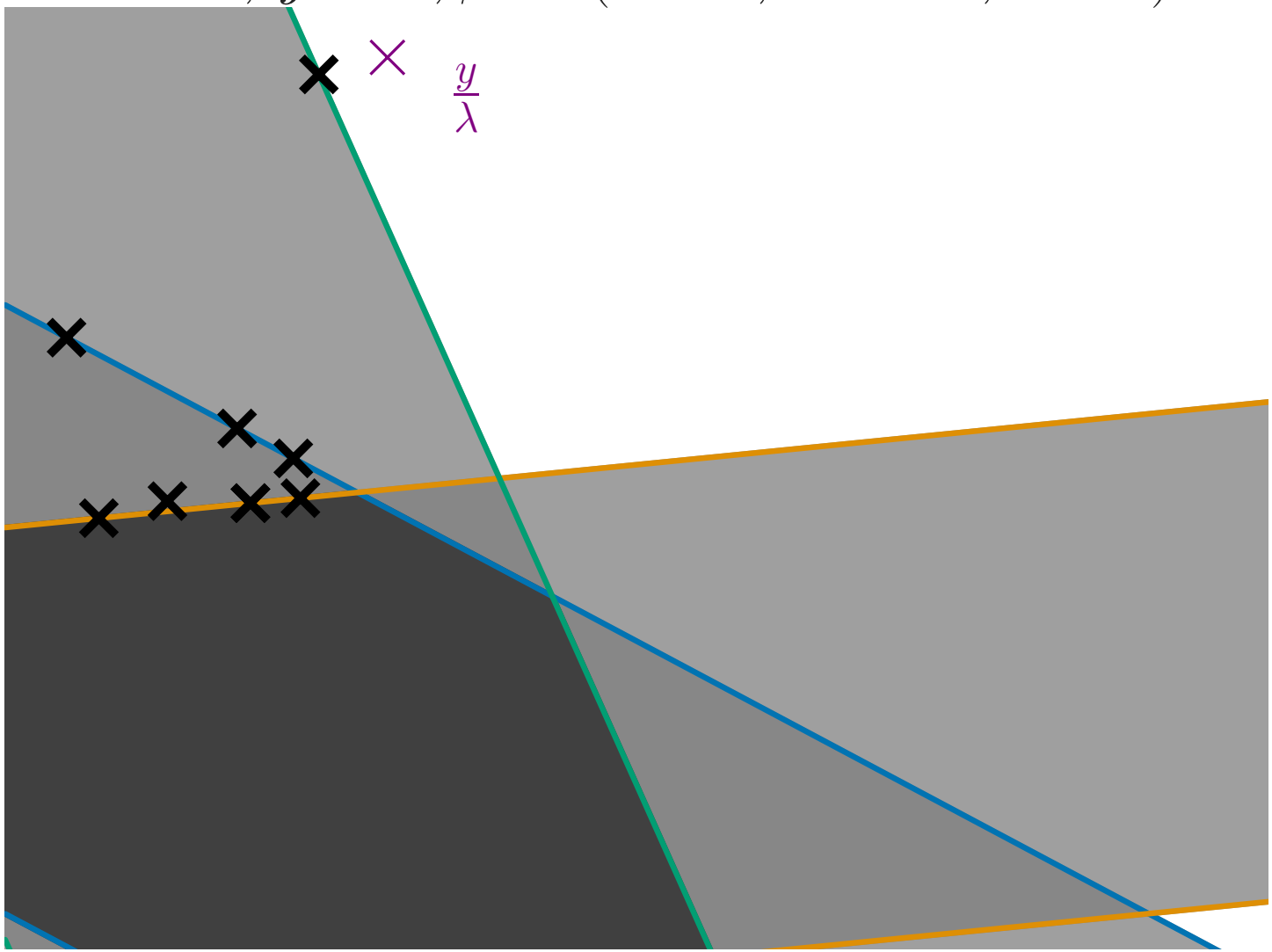
Residuals regularity, sign identification

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



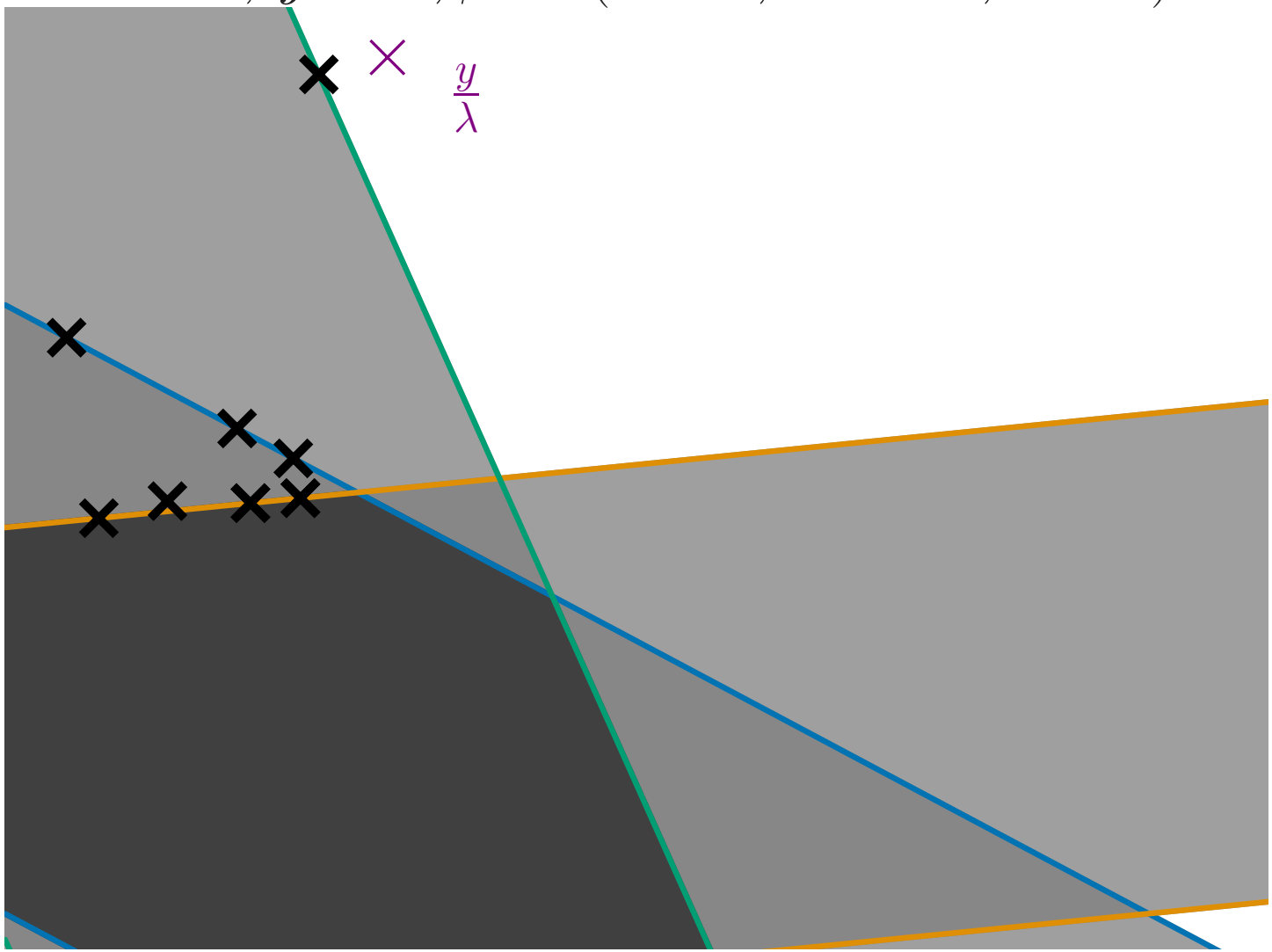
Residuals regularity, sign identification

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



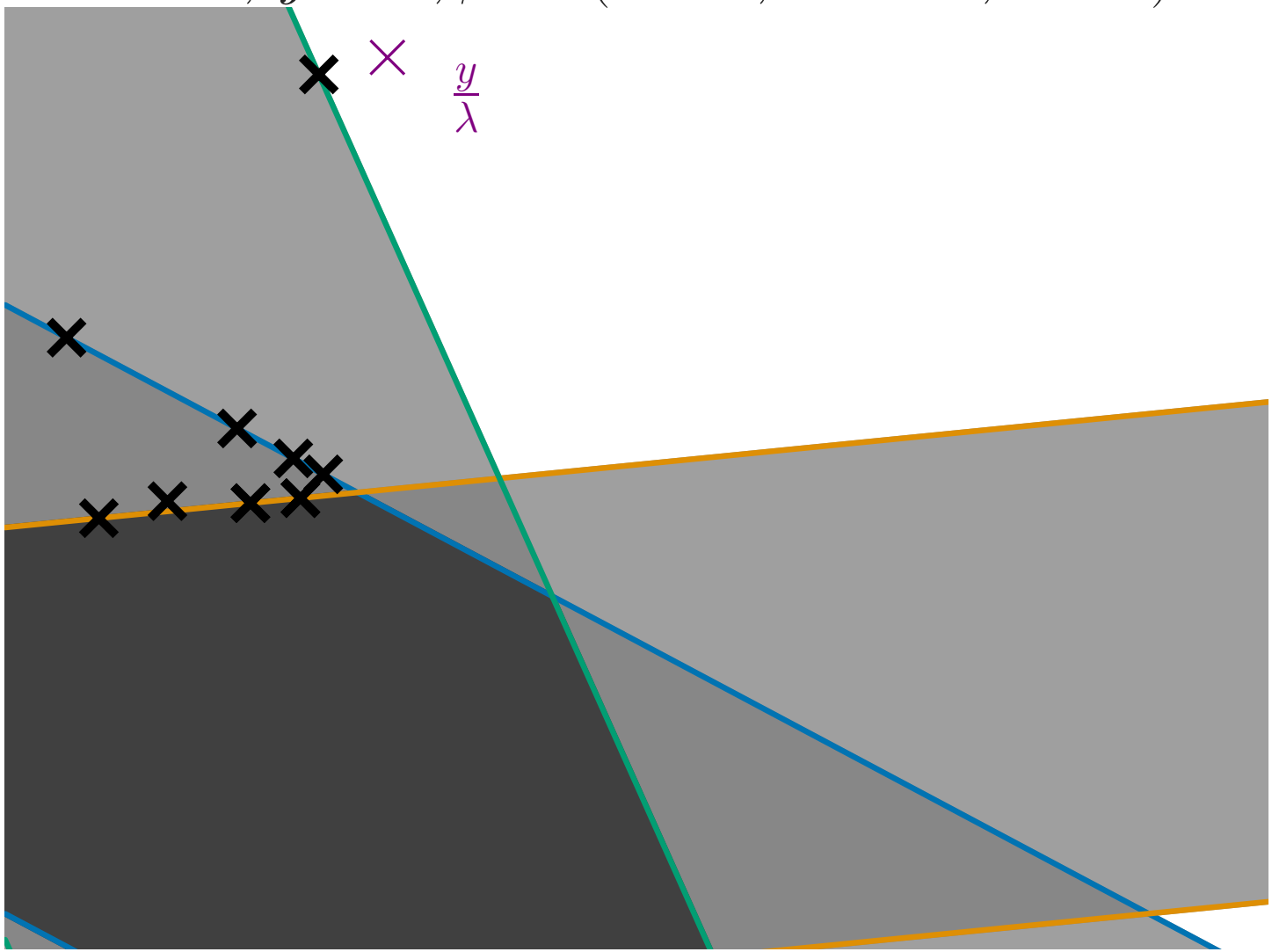
Residuals regularity, sign identification

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



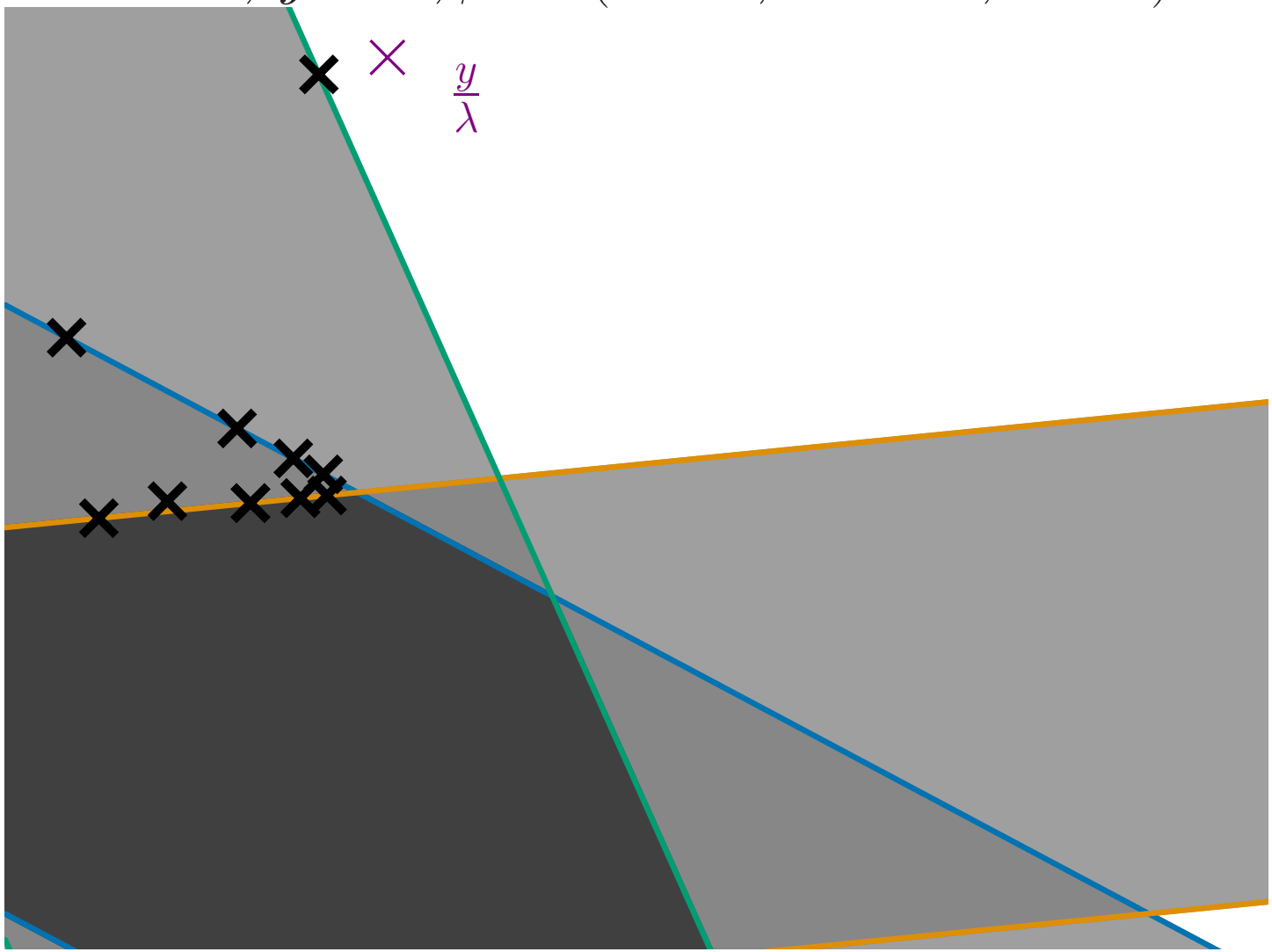
Residuals regularity, sign identification

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



Residuals regularity, sign identification

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



VAR regularity in residuals

Theorem[4]

Under uniqueness assumption, ISTA/CD achieves sign id.: $\operatorname{sign} \beta_j^{(t)} = \operatorname{sign} \hat{\beta}_j$. Then, Lasso residuals are Vector AutoRegressive (**VAR**):

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

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

We do not know when the sign is identified, n points is high

Need a cheaper solution $\hookrightarrow$ **extrapolation**

[4] M. Massias, A. Gramfort, and J. Salmon. "Celer: a fast solver for the Lasso with dual extrapolation". In: *ICML*. 2018, pp. 3321–3330.

Simple example: extrapolation in 1D

1D autoregressive process:

$$x^{(t)} = ax^{(t-1)} + b \underset{t\to\infty}{\to} x^*$$

we have

$$x^{(t)} - x^* = a(x^{(t-1)} - x^*)$$
$$x^{(t-1)} - x^* = a(x^{(t-2)} - x^*)$$

"Aitken's Δ^2": 2 unknowns, so 2 eqs or 3 points $x^{(t)}, x^{(t-1)}, x^{(t-2)}$ are enough to find x^*![5]

[5] A. Aitken. "On Bernoulli's numerical solution of algebraic equations". In: *Proceedings of the Royal Society of Edinburgh* 46 (1926), pp. 289–305.

Aitken application

$$\lim_{t\to\infty} \sum_{i=0}^{t} \frac{(-1)^i}{2i+1} = \frac{\pi}{4} = 0.785398...$$

t	$\sum_{i=0}^{t} \frac{(-1)^i}{2i+1}$	Δ^2
0	1.0000	–
1	0.66667	–
2	0.86667	**0.79**167
3	**0.72**381	**0.78**333
4	0.83492	**0.78**631
5	**0.74**401	**0.78**492
6	0.82093	**0.785**68
7	**0.75**427	**0.785**22
8	0.81309	**0.785**52
9	**0.76**046	**0.785**31

Generalization[6] to VAR $r^{(t)} \in \mathbb{R}^n$

- fix $K = 5$ (small)
- keep track of K past residuals $r^{(t)}, \ldots, r^{(t+1-K)}$
- $U^{(t)} = [r^{(t+1-K)} - r^{(t-K)}, \ldots, r^{(t)} - r^{(t-1)}] \in \mathbb{R}^{n \times K}$
- solve $(U^{(t)})^\top U^{(t)} z = \mathbf{1}_K$
- $c = z / z^\top \mathbf{1}_K$

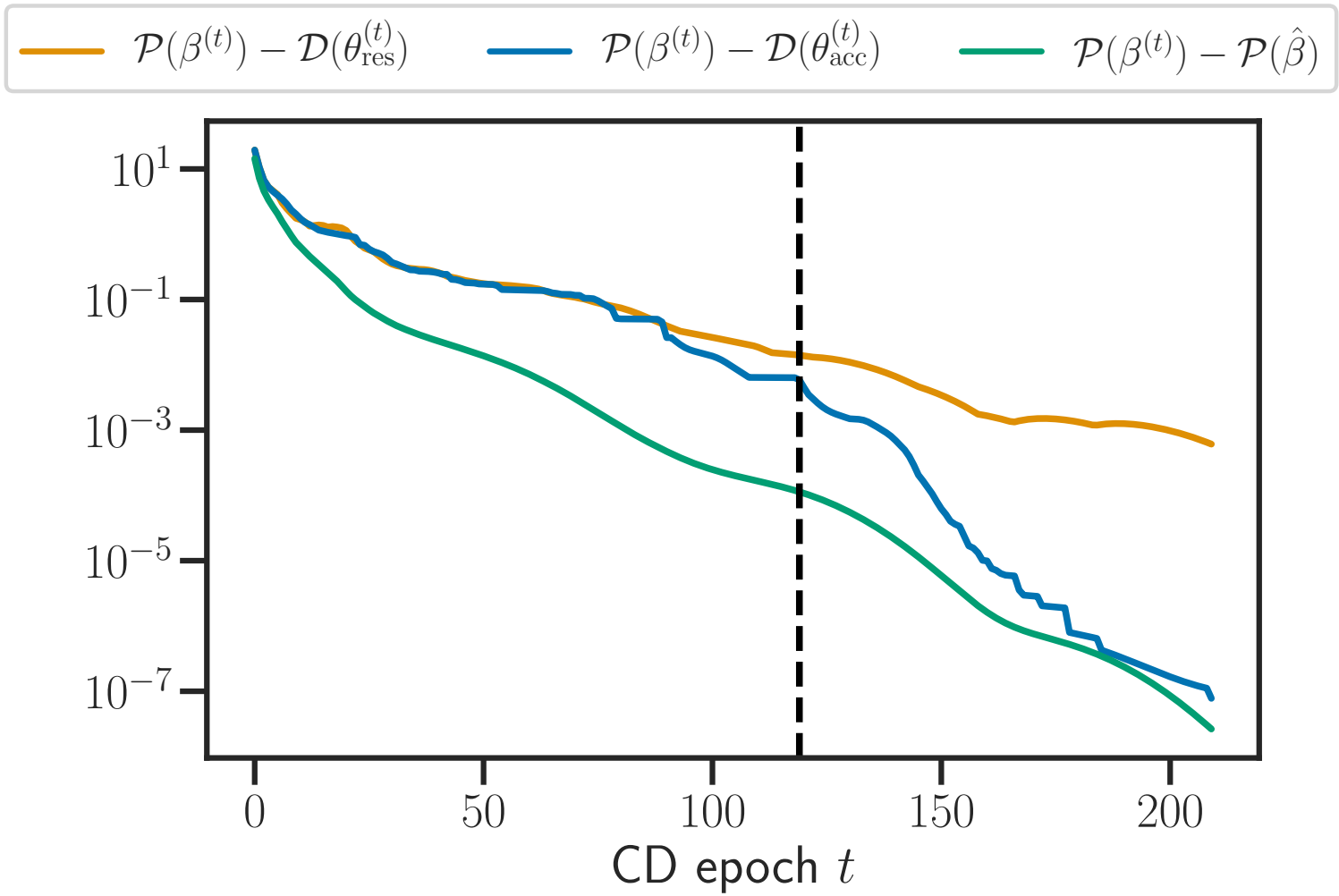
$$r_{\text{accel}}^{(t)} \triangleq \sum_{k=1}^{K} c_k r^{(t+1-k)}$$

$$\boxed{\theta_{\text{accel}}^{(t)} \triangleq r_{\text{accel}}^{(t)} / \max(\lambda, \|X^\top r_{\text{accel}}^{(t)}\|_\infty)}$$

Cost: $\mathcal{O}(K^3 + K^2 n + np)$

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

Dual extrapolation for the Lasso



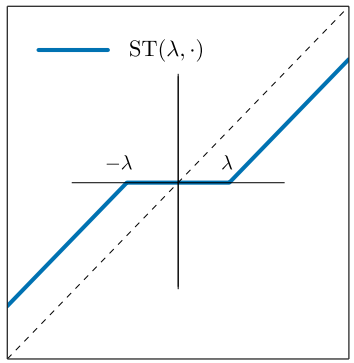
Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

VAR after sign identification

Wlog, support of $\hat{\beta}$: $\{1, \ldots, S\}$ (other coordinates stay at 0)

Consider 1 epoch of CD:

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



Decomposed into non-zero coordinate updates

$$\beta^{(t)} = \tilde{\beta}^{(0)} \xrightarrow{1} \tilde{\beta}^{(1)} \xrightarrow{2} \ldots \xrightarrow{S} \tilde{\beta}^{(S)} = \beta^{(t+1)}$$

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

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

VAR after sign identification

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

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

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

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

Table of Contents

VAR for other GLMs

sparse Log. reg. $\displaystyle\arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \beta^\top x_i)\right) + \lambda \|\beta\|_1$

Multi-task Lasso $\displaystyle\arg\min_{\mathrm{B} \in \mathbb{R}^{p \times T}} \frac{1}{2} \|Y - X\mathrm{B}\|^2 + \lambda \|\mathrm{B}\|_{2,1}$

Log. reg. CD update after sign ID:

$$\tilde{\beta}_s^{(s)} = \tilde{\beta}_s^{(s-1)} - \frac{\gamma}{\|X_s\|^2} X_s^\top \nabla F(X\tilde{\beta}^{(s-1)}) - \frac{\gamma}{\|X_s\|^2} \lambda \operatorname{sign}(\hat{\beta}_s)$$

∇F not linear in $X\beta$ if F is not a quadratic

Structure for other GLMs

Solution: linearization of ∇F around optimum

$$\nabla F(X\beta) = \nabla F(X\hat{\beta}) + \underbrace{D}_{n\times n,\text{ diagonal}} (X\beta - X\hat{\beta}) + o(X\beta - X\hat{\beta})$$

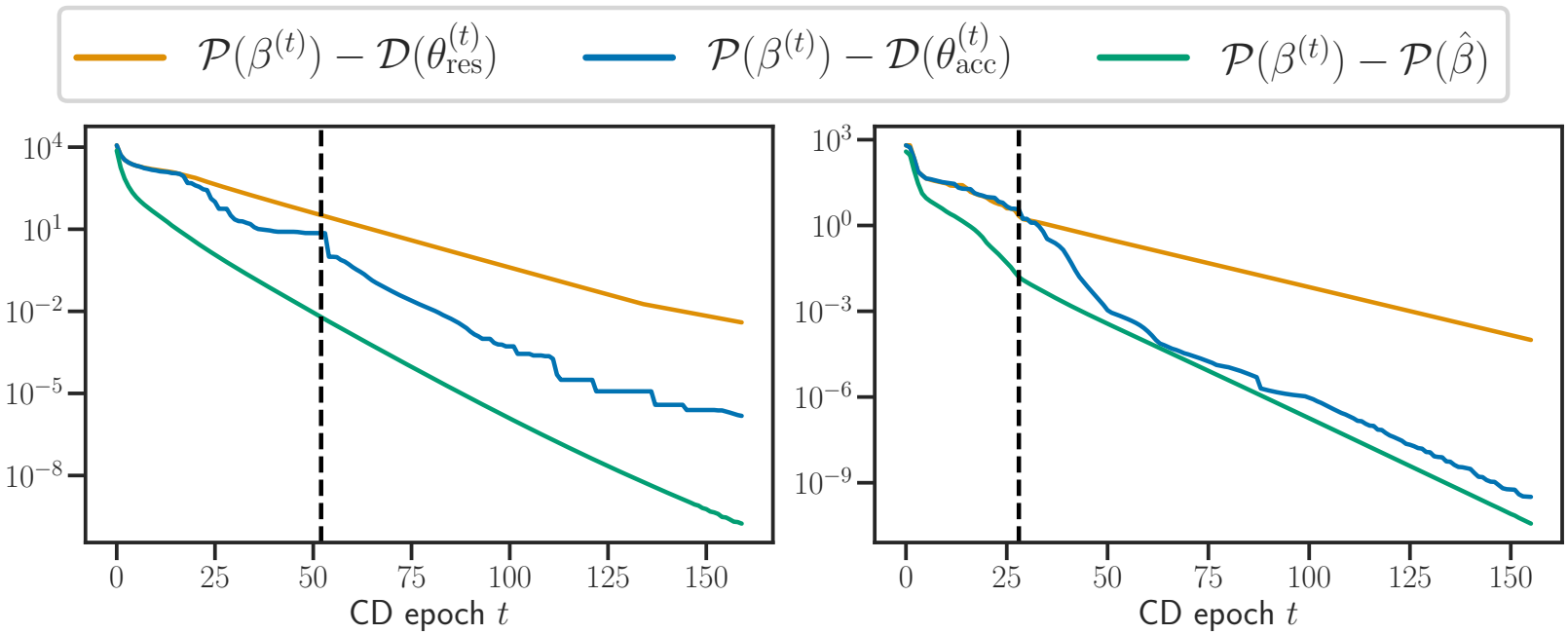
Leads to **asymptotic VAR sequence**:

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

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

Applicability to other models

Asymptotic VAR structure is still exploitable[7]



logreg, *rcv1* dataset:

$p = 20\text{k}, \quad n = 20\text{k}$

$\lambda = \lambda_{\max}/20 \;\; (\|\hat{\beta}\|_0 = 395)$

MTL, MEG data:

$p = 7498, \quad n = 305$

$\lambda = \lambda_{\max}/10 \;\; (\|\hat{\mathrm{B}}\|_{2,0} = 45)$

[7] M. Massias et al. "Dual extrapolation for sparse Generalized Linear Models". In: *submission to JMLR* (2019).

Table of Contents

Speeding-up solvers

Two approaches:

- safe screening[8,9] (**backward approach**): remove feature j when it is certified that $\hat{\beta}_j = 0$
- working set[10] (**forward approach**): focus on j's for which it is very likely that $\hat{\beta}_j \neq 0$.

[8] 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.

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

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

Key to identifying features

Equicorrelation set[11]

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

- 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 $\#E \ll p$

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

Duality again: gap screening

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

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

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

[12] 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 know in advance $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

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

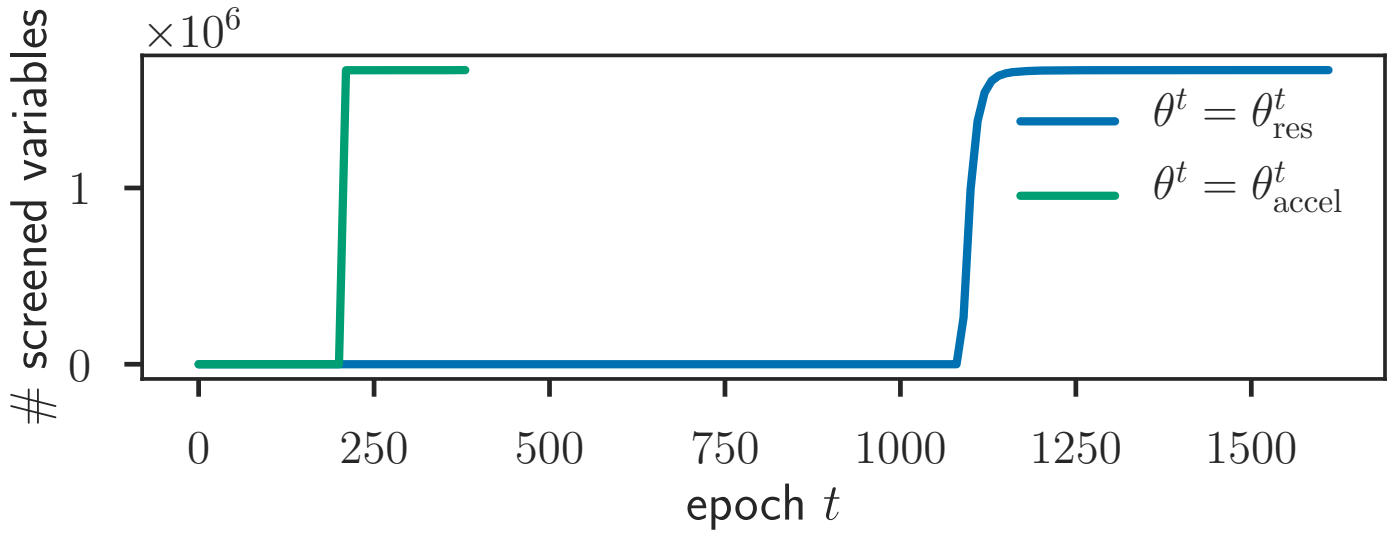
$$\boxed{\forall (\beta, \theta) \in \mathbb{R}^p \times \Delta_X, \quad |X_j^\top \theta| < 1 - \|X_j\| \rho \Rightarrow \hat{\beta}_j = 0}$$

[12] 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[13]

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

better dual point $\Rightarrow$ better safe screening

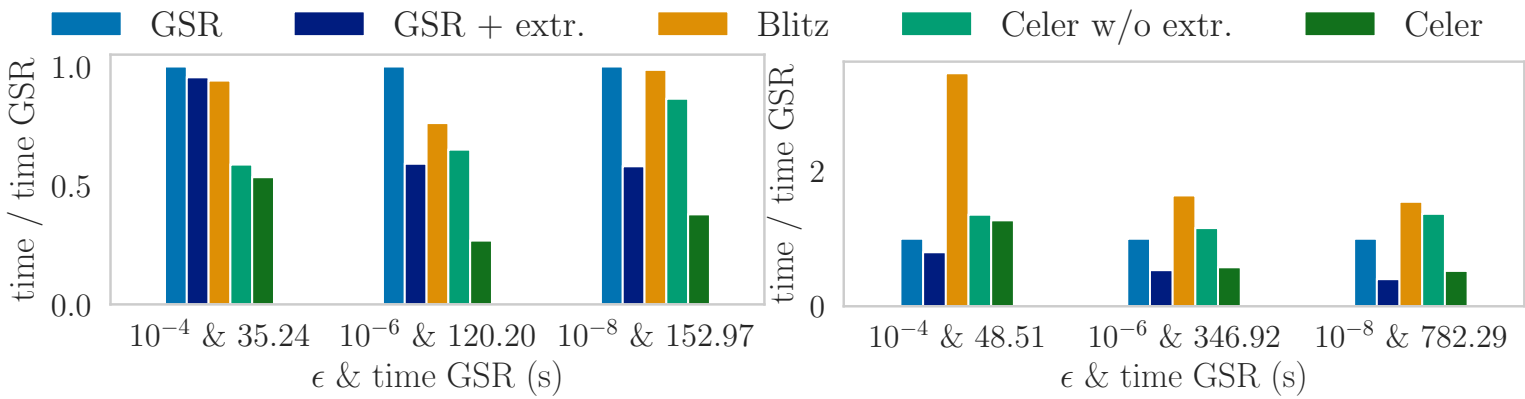


Finance dataset: $p = 1.5 \times 10^6, n = 1.5 \times 10^4$, $\lambda = \lambda_{\max}/5$

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

Working sets

Screening can be used aggressively to define WS, therefore a **better dual point also helps**:



news20 dataset, coarse and fine Lasso paths computation

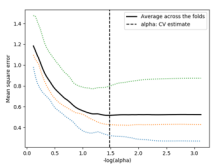
Lasso: $-\nabla F = y - X\beta$, hence: **C**onstraint**E**limination for the **L**asso with **E**xtrapolated **R**esiduals

Online code

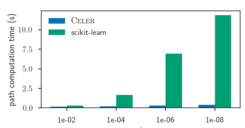
Fast & pip-installable Cython code, continuous integration, bug tracker, code coverage

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

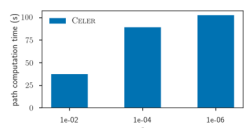
Examples Gallery¶



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 $X\beta^{(t)} \hookrightarrow$ 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
Paper: https://arxiv.org/abs/1907.05830

References I

- Aitken, A. "On Bernoulli's numerical solution of algebraic equations". In: *Proceedings of the Royal Society of Edinburgh* 46 (1926), pp. 289–305.
- 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.

References II

- Mairal, J. "Sparse coding for machine learning, image processing and computer vision". PhD thesis. École normale supérieure de Cachan, 2010.
- Massias, M., A. Gramfort, and J. Salmon. "Celer: a fast solver for the Lasso with dual extrapolation". In: *ICML*. 2018, pp. 3321–3330.
- Massias, M. et al. "Dual extrapolation for sparse Generalized Linear Models". In: *submission to JMLR* (2019).
- 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.

References III

- 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

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