Celer: Fast solver for the Lasso with dual extrapolation

Mathurin Massias
https://mathurinm.github.io
INRIA

Joint work with:
Alexandre Gramfort (INRIA)
Joseph Salmon (Télécom ParisTech)

To appear in ICML 2018

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
- $\lambda > 0$: trade-off parameter between data-fit and regularization
- sparsity: for λ large enough, $\|\hat{\beta}\|_0 \ll p$

Rem: uniqueness is not guaranteed, more later

---

[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

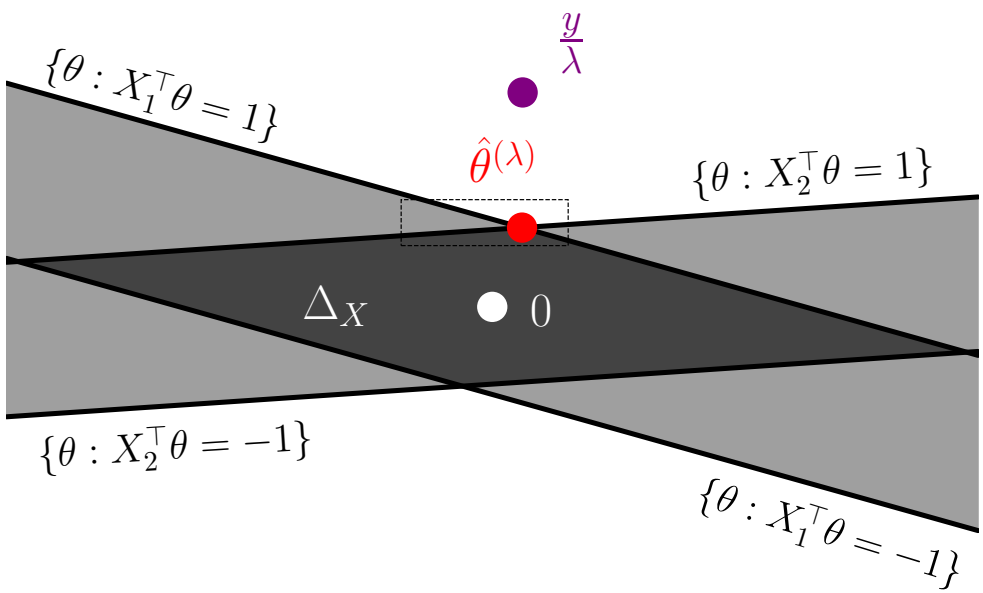
$$\hat{\theta} = \underset{\theta \in \Delta_X}{\arg\max} \underbrace{\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \|y/\lambda - \theta\|^2}_{\mathcal{D}(\theta)}$$

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

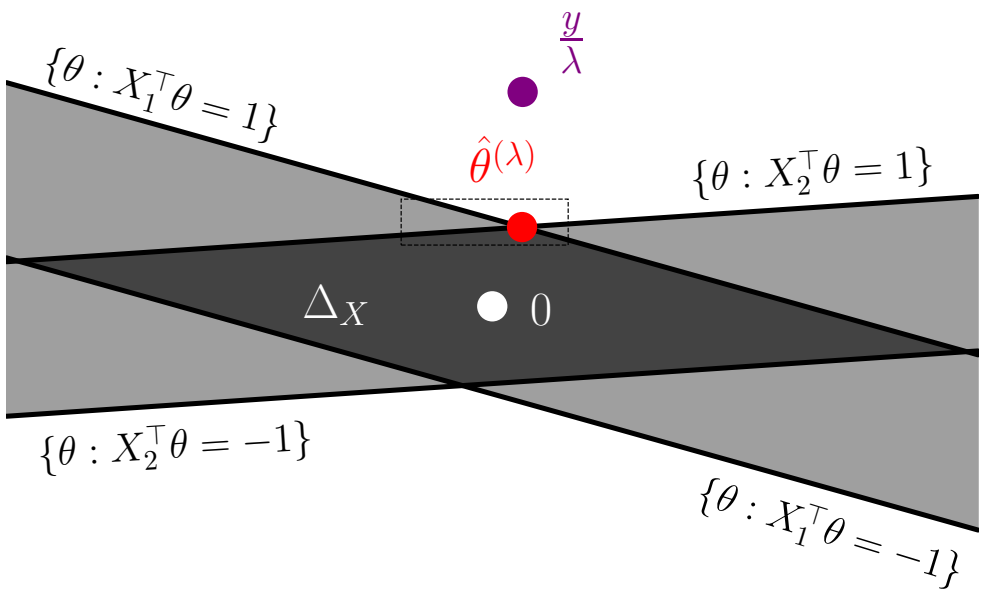


Toy visualization example: $n = p = 2$

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

Solving the Lasso

So-called *smooth + separable* problem

- In signal processing: use ISTA/FISTA[3]
- In ML: state-of-the-art algorithm when X is not an implicit operator: coordinate descent (CD)[4,5]

  Iterative algorithm: minimize $\mathcal{P}(\beta) = \mathcal{P}(\beta_1, \dots, \beta_p)$ *w.r.t.* β_1, then β_2, etc.

[3] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM J. Imaging Sci.* 2.1 (2009), pp. 183–202.

[4] J. Friedman et al. "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332.

[5] P. Tseng. "Convergence of a block coordinate descent method for nondifferentiable minimization". In: *J. Optim. Theory Appl.* 109.3 (2001), pp. 475–494.

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \ldots, T$ **do**

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \dots, T$ **do**

$$\beta_1^t \leftarrow \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{t-1}, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \dots, T$ **do**

$$\beta_1^t \leftarrow \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{t-1}, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_2^t \leftarrow \underset{\beta_2 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2 \quad , \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \dots, T$ **do**

$$\beta_1^t \leftarrow \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{t-1}, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_2^t \leftarrow \underset{\beta_2 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_3^t \leftarrow \underset{\beta_3 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2^t, \beta_3, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \dots, T$ **do**

$$\beta_1^t \leftarrow \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{t-1}, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_2^t \leftarrow \underset{\beta_2 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_3^t \leftarrow \underset{\beta_3 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2^t, \beta_3, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$\vdots$

Solving the Lasso: cyclic CD

To minimize $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$:

Initialisation: $\beta^0 = 0 \in \mathbb{R}^p$

for $t = 1, \dots, T$ **do**

$$\beta_1^t \leftarrow \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{t-1}, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_2^t \leftarrow \underset{\beta_2 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2, \beta_3^{t-1}, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

$$\beta_3^t \leftarrow \underset{\beta_3 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2^t, \beta_3, \dots, \beta_{p-1}^{t-1}, \beta_p^{t-1})$$

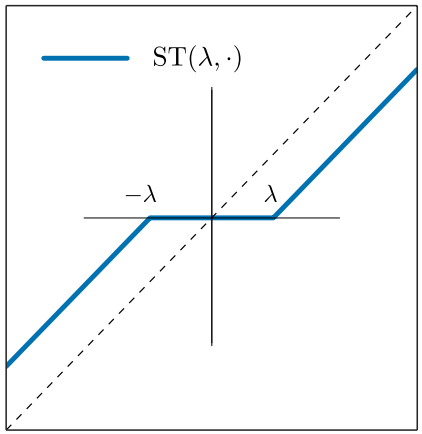
$$\vdots$$

$$\beta_p^t \leftarrow \underset{\beta_p \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^t, \beta_2^t, \beta_3^t, \dots, \beta_{p-1}^t, \beta_p)$$

CD update: soft-thresholding

Coordinate-wise minimization is easy:

$$\beta_j \leftarrow \mathrm{ST}\left(\frac{\lambda}{\|X_j\|^2}, \beta_j + \frac{X_j^\top (y - X\beta)}{\|X_j\|^2}\right)$$



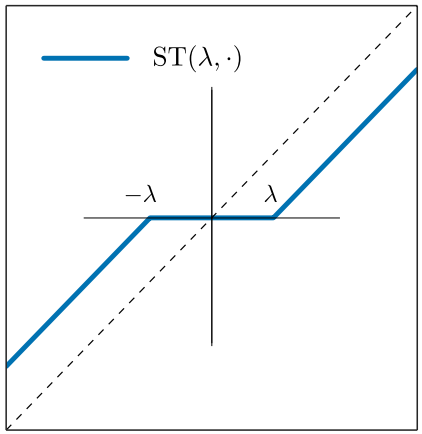
1 update is $\mathcal{O}(n)$

Variants: minimize *w.r.t.* β_j with j chosen at random, or **shuffle** order every epoch (1 epoch $= p$ updates)

CD update: soft-thresholding

Coordinate-wise minimization is easy:

$$\beta_j \leftarrow \mathrm{ST}\left(\frac{\lambda}{\|X_j\|^2}, \beta_j + \frac{X_j^\top (y - X\beta)}{\|X_j\|^2}\right)$$



1 update is $\mathcal{O}(n)$

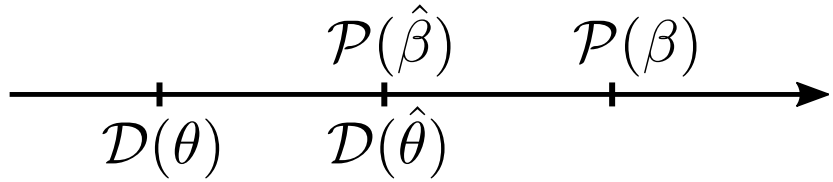
Variants: minimize *w.r.t.* β_j with j chosen at random, or **shuffle** order every epoch (1 epoch $= p$ updates)

When do we stop?

Duality gap as a stopping criterion

For any primal-dual pair (β, θ):

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



The **duality gap** $\mathcal{P}(\beta) - \mathcal{D}(\theta) =: \text{gap}(\beta, \theta)$ is an upper bound of the **suboptimality gap** $\mathcal{P}(\beta) - \mathcal{P}(\hat{\beta})$:

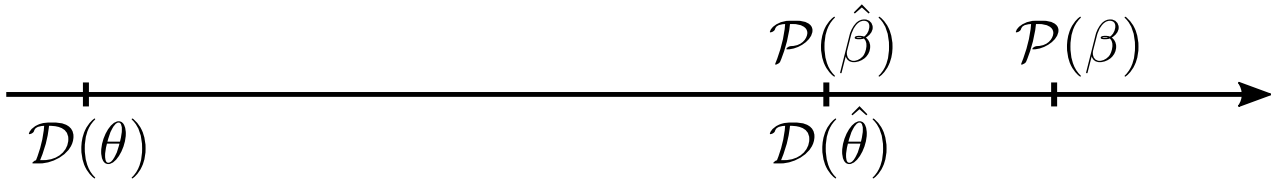
$$\forall \beta, (\exists \theta \in \Delta_X, \text{gap}(\beta, \theta) \leq \epsilon) \Rightarrow \mathcal{P}(\beta) - \mathcal{P}(\hat{\beta}) \leq \epsilon$$

i.e., β is an ϵ-solution

Duality gap as a stopping criterion

For any primal-dual pair (β, θ):

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



The **duality gap** $\mathcal{P}(\beta) - \mathcal{D}(\theta) =: \text{gap}(\beta, \theta)$ is an upper bound of the **suboptimality gap** $\mathcal{P}(\beta) - \mathcal{P}(\hat{\beta})$:

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

i.e., β is an ϵ-solution

Choice of dual point

Primal-dual link at optimum:

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

[6] 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[6]: at epoch t, corresponding to iterate β^t and residuals $r^t := y - X\beta^t$, take

$$\theta = \theta^t_{\text{res}} := r^t/\lambda$$

[6] 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[6]: at epoch t, corresponding to iterate β^t and residuals $r^t := y - X\beta^t$, take

$$\theta = \theta^t_{\text{res}} := r^t/\lambda$$

It is not necessarily feasible!

[6] 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[6]: at epoch t, corresponding to iterate β^t and residuals $r^t := y - X\beta^t$, take

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

residuals rescaling

[6] 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[6]: at epoch t, corresponding to iterate β^t and residuals $r^t := y - X\beta^t$, take

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

residuals rescaling

- converges to $\hat{\theta}$ (provided β^t converges to $\hat{\beta}$)
- $\mathcal{O}(np)$ to compute ($=$ 1 epoch of CD)
  $\rightarrow$ rule of thumb: compute θ^t_{res} and gap every $f = 10$ epochs

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

Table of Contents

Speeding up solvers

Lasso motivation: we expect sparse solutions/small supports

$$\mathcal{S}_{\hat{\beta}} := \left\{ j \in [p] \,:\, \hat{\beta}_j \neq 0 \right\}$$

"the solution restricted to its support solves the problem restricted to features in this support"

$$\hat{\beta}_{\mathcal{S}_{\hat{\beta}}} \in \underset{b \in \mathbb{R}^{\|\hat{\beta}\|_0}}{\arg\min} \frac{1}{2} \|y - X_{\mathcal{S}_{\hat{\beta}}} b\|^2 + \lambda \|b\|_1$$

Usually $\|\hat{\beta}\|_0 \ll p$ so the second problem is much simpler.

Technical detail

- The primal solution/support might not be unique!
- But $\hat{\theta}$ is unique and so is the *equicorrelation set*[7]:

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

- For any primal solution, $\mathcal{S}_{\hat{\beta}} \subset E$

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

Technical detail

- The primal solution/support might not be unique!
- But $\hat{\theta}$ is unique and so is the *equicorrelation set*[7]:

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

- For any primal solution, $\mathcal{S}_{\hat{\beta}} \subset E$

Grail of sparse solvers: identify E, solve on E

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

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

Speeding-up solvers

Two approaches:

- safe screening[8,9] (**backward approach**): remove feature j when it is certified that $j \notin E$
- working set[10] (**forward approach**): focus on j's very likely to be in E

Rem: hybrid approach: strong rules[11]

---

[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.

[11] R. Tibshirani et al. "Strong rules for discarding predictors in lasso-type problems". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.

Duality comes into play: gap screening

We want to identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

... but we can't get it without $\hat{\beta}$

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

$$\sup_{\theta \in \mathcal{C}} |X_j^\top \theta| < 1 \Rightarrow |X_j^\top \hat{\theta}| < 1$$

[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 comes into play: gap screening

We want to identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

... but we can't get it without $\hat{\beta}$

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

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

[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 comes into play: gap screening

We want to identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

... but we can't get it without $\hat{\beta}$

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 comes into play: gap screening

We want to identify $E = \left\{ j \in [p] \, : \, |X_j^\top \hat{\theta}| = 1 \right\}$

... but we can't get it without $\hat{\beta}$

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 $r = \sqrt{\frac{2}{\lambda^2} \text{gap}(\beta, \theta)}$

$$\boxed{\forall (\beta, \theta), |X_j^\top \theta| < 1 - \|X_j\| r \Rightarrow \hat{\beta}_j = 0}$$

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

Working/active set

Algorithm: Generic WS algorithm

Initialization: $\beta^0 = 0 \in \mathbb{R}^p$

for $it = 1, \dots, it_{\max}$ **do**

- define working set $\mathcal{W}_{it} \subset [p]$
- approximately solve Lasso restricted to features in $\mathcal{W}_{it}$
- update $\beta_{\mathcal{W}_{it}}$

Table of Contents

Back to dual choice

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

Two drawbacks of residuals rescaling:

- ignores information from previous iterates
- workload "imbalanced": more efforts in primal than in dual

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

Back to dual choice

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

Two drawbacks of residuals rescaling:

- ignores information from previous iterates
- workload "imbalanced": more efforts in primal than in dual

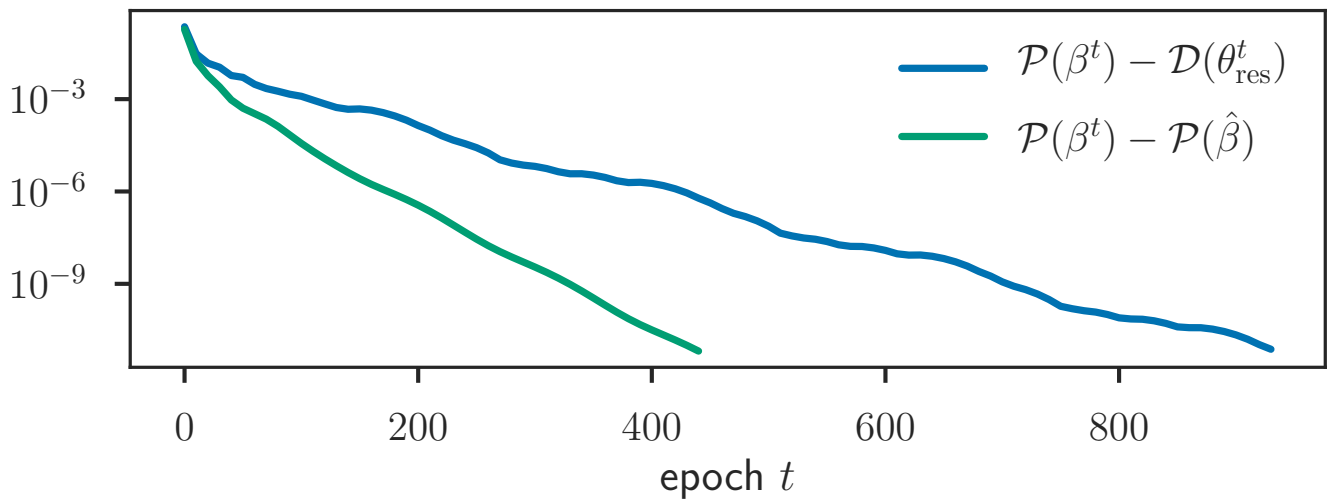
$\lambda_{\max} = \|X^\top y\|_\infty$ is the smallest λ giving $\hat{\beta} = 0$

Back to dual choice

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

Two drawbacks of residuals rescaling:

- ignores information from previous iterates
- workload "imbalanced": more efforts in primal than in dual



Leukemia dataset ($p = 7129, n = 72$), for $\lambda = \lambda_{\max}/20$

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

Acceleration through residuals extrapolation[13]

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

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

Acceleration through residuals extrapolation[13]

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

extrapolation!

$\rightarrow$ use the same idea to infer $\lim_{t \to \infty} r^t = \lambda \hat{\theta}$

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

Extrapolation justification

If $(x_t)_{t\in\mathbb{N}}$ follows a converging autoregressive process (AR):

$$x_t = ax_{t-1} - b \qquad (|a| < 1) \quad \text{with} \quad \lim_{t\to\infty} x_t = x^*$$

we have

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

Aitken's Δ^2: 2 unknowns, so 2 equations/3 points x_t, x_{t-1}, x_{t-2} are enough to find x^*![14]

[14] 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 to $x_t \in \mathbb{R}^n$

AMPE (Approximate Minimal Polynomial Extrapolation): applies to Vector autoregressive (VAR) process

i.e., $x_t \in \mathbb{R}^n$, and $a \in \mathbb{R}$ becomes $A \in \mathbb{R}^{n \times n}$

More difficult to eliminate A (unobserved), the idea is to approximate its minimal polynomial.

Extrapolated residuals[15]

- fix $K = 5$ (small)
- keep track of K past residuals $r^t, \ldots, r^{t+1-K}$

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated residuals[15]

- fix $K = 5$ (small)
- keep track of K past residuals $r^t, \dots, r^{t+1-K}$
- form $U^t = [r^{t+1-K} - r^{t-K}, \dots, r^t - r^{t-1}] \in \mathbb{R}^{n \times K}$

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated residuals[15]

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

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated residuals[15]

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

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated residuals[15]

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

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated residuals[15]

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

$$r^t_{\text{accel}} = \begin{cases} r^t, & \text{if } t \leq K \\ \sum_{k=1}^{K} c_k r^{t+1-k}, & \text{if } t > K \end{cases}$$

(affine combination, goes outside convex hull of $\{r^t, \ldots, r^{t+1-K}\}$)

[15] M. Massias, A. Gramfort, and J. Salmon. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].

Extrapolated dual point

$$r^t_{\text{accel}} = \begin{cases} r^t, & \text{if } t \leq K \\ \sum_{k=1}^{K} c_k r^{t+1-k}, & \text{if } t > K \end{cases}$$

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

Guarantees?

- convergence of θ^t_{accel}?
- $(U^t)^\top U^t z = \mathbf{1}_K \rightarrow$ linear system solving?

Guarantees?

- convergence of θ^t_{accel}?
- $(U^t)^\top U^t z = \mathbf{1}_K \rightarrow$ linear system solving?
- $c = \dfrac{z}{z^\top \mathbf{1}_K} \rightarrow$ what if $z^\top \mathbf{1}_K = 0$?

Guarantees?

- convergence of θ^t_{accel}?
- $(U^t)^\top U^t z = \mathbf{1}_K \rightarrow$ linear system solving?
- $c = \dfrac{z}{z^\top \mathbf{1}_K} \rightarrow$ what if $z^\top \mathbf{1}_K = 0$?

Guarantees?

- convergence of θ^t_{accel}?
- $(U^t)^\top U^t z = \mathbf{1}_K \to$ linear system solving?
- $c = \dfrac{z}{z^\top \mathbf{1}_K} \to$ what if $z^\top \mathbf{1}_K = 0$?

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

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

Guarantees?

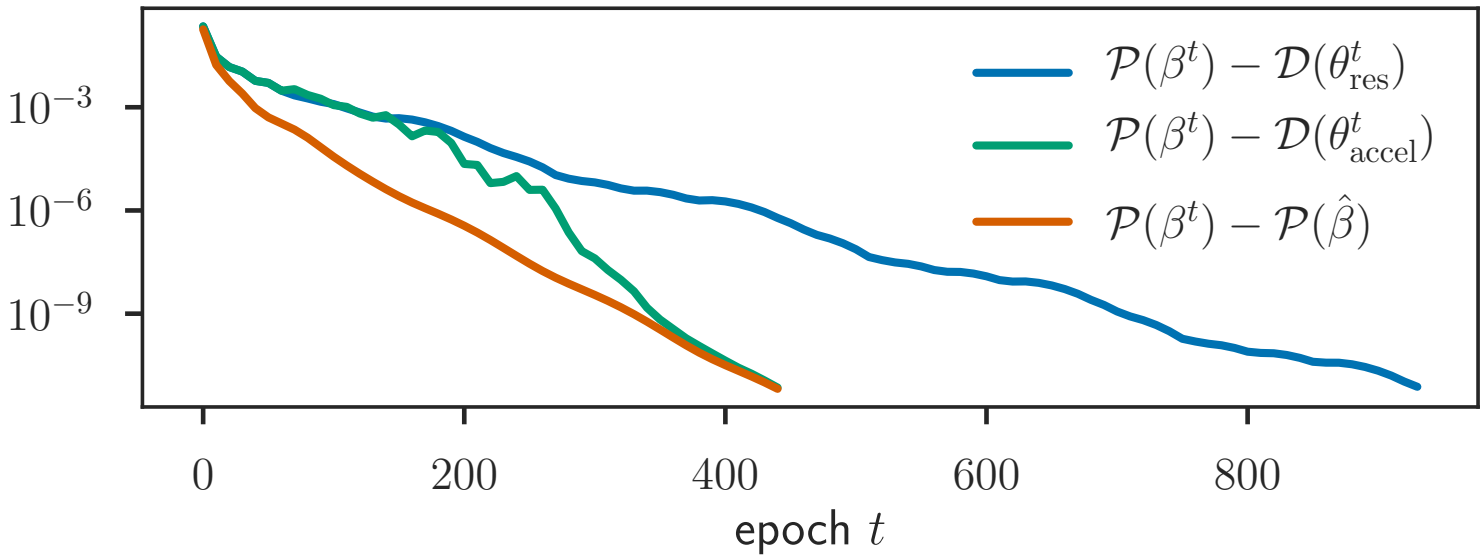
- convergence of θ^t_{accel}?
- $(U^t)^\top U^t z = \mathbf{1}_K \to$ linear system solving?
- $c = \dfrac{z}{z^\top \mathbf{1}_K} \to$ what if $z^\top \mathbf{1}_K = 0$?

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

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

Final cost of 10 CD epochs + gap computation $\approx$ 12 np vs 11 np in classical approach

Does it work?



Leukemia dataset ($p = 7129, n = 72$), for $\lambda = \lambda_{\max}/20$
(consistent finding across datasets)

- θ_{res} is bad
- θ_{accel} gives a tighter bound
- θ_{accel} does not behave erratically

Which algorithm to produce β^t?

Key assumption for extrapolation[16]: r^t follows a VAR.

- True for ISTA and the Lasso, once support is identified[17] (but ISTA/FISTA slow on our statistical scenarios)

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

[17] J. Liang, J. Fadili, and G. Peyré. "Local Linear Convergence of Forward–Backward under Partial Smoothness". In: *NIPS*. 2014, pp. 1970–1978.

Which algorithm to produce β^t?

Key assumption for extrapolation[16]: r^t follows a VAR.

- True for ISTA and the Lasso, once support is identified[17] (but ISTA/FISTA slow on our statistical scenarios)
- Conjecture: it is also true for cyclic CD

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

[17] J. Liang, J. Fadili, and G. Peyré. "Local Linear Convergence of Forward–Backward under Partial Smoothness". In: *NIPS*. 2014, pp. 1970–1978.

Which algorithm to produce β^t?

Key assumption for extrapolation[16]: r^t follows a VAR.

- True for ISTA and the Lasso, once support is identified[17] (but ISTA/FISTA slow on our statistical scenarios)
- Conjecture: it is also true for cyclic CD

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

[17] J. Liang, J. Fadili, and G. Peyré. "Local Linear Convergence of Forward–Backward under Partial Smoothness". In: *NIPS*. 2014, pp. 1970–1978.

Which algorithm to produce β^t?

Key assumption for extrapolation[16]: r^t follows a VAR.

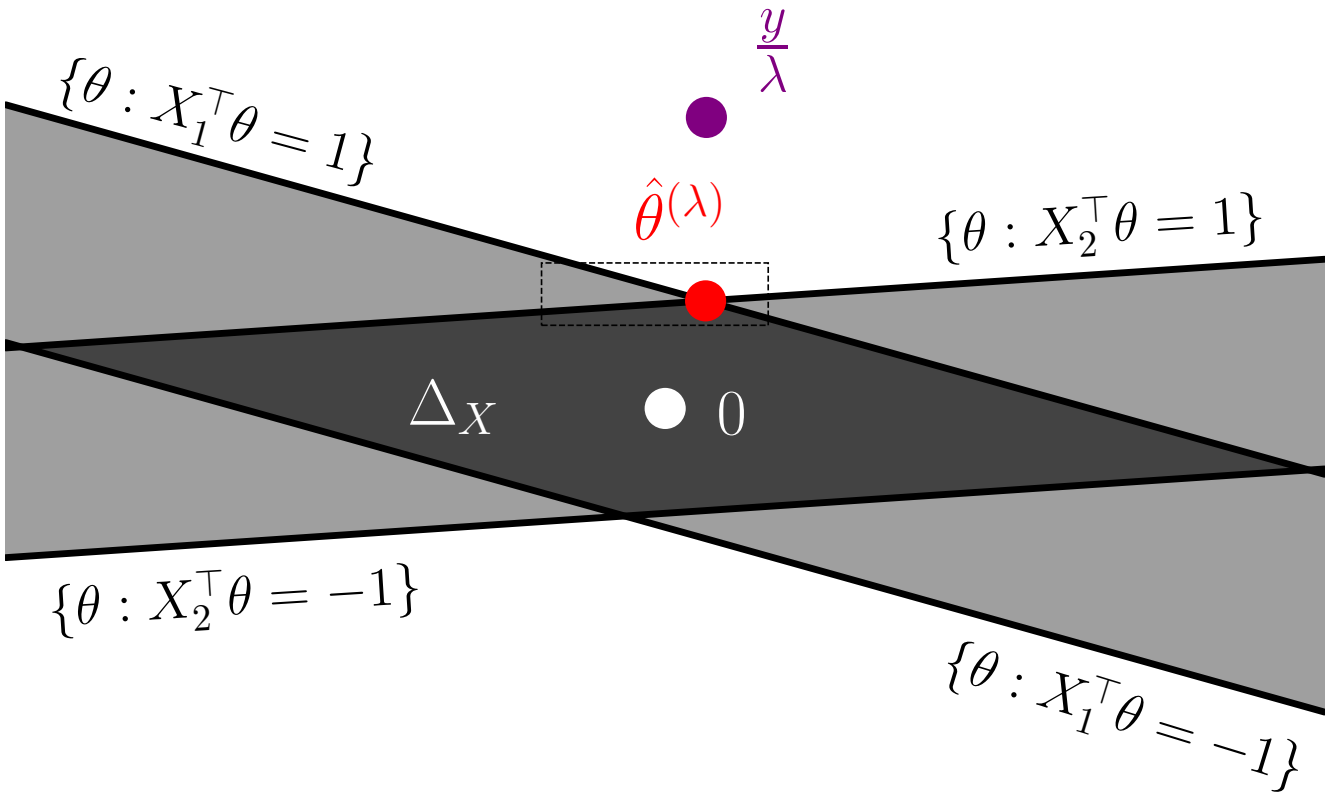
- True for ISTA and the Lasso, once support is identified[17] (but ISTA/FISTA slow on our statistical scenarios)
- Conjecture: it is also true for cyclic CD

Rem: : Shuffle CD breaks the regularity

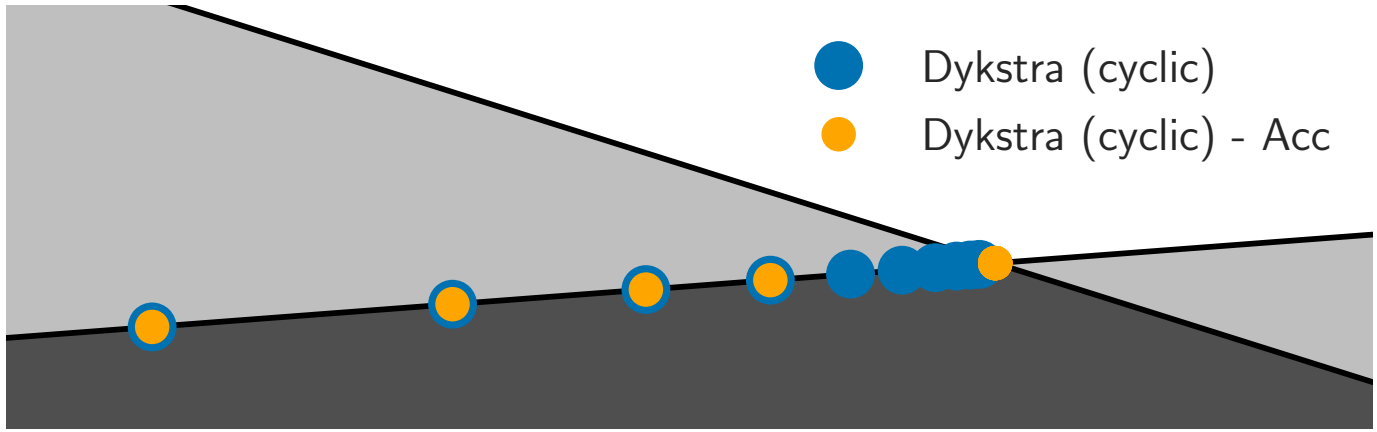
[16] D. Scieur, A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NIPS*. 2016, pp. 712–720.

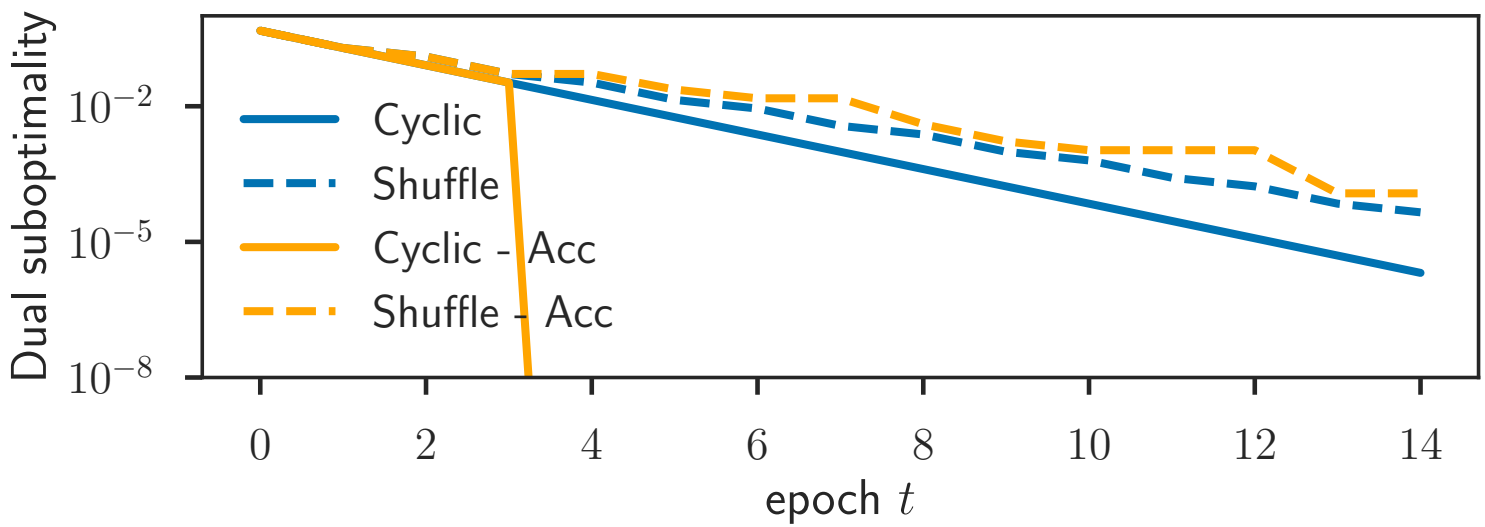
[17] J. Liang, J. Fadili, and G. Peyré. "Local Linear Convergence of Forward–Backward under Partial Smoothness". In: *NIPS*. 2014, pp. 1970–1978.

2D example again

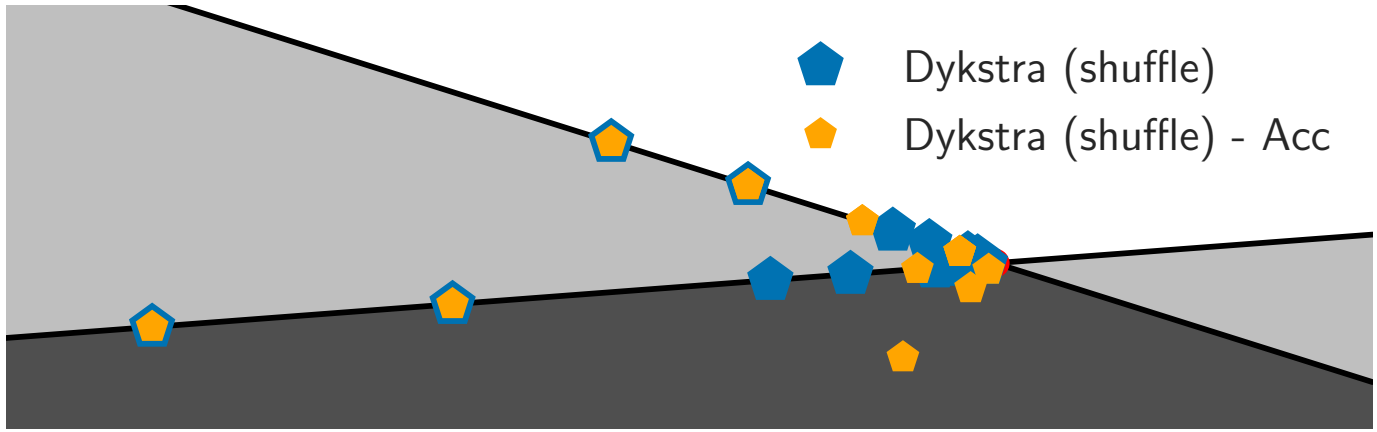


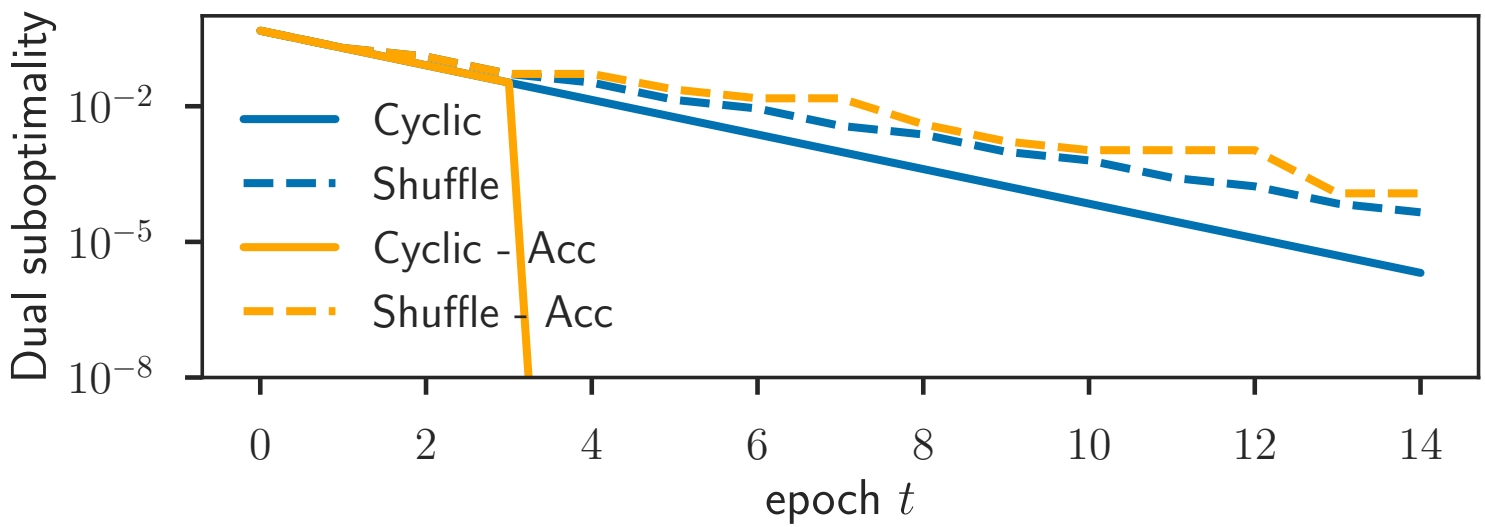
Toy dual zoom: cyclic





Toy dual zoom: shuffle





Better safe screening

Recall Gap Safe screening rule:

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

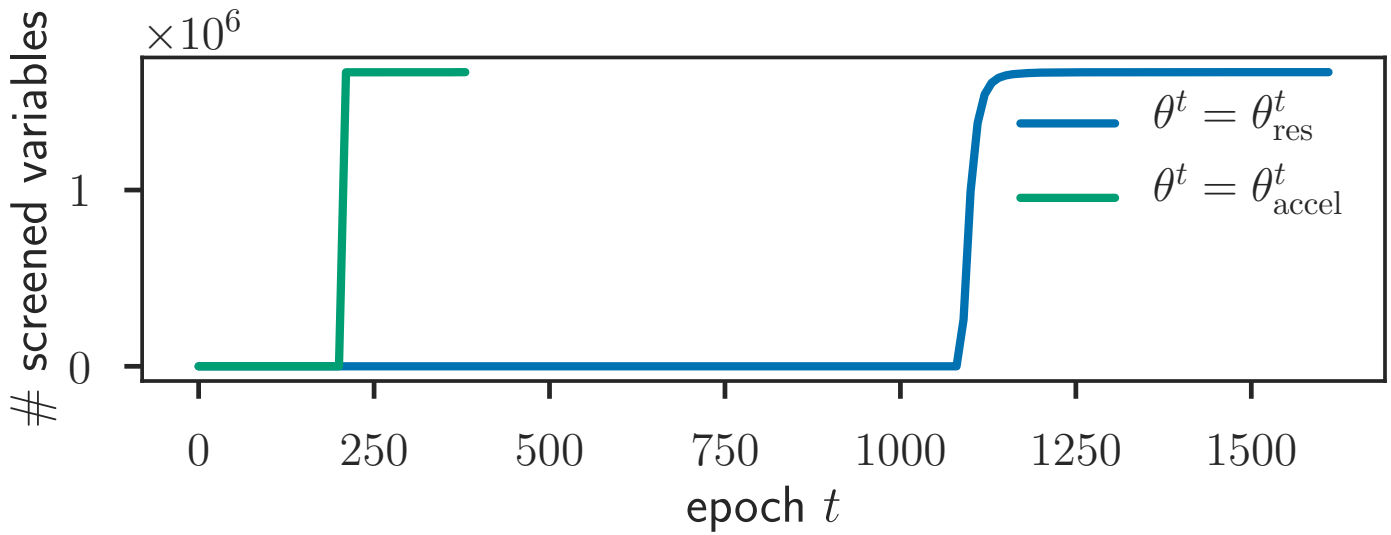
better dual point $\Rightarrow$ better safe screening

Better safe screening

Recall Gap Safe screening rule:

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

better dual point $\Rightarrow$ better safe screening



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

Screening vs Working sets

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

Screening vs Working sets

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

$$\Leftrightarrow$$

$$d_j(\theta) > \sqrt{\frac{2}{\lambda^2} \mathsf{gap}(\beta, \theta)} \Rightarrow \hat{\beta}_j = 0$$

$$\text{with } d_j(\theta) := \frac{1 - |X_j^\top \theta|}{\|X_j\|}$$

$d_j(\theta)$ larger than threshold $\rightarrow$ exclude feature j

Screening vs Working sets

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

$$\Leftrightarrow$$

$$d_j(\theta) > \sqrt{\frac{2}{\lambda^2} \mathsf{gap}(\beta, \theta)} \Rightarrow \hat{\beta}_j = 0$$

$$\text{with } d_j(\theta) := \frac{1 - |X_j^\top \theta|}{\|X_j\|}$$

$d_j(\theta)$ larger than threshold $\rightarrow$ exclude feature j

Alternative: Solve subproblem with small $d_j(\theta)$ only (WS)

3 questions for working sets

- how to prioritize features?

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$
- how many features in WS?

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$
- how many features in WS? $\rightarrow$ start at 100, double at each WS definition. Features cannot leave the WS

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$
- how many features in WS? $\rightarrow$ start at 100, double at each WS definition. Features cannot leave the WS
- solve the subproblem with which precision?

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$
- how many features in WS? $\rightarrow$ start at 100, double at each WS definition. Features cannot leave the WS
- solve the subproblem with which precision? $\rightarrow$ use same as required for whole problem

3 questions for working sets

- how to prioritize features? $\rightarrow$ use $d_j(\theta)$
- how many features in WS? $\rightarrow$ start at 100, double at each WS definition. Features cannot leave the WS
- solve the subproblem with which precision? $\rightarrow$ use same as required for whole problem

Not so fancy, but guarantees convergence.

Pruning variant: working set can decrease in size & features can leave the working set

Similarities[18,19]

$$d_j(\theta) := \frac{1 - |X_j^\top \theta|}{\|X_j\|}$$

[18] J. Fan and J. Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008), pp. 849–911.

[19] S. Stich, A. Raj, and M. Jaggi. "Safe Adaptive Importance Sampling". In: *NIPS*. 2017.

Similarities[18,19]

$$\boxed{d_j(\theta) := \frac{1 - |X_j^\top \theta|}{\|X_j\|}}$$

Lasso case with $\theta = \theta_{\text{res}}$ and normalized X_j's:

$$1 - d_j(\theta) \propto |X_j^\top r^t|$$

small $d_j(\theta)$ = high correlation with residuals/high norm of partial gradient of data-fitting term...

---

[18] J. Fan and J. Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008), pp. 849–911.

[19] S. Stich, A. Raj, and M. Jaggi. "Safe Adaptive Importance Sampling". In: *NIPS*. 2017.

Similarities[18,19]

$$\boxed{d_j(\theta) := \frac{1 - |X_j^\top \theta|}{\|X_j\|}}$$

Lasso case with $\theta = \theta_{\text{res}}$ and normalized X_j's:

$$1 - d_j(\theta) \propto |X_j^\top r^t|$$

small $d_j(\theta)$ = high correlation with residuals/high norm of partial gradient of data-fitting term...

BUT our strength is that we can use any θ, in particular θ_{accel}

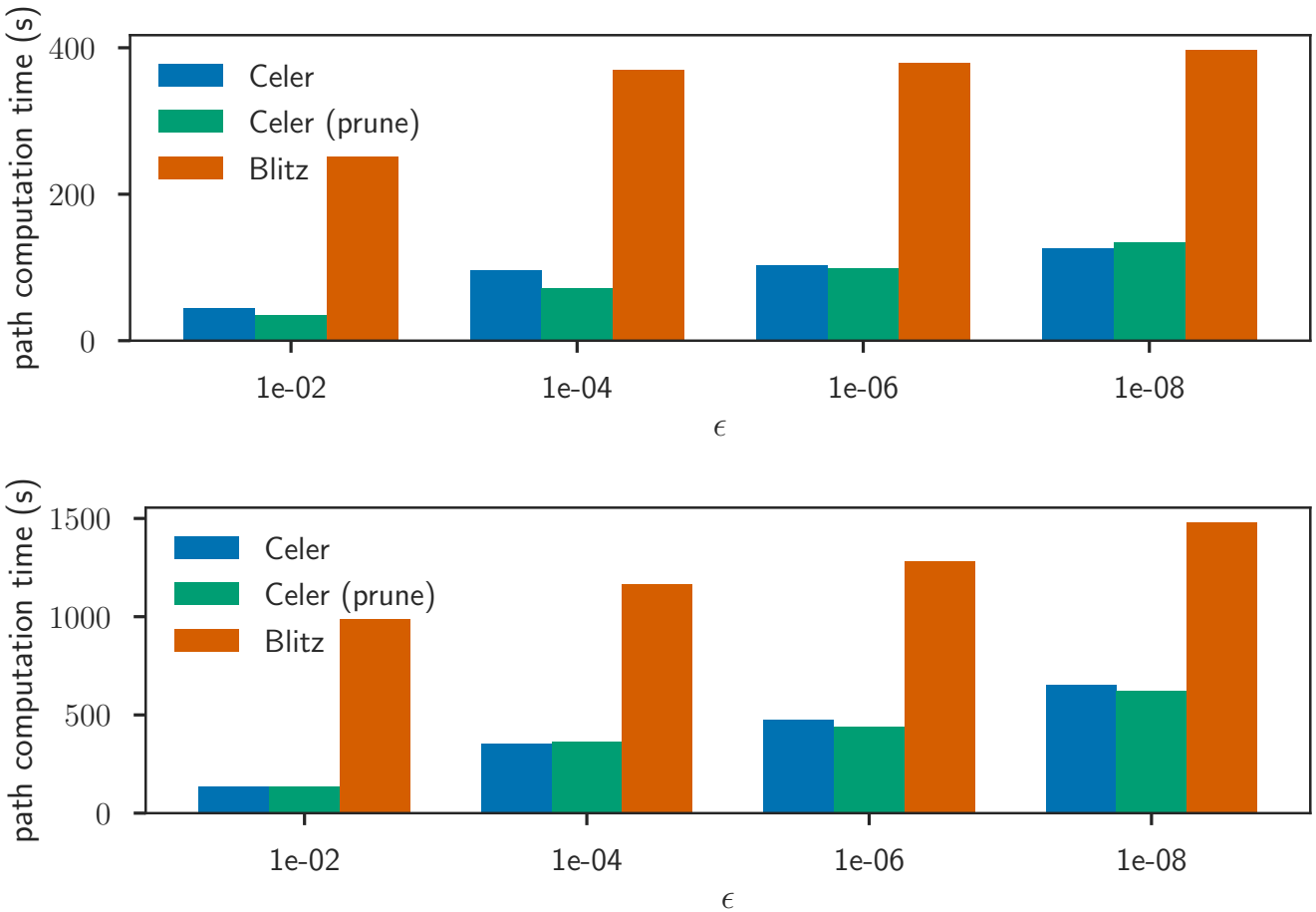
---

[18] J. Fan and J. Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008), pp. 849–911.

[19] S. Stich, A. Raj, and M. Jaggi. "Safe Adaptive Importance Sampling". In: *NIPS*. 2017.

Comparison

State-of-the-art WS solver for sparse problems: Blitz[20]

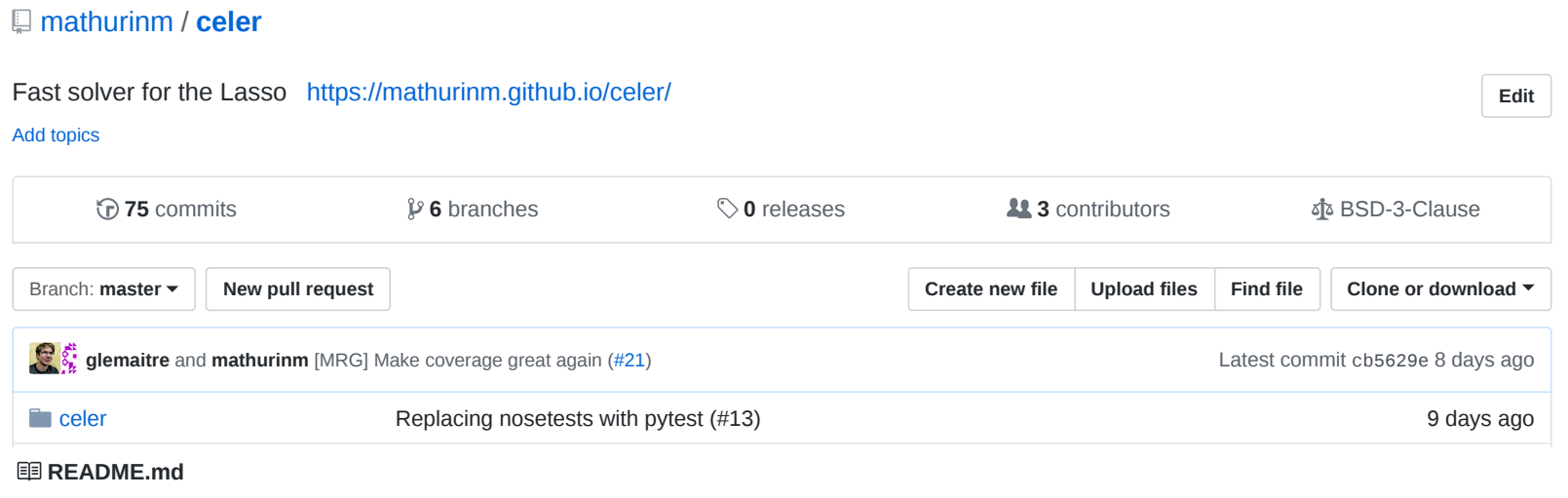


Finance dataset, Lasso path of 10 (top) or 100 (bottom) λ's from $\lambda_{\max}$ to $\lambda_{\max}/100$

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

Reusable science

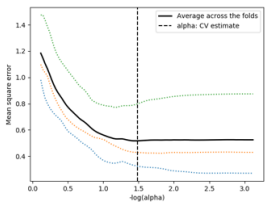
https://github.com/mathurinm/celer: code with continuous integration, code coverage, bug tracker



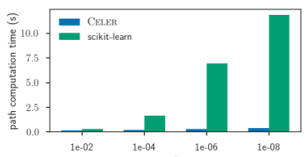
Examples gallery

https://mathurinm.github.io/celer: documentation (examples, API)

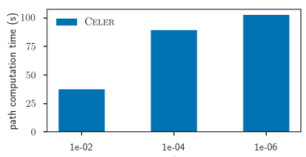
Examples Gallery¶



Run LassoCV for cross-validation on Leukemia



Lasso path computation on Leukemia dataset



Lasso path computation on Finance/log1p

Drop-in sklearn replacement

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

celer.Lasso

class `celer.` **Lasso** (***alpha=1.0, max_iter=100, gap_freq=10, max_epochs=50000, p0=10, verbose=0, tol=1e-06, prune=0, fit_intercept=True***)

Lasso scikit-learn estimator based on Celer solver

The optimization objective for Lasso is:

```
(1 / (2 * n_samples)) * ||y - X beta||^2_2 + alpha * ||beta||_1
```

Parameters:

alpha : float, optional

> Constant that multiplies the L1 term. Defaults to 1.0. `alpha = 0` is equivalent to an ordinary least square. For numerical reasons, using `alpha = 0` with the `Lasso` object is not advised.

max_iter : int, optional

> The maximum number of iterations (subproblem definitions)

gap_freq : int

> Number of coordinate descent epochs between each duality gap computations.

Drop-in sklearn replacement

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

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

celer.Lasso

class `celer.` **Lasso** (***alpha=1.0, max_iter=100, gap_freq=10, max_epochs=50000, p0=10, verbose=0, tol=1e-06, prune=0, fit_intercept=True***)

Lasso scikit-learn estimator based on Celer solver

The optimization objective for Lasso is:

```
(1 / (2 * n_samples)) * ||y - X beta||^2_2 + alpha * ||beta||_1
```

Parameters:

alpha : float, optional

Constant that multiplies the L1 term. Defaults to 1.0. `alpha = 0` is equivalent to an ordinary least square. For numerical reasons, using `alpha = 0` with the `Lasso` object is not advised.

max_iter : int, optional

The maximum number of iterations (subproblem definitions)

gap_freq : int

Number of coordinate descent epochs between each duality gap computations.

Conclusion

Duality matters at several levels for the Lasso:

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

Key improvement: residuals rescaling $\rightarrow$ residuals extrapolation

Future works:

- Can it work for sparse logreg, group Lasso?
- Can we prove convergence of θ_{accel} and give rates?

Feedback welcome on the online code!

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.
- Beck, A. and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM J. Imaging Sci.* 2.1 (2009), pp. 183–202.
- 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.
- Fan, J. and J. Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008), pp. 849–911.

References II

- Friedman, J. et al. "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332.
- Johnson, T. B. and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.
- Liang, J., J. Fadili, and G. Peyré. "Local Linear Convergence of Forward–Backward under Partial Smoothness". In: *NIPS*. 2014, pp. 1970–1978.
- 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. "Dual Extrapolation for Faster Lasso Solvers". In: *ArXiv e-prints* (2018). arXiv: 1802.07481 [stat.ML].
- Ndiaye, E. et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.

References III

- Scieur, D., A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NIPS*. 2016, pp. 712–720.
- Stich, S., A. Raj, and M. Jaggi. "Safe Adaptive Importance Sampling". In: *NIPS*. 2017.
- 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.
- Tibshirani, R. et al. "Strong rules for discarding predictors in lasso-type problems". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.
- Tseng, P. "Convergence of a block coordinate descent method for nondifferentiable minimization". In: *J. Optim. Theory Appl.* 109.3 (2001), pp. 475–494.