Beyond L1: Faster and better sparse models with `skglm`

Pierre-Antoine Bannier

with Q. Bertrand, Q. Klopfenstein, G. Gidel and M. Massias

https://arxiv.org/abs/2204.07826

The optimization problem at hand

$$\boxed{\min f(\beta) + \sum_{j=1}^{p} g_j(\beta_j) = f(\beta) + g(\beta)}$$

- f convex + classical assumptions
- g_j classical assumptions + not necessarily convex

Focus on finding a critical point[1]:

$$-\nabla f(\beta) \in \partial g(\beta)$$

Def.: *generalized support* of $\beta \in \mathbb{R}^p$ = set of indices $j \in [p]$ s.t. g_j is differentiable at β_j:

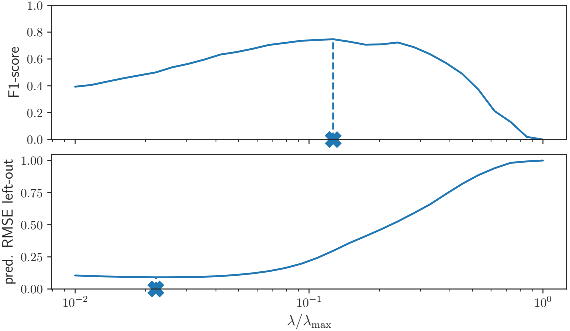
$$\mathrm{gsupp}(\beta) = \{j \in [p] : \partial g_j(\beta_j) \text{ is a singleton}\}$$

Ex: non-zero coefficients for ℓ_1, support vectors for SVM

[1] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.

Limitations of convex penalties

Amplitude bias leads to *estimation-prediction* dilemma:

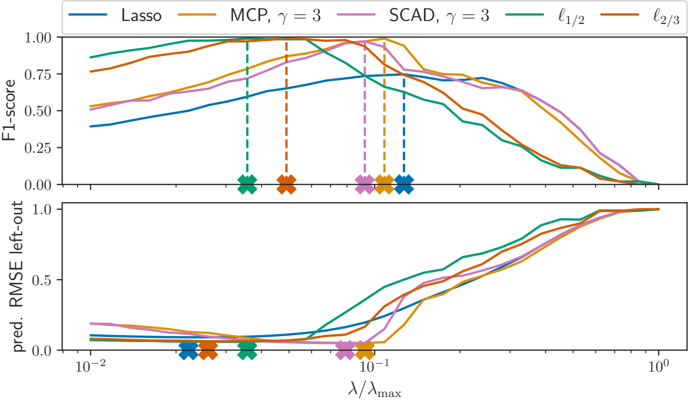


Top: support recovery, bottom: left-out prediction

$\hookrightarrow$ **go non convex**[2]

[2] E. Soubies, L. Blanc-Féraud, and G. Aubert. "A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization". In: *SIAM Journal on Optimization* 27.3 (2017), pp. 2034–2060.

Performance of non-convex penalties



- solve estimation-prediction dilemma[3]
- achieve perfect support recovery

$\hookrightarrow$ need fast algorithms, not tailored to L1 or quadratics

[3] C.-H. Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of statistics* 38.2 (2010), pp. 894–942.

The limitations of current algorithms

Most popular packages for sparse generalized linear models

Name	Acceleration	Huge scale	Nncvx	Modular
`glmnet` [Friedman et al., 2010]	✗	✗	✗	✗ (Fortran)
`scikit-learn` [Pedregosa et al., 2011]	✗	✗	✗	✗ (Cython)
`lightning` [Blondel and Pedregosa, 2016]	✗	✗	✗	✓ (Cython)
`celer` [Massias et al., 2018]	✓	✓	✗	✗ (Cython)
`picasso` [Ge et al., 2019]	✗	✗	✓	✗ (`C++`)
`pyGLMnet` [Jas et al., 2020]	✗	✗✗	✗	✓ (Python)

- Fast algorithms have a limited number of penalties
- Code is not easily maintainable (legacy code in Fortran and C)

$\hookrightarrow$ need for a **modular**, **generic** and **fast** solver for sparse GLMs

The ingredients of `skglm`

1. Working set strategy,[4] able to handle a large class of convex and non-convex penalties (outer)

2. Anderson accelerated[5] coordinate descent for non-convex problems (inner)

[4] T. B. Johnson and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. vol. 37. 2015, pp. 1171–1179.

[5] D. G. Anderson. "Iterative procedures for nonlinear integral equations". In: *Journal of the ACM* 12.4 (1965), pp. 547–560.

Ingredient #1: Working sets

Working set: identify the generalized support of the solution, ignore other coefficients

`skglm` ranks features with a violation of the optimality condition:

$$\mathrm{score}^{\partial}_j = \mathrm{dist}(-\nabla_j f(\hat{\beta}), \partial g_j(\hat{\beta})) \ .$$

Take largest n_k features violating condition, solve restricted problem, increase n_k.

Prop.: If inner solver converges to a critical point, the whole algorithm converges .[6]

---

[6] Q. Bertand et al. *Beyond L1: Faster and Better Sparse Models with skglm*. 2022.

Ingredient #2: Anderson acceleration

Need linear iterations, vector autoregressive structure:

$$\beta^{(k+1)} = A\beta^{(k)} + b, \qquad \text{with } \lambda_{\max}(A) < 1 \ .$$

For coordinate descent (matrix not symmetrical[7]):

$$\beta^{(k+1)} = \underbrace{\left(\mathrm{Id}_p - \frac{e_n e_n^\top}{X_{nn}} X\right) \dots \left(\mathrm{Id}_p - \frac{e_1 e_1^\top}{X_{11}} X\right)}_{T^{\mathrm{CD}}} \beta^{(k)} + b^{\mathrm{CD}} \ .$$

(7) M. Massias et al. "Dual extrapolation for sparse generalized linear models". In: *J. Mach. Learn. Res.* (2020).

Ingredient #2: Anderson acceleration

Anderson extrapolation coefficient found by solving:

$$\min_{c^\top \mathbf{1}_K = 1} \left\| \sum_{k=1}^{K} c_k (\beta^{(k)} - \beta^{(k-1)}) \right\| \quad .$$

Introducing $U = (\beta^{(1)} - \beta^{(0)}, \dots, \beta^{(K)} - x^{(K-1)}) \in \mathbb{R}^{d \times K}$:

$$\min_{c^\top \mathbf{1}_K = 1} \|Uc\|^2 \quad .$$

Closed-form (cost: $K^3 + K^2 d$, but K small in practice):

$$\boxed{c = \frac{(U^\top U)^{-1} \mathbf{1}_K}{\mathbf{1}_K^\top (U^\top U)^{-1} \mathbf{1}_K}}$$

Theoretical guarantees of support identification

Assumptions:

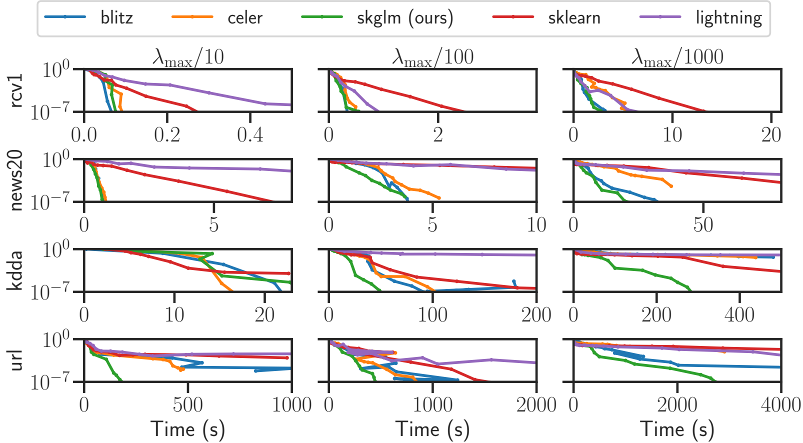
- α-semi convex penalties g_j/L_j (common for most non-convex penalties)
- convergence to a non-degenerated critical point $\hat{\beta} \in \mathbb{R}^p$: $\forall j \notin \operatorname{gsupp}(\hat{\beta})$, $-\nabla f_j(\hat{\beta}) \in \operatorname{int}(\partial g_j(\hat{\beta}_j))$.
- local ($\mathcal{C}^3$) regularity assumptions ($+$ piecewise quadratic g_j)

Guarantees:

- Identification of the generalized support with working set makes the problem easier
- Very fast local accelerated convergence rate with Anderson acceleration

Experiments: Lasso

State-of-the-art on convex problems (n, p in the millions)



Conclusion and future work

- Flexible solver for large scale non convex sparse models
- sklearn-compliant package released:
  https://github.com/scikit-learn-contrib/skglm
- Model identification, Anderson acceleration
- Paper: https://arxiv.org/abs/2204.07826

Bibliography I

- Anderson, D. G. "Iterative procedures for nonlinear integral equations". In: *Journal of the ACM* 12.4 (1965), pp. 547–560.
- Bauschke, H. H. and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.
- Bertand, Q. et al. *Beyond L1: Faster and Better Sparse Models with skglm*. 2022.
- Johnson, T. B. and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. Vol. 37. 2015, pp. 1171–1179.
- Massias, M. et al. "Dual extrapolation for sparse generalized linear models". In: *J. Mach. Learn. Res.* (2020).
- Soubies, E., L. Blanc-Féraud, and G. Aubert. "A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization". In: *SIAM Journal on Optimization* 27.3 (2017), pp. 2034–2060.

Bibliography II

- Zhang, C.-H. "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of statistics* 38.2 (2010), pp. 894–942.