`skglm`: an algorithm for non-smooth convex and non-convex optimization

Pierre-Antoine Bannier (X-HEC, M2 student)

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

https://arxiv.org/abs/2204.07826

Accepted yesterday at NeurIPS 2022

Sparse regression: a wide variety of applications

- Modern applications: $\#$ samples $\ll$ $\#$ features
- Solution: assume parameters are sparse
- Extensively studied theoretical properties[1]
- Available implementations of fast algorithms for $\#$ features up to 10^6: `sklearn`, `glmnet`,[2] `liblinear`[3]

---

[1] T. J. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[2] J. Friedman, T. Hastie, and R. Tibshirani. "glmnet: Lasso and elastic-net regularized generalized linear models". In: *R package version* 1.4 (2009).

[3] R. E. Fan et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.

What is sparse regression?

- supervised learning framework:

$$\text{i.i.d. dataset } (x_i, y_i)_{i \in [n]} \in \mathbb{R}^p \times \mathcal{Y}$$

- generalized linear models: parameters of the distribution depend linearly on x_i:

$$y_i | x_i \sim \phi(x_i^\top \beta)$$

- separable sparsity-inducing penalty $g = \sum_j g_j$ modulated by a regularization parameter $\lambda \in \mathbb{R}_+^*$

- inference through negative log-likelihood minimization:

$$\boxed{\beta^\star \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} f_i(x_i^\top \beta) + \sum_{i=1}^{p} g_j(\beta_j)}$$

Some well-known sparse GLMs

Lasso[4]:

$$\beta^* \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Sparse logistic regression:

$$\beta^* \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log(1 + \exp(-x_i^\top \beta y_i)) + \lambda \|\beta\|_1$$

SVM with hinge loss:

$$\alpha^* \in \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top G \alpha - \sum_{i=1}^{n} \alpha_i + \iota_{[0,C]}(\alpha_i)$$

$$\text{with } G_{ij} = y_i y_j x_i^\top x_j$$

---

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

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[5]:

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

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

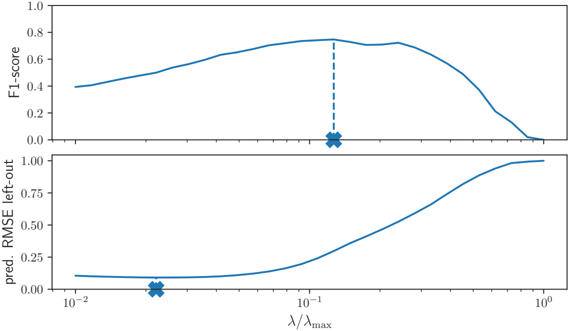
$$\text{gsupp}(\beta) = \{j \in [p] : \partial g_j(\beta_j) \text{ is a singleton}\}$$

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

[5] 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:

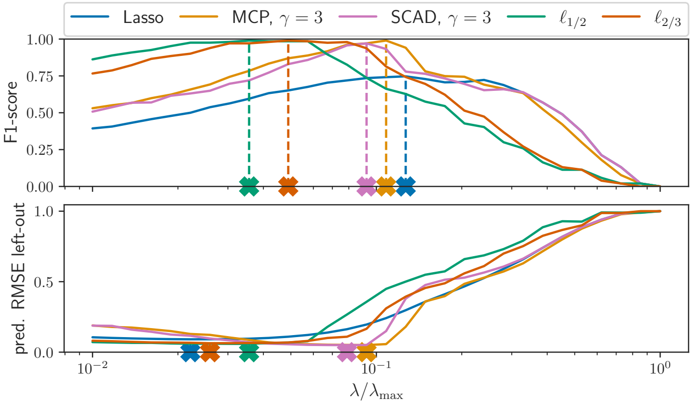


Top: support recovery, bottom: left-out prediction

$\hookrightarrow$ **go non convex**[6]

[6] 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[7]
- achieve perfect support recovery

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

[7] 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,[8] able to handle a large class of convex and non-convex penalties (outer)

2. Anderson accelerated[9] coordinate descent for non-convex problems (inner)

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

[9] 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:

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

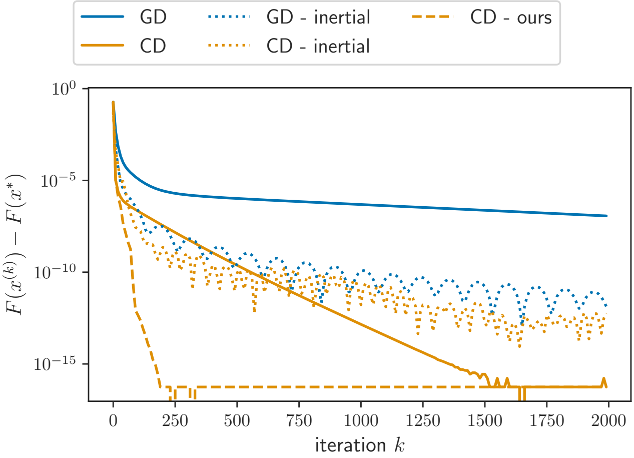
Take largest n_k features violating condition, solve restricted problem, increase n_k.

Prop.: If inner solver converges to a critical point, the whole algorithm converges .[10]

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

Ingredient #2: Anderson acceleration

Classical acceleration à la Nesterov[11] of CD is tricky:



Anderson CD[12] results in practical gains

[11] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

[12] Q. Bertrand and M. Massias. "Anderson acceleration of coordinate descent". In: *AISTATS*. 2021.

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[13]):

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

[13] 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:

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

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

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

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 $\hat{\beta} \in \mathbb{R}^p$: $\forall j \notin \operatorname{gsupp}(\hat{\beta})$, $-\nabla f_j(\hat{\beta}) \in \operatorname{int}(\partial g_j(\hat{\beta}_j))$.

Proposition (Model identification)

*CD **identifies the model** in finitely many iterations: for $\mathcal{S} = \operatorname{gsupp}(\hat{\beta})$, there exists $K > 0$ such that for all $k \geq K$,*

$$\beta^{(k)}_{\mathcal{S}^c} = \hat{\beta}_{\mathcal{S}^c}$$

Identification of the generalized support makes the problem easier

Improved convergence rate

Under local ($\mathcal{C}^3$) regularity assumptions ($+$ piecewise quadratic g_j)

$\hookrightarrow$ **Very fast local** convergence rates

Proposition (Accelerated rates)

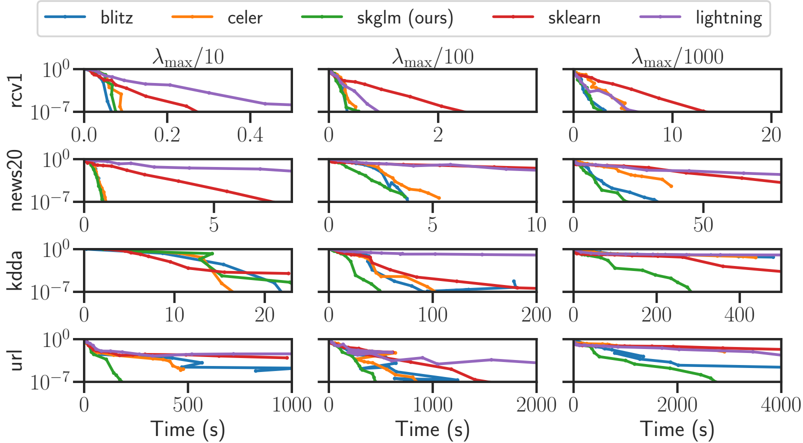
There exists $K \in \mathbb{N}$, *and a* $\mathcal{C}^1$ *function* $\psi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ *such that, for all* $k \in \mathbb{N}, k \geq K$*:* $\beta_j^{(k)} = \hat{\beta}_j$ *for all* $j \in \mathcal{S}^c$

Let $T \triangleq \mathcal{J}\psi(\hat{\beta})$, $H \triangleq \nabla^2_{\mathcal{S},\mathcal{S}} f(\hat{\beta}) + \nabla^2_{\mathcal{S},\mathcal{S}} g(\hat{\beta})$,
$\zeta \triangleq (1 - \sqrt{1 - \rho(T)})/(1 + \sqrt{1 - \rho(T)})$ *and*
$B \triangleq (T - \mathrm{Id})^\top (T - \mathrm{Id})$.
Then $\rho(T) < 1$ *and the iterates of Anderson extrapolation enjoy local accelerated convergence rate:*

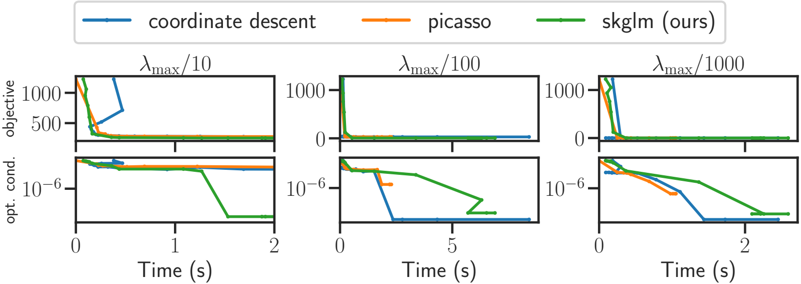
$$\|\beta_{\mathcal{S}}^{(k-K)} - \hat{\beta}_{\mathcal{S}}\|_B \leq \left(\sqrt{\kappa(H)}\frac{2\zeta^{M-1}}{1+\zeta^{2(M-1)}}\right)^{(k-K)/M} \|\beta_{\mathcal{S}}^{(K)} - \hat{\beta}_{\mathcal{S}}\|_B$$

Experiments: Lasso

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



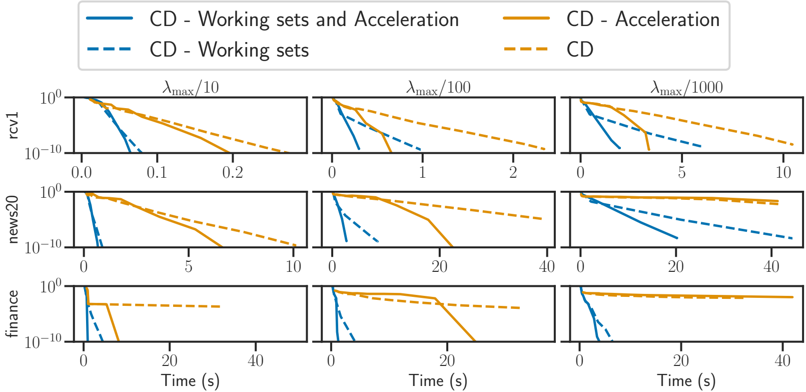
Experiments: MCP



Code integrated in scikit-learn:

https://github.com/scikit-learn-contrib/skglm

Ablation study



Both Anderson acceleration and working sets are useful

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

Appendix #1: Another ranking strategy for working sets

For non-convex ℓ_p $(p < 1)$ penalties, $\partial g_j(\hat{\beta}) = \mathbb{R}$

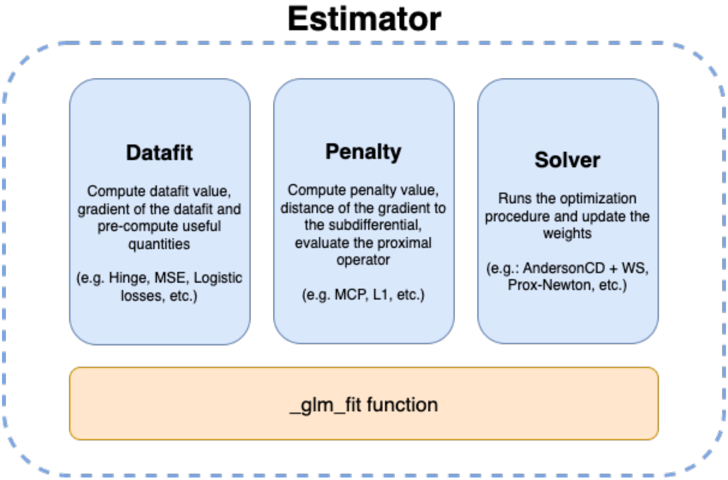
$\hookrightarrow$ need another ranking strategy

Alternative strategy based on violation of the fixed point iterate:

$$\mathrm{score}_j = |\hat{\beta}_j - \mathbf{prox}\left(\hat{\beta}_j - \frac{1}{L_j}\nabla_j f(\hat{\beta})\right)| \enspace .$$

Appendix #2: how is `skglm` designed?

skglm is written entirely in Python (JIT-compiled with Numba)



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.
- Bertrand, Q. and M. Massias. "Anderson acceleration of coordinate descent". In: *AISTATS*. 2021.
- Fan, R. E. et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.
- Friedman, J., T. Hastie, and R. Tibshirani. "glmnet: Lasso and elastic-net regularized generalized linear models". In: *R package version* 1.4 (2009).
- Hastie, T. J., R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

Bibliography II

- 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).
- Nesterov, Y. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.
- 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.
- Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

Bibliography III

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