Anderson acceleration of coordinate descent

- Quentin Bertrand (MILA, Montréal, Canada)
- **Mathurin Massias** (Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France)

Keywords: optimization, acceleration, Anderson, coordinate descent

Abstract: Gradient descent is the workhorse of modern convex optimization. For composite problems, proximal gradient descent retains the nice properties enjoyed by the latter. In both techniques, inertial acceleration achieves accelerated convergence rates [1]. Coordinate descent is a variant of gradient descent, which updates the iterates one coordinate at a time [2]. Proximal coordinate descent has been applied to numerous Machine Learning problems, in particular the Lasso, elastic net or sparse logistic regression. It is used in preeminent packages such as scikit-learn, glmnet, libsvm or lightning. On the theoretical side, inertial accelerated versions of coordinate descent [3, 4] achieve accelerated rates.

To obtain accelerated rates, Anderson extrapolation [5] is an alternative to inertia: it provides acceleration by exploiting the iterates' structure. Anderson acceleration enjoys accelerated rates on quadratics, but theoretical guarantees in the nonquadratic case are weaker [6]. Interestingly, numerical performances still show significant improvements on nonquadratic objectives. Anderson acceleration has been adapted to various algorithms such as Douglas-Rachford, ADMM or proximal gradient descent [7]. Among main benefits, the practical version of Anderson acceleration is memory efficient, easy to implement, line search free, has a low cost per iteration and does not require knowledge of the strong convexity constant. Finally, it introduces a single additional parameter, which often does not require tuning.

In this work:

- We propose an Anderson acceleration scheme for coordinate descent, which outperforms inertial and extrapolated gradient descent, as well as inertial and randomized coordinate descent.
- The acceleration is obtained even though the iteration matrix is not symmetric, a notable problem in the analysis of Anderson extrapolation.
- We empirically highlight that the proposed acceleration technique can generalize in the nonquadratic case and significantly improve proximal coordinate descent algorithms, which are state-of-the-art first order methods on the considered problems.

A longer version of this work has been accepted at the AISTATS21 conference [8].

References:

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[2] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

[3] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. In *NeurIPS*, pages 3059–3067. 2014.

[4] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[5] D. G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4): 547–560, 1965.

[6] D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

[7] V. V. Mai and M. Johansson. Anderson acceleration of proximal gradient methods. In *ICML*. 2019.

[8] Q. Bertrand and M. Massias. Anderson acceleration of coordinate descent. In *AISTATS*, 2021.