

Anderson acceleration of coordinate descent

Quentin Bertrand (Inria)
<https://qb3.github.io>

Mathurin Massias (University of Genova)
<https://mathurinm.github.io/>

Why (proximal) coordinate descent?

State-of-the art solvers^{1,2} for optimization-based estimators:

$$\arg \min_{x \in \mathbb{R}^p} \underbrace{f(Ax)}_{\text{smooth}} + \underbrace{\sum_{j=1}^p g_j(x_j)}_{\text{separable}}$$

Examples:

- ▶ Lasso $\arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$
- ▶ Elastic net $\arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1 + \frac{\rho}{2} \|x\|_2^2$
- ▶ (dual) SVM

¹F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *JMLR* 12 (2011), pp. 2825–2830.

²J. Friedman et al. "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332.

Why (proximal) coordinate descent?

State-of-the art solvers^{1,2} for optimization-based estimators:

$$\arg \min_{x \in \mathbb{R}^p} \underbrace{f(Ax)}_{\text{smooth}} + \underbrace{\sum_{j=1}^p g_j(x_j)}_{\text{separable}}$$

Examples:

- ▶ Lasso $\arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$
- ▶ Elastic net $\arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1 + \frac{\rho}{2} \|x\|_2^2$
- ▶ (dual) SVM

¹F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *JMLR* 12 (2011), pp. 2825–2830.

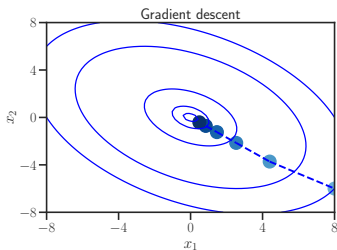
²J. Friedman et al. "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332.

CD on least squares

$$\arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|^2, A \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n$$

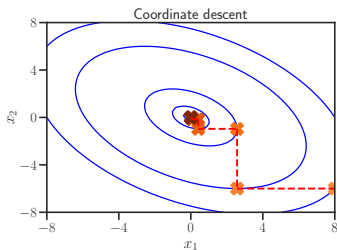
Algorithm: Gradient descent

init : $x \in \mathbb{R}^p$
for $k = 0, 1, \dots$, **do**
 $x \leftarrow x - \frac{A^\top(Ax - y)}{\|A\|_2^2}$
return x



Algorithm: CD

init : $x \in \mathbb{R}^p$
for $k = 0, 1, \dots$, **do**
 Select $j \in [p]$
 $x_j \leftarrow x_j - \frac{A_{:j}^\top(Ax - y)}{\|A_{:j}\|_2^2}$
return x



Why CD works well?

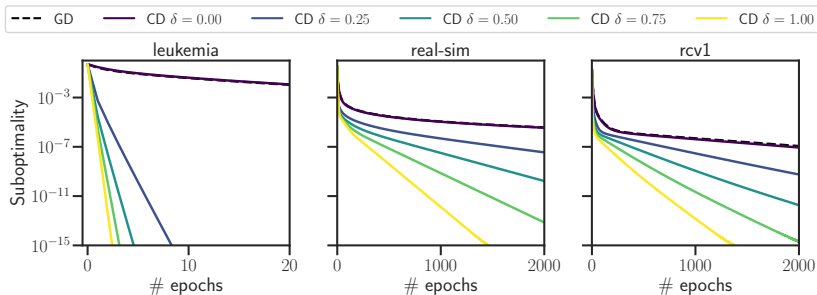
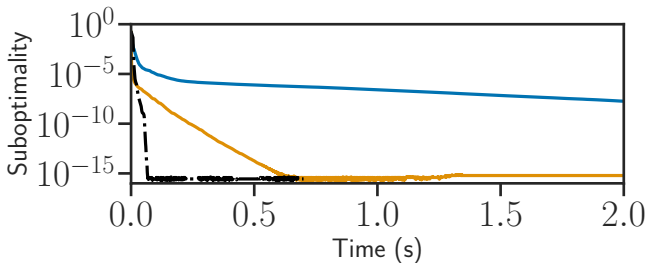
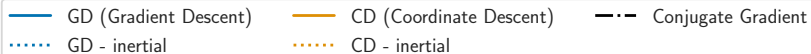


Figure: Influence of the step size for coordinate descent, OLS.

Gradient descent is compared against coordinate descent with step sizes $\gamma_j = \delta/L_j + (1 - \delta)/L$, for multiple values of δ .

Large step sizes: better convergence

Acceleration of CD



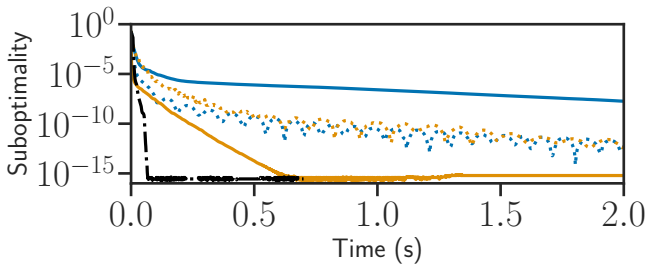
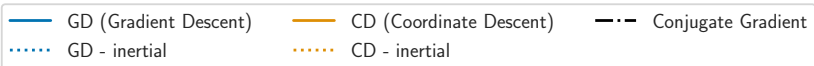
Least squares on *rcv1* ($n = p \approx 20k$)

Nesterov-like inertial CD^{3,4} slows down convergence

³Q. Lin, Z. Lu, and L. Xiao. "An Accelerated Proximal Coordinate Gradient Method". In: *NeurIPS*. 2014, pp. 3059–3067.

⁴O. Fercoq and P. Richtárik. "Accelerated, parallel, and proximal coordinate descent". In: *SIAM Journal on Optimization* 25.4 (2015), pp. 1997–2023.

Acceleration of CD



Least squares on *rcv1* ($n = p \approx 20k$)

Nesterov-like **inertial CD**^{3,4} **slows down** convergence

³Q. Lin, Z. Lu, and L. Xiao. "An Accelerated Proximal Coordinate Gradient Method". In: *NeurIPS*. 2014, pp. 3059–3067.

⁴O. Fercoq and P. Richtárik. "Accelerated, parallel, and proximal coordinate descent". In: *SIAM Journal on Optimization* 25.4 (2015), pp. 1997–2023.

Anderson acceleration: intuition

How to accelerate fixed point algorithms

$$x^{(k+1)} = Tx^{(k)} + b ?$$

Idea: search a fixed point of the form

$$x^* = \sum_{i=1}^k c_i x^{(k-1)}$$

Anderson acceleration: intuition

How to accelerate fixed point algorithms

$$x^{(k+1)} = Tx^{(k)} + b ?$$

Idea: search a fixed point of the form

$$x^* = \sum_{i=1}^k c_i x^{(k-1)}$$

One should have:

$$\sum_{i=1}^k c_i x^{(k-1)} \approx T \sum_{i=0}^{k-1} c_i x^{(k-1)} + b$$

Anderson acceleration: intuition

How to accelerate fixed point algorithms

$$x^{(k+1)} = Tx^{(k)} + b ?$$

Idea: search a fixed point of the form

$$x^* = \sum_{i=1}^k c_i x^{(k-1)}$$

One should have:

$$\sum_{i=1}^k c_i x^{(k-1)} \approx T \sum_{i=0}^{k-1} c_i x^{(k-1)} + b$$

Choose c_i such that

$$c \in \arg \min_{\sum_i c_i = 1} \left\| \sum_{i=1}^k c_i x^{(k-1)} - T \sum_{i=1}^k c_i x^{(k-1)} - b \right\|^2$$

$$\in \arg \min_{\sum_i c_i = 1} \left\| \sum_{i=1}^k c_i x^{(k-1)} - \sum_{i=1}^k c_i x^{(k)} \right\|^2 = \left\| \sum_{i=1}^k c_i (x^{(k-1)} - x^{(k)}) \right\|^2$$

Anderson acceleration: intuition

How to accelerate fixed point algorithms

$$x^{(k+1)} = Tx^{(k)} + b ?$$

Idea: search a fixed point of the form

$$x^* = \sum_{i=1}^k c_i x^{(k-1)}$$

One should have:

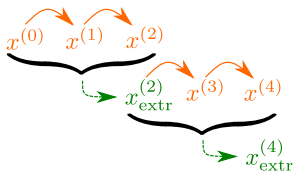
$$\sum_{i=1}^k c_i x^{(k-1)} \approx T \sum_{i=0}^{k-1} c_i x^{(k-1)} + b$$

Choose c_i such that

$$c \in \arg \min_{\sum_i c_i = 1} \left\| \sum_{i=1}^k c_i x^{(k-1)} - T \sum_{i=1}^k c_i x^{(k-1)} - b \right\|^2$$

$$\in \arg \min_{\sum_i c_i = 1} \left\| \sum_{i=1}^k c_i x^{(k-1)} - \sum_{i=1}^k c_i x^{(k)} \right\|^2 = \left\| \sum_{i=1}^k c_i (x^{(k-1)} - x^{(k)}) \right\|^2$$

Anderson acceleration: algorithm^{5,6}

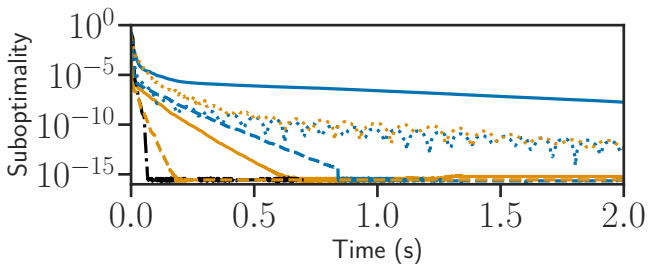


```
init   :  $x^{(0)} \in \mathbb{R}^P$ 
for  $k = 1, \dots$  do
     $x^{(k)} = Tx^{(k-1)} + b$  // regular iter.
    if  $k = 0 \pmod K$  then
         $U = [x^{(k-K+1)} - x^{(k-K)}, \dots, ]$ 
         $c = (U^\top U)^{-1} \mathbf{1}_K$ 
         $x_{\text{extr}}^{(k)} = \sum_i^K c_i x^{(k-K+i)} / \sum_i c_i$ 
         $x^{(k)} = x_{\text{extr}}^{(k)}$  // base sequence changes
return  $x^{(k)}$ 
```

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

⁶D. Scieur. "Generalized Framework for Nonlinear Acceleration". In: *arXiv preprint arXiv:1903.08764* (2019).

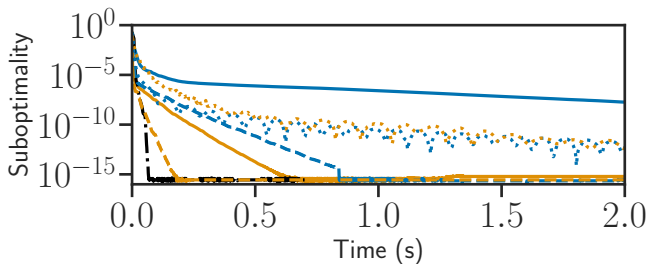
Acceleration of CD II



Least squares on *rcv1* ($n = p \approx 20k$)

► Anderson acceleration provides speedups for CD

Acceleration of CD II



Least squares on *rcv1* ($n = p \approx 20k$)

- ▶ Anderson acceleration provides speedups for CD

Theoretical properties

Proposition (Symmetric T)

Let the iteration matrix T be symmetric semi-definite positive, with spectral radius $\rho = \rho(T) < 1$. Let x^* be the limit of the sequence $(x^{(k)})$. Let $\zeta = (1 - \sqrt{1 - \rho}) / (1 + \sqrt{1 - \rho})$. Then the iterates of Anderson acceleration satisfy,^a with $B = (\text{Id} - T)^2$:

$$\|x_{\text{extr}}^{(k)} - x^*\|_B \leq \left(\frac{2\zeta^{K-1}}{1 + \zeta^{2(K-1)}} \right)^{k/K} \|x^{(0)} - x^*\|_B .$$

^aD. Scieur. "Generalized Framework for Nonlinear Acceleration". In: *arXiv preprint arXiv:1903.08764* (2019).

Symmetric T : gradient descent ✓

Coordinate descent?

Theoretical properties

Proposition (Symmetric T)

Let the iteration matrix T be symmetric semi-definite positive, with spectral radius $\rho = \rho(T) < 1$. Let x^* be the limit of the sequence $(x^{(k)})$. Let $\zeta = (1 - \sqrt{1 - \rho}) / (1 + \sqrt{1 - \rho})$. Then the iterates of Anderson acceleration satisfy,^a with $B = (\text{Id} - T)^2$:

$$\|x_{\text{extr}}^{(k)} - x^*\|_B \leq \left(\frac{2\zeta^{K-1}}{1 + \zeta^{2(K-1)}} \right)^{k/K} \|x^{(0)} - x^*\|_B .$$

^aD. Scieur. "Generalized Framework for Nonlinear Acceleration". In: *arXiv preprint arXiv:1903.08764* (2019).

Symmetric T : gradient descent ✓

Coordinate descent?

Coordinate descent (CD)

- ▶ Quadratic problem, with $b \in \mathbb{R}^p$, $H \in \mathbb{S}_{++}^p$, $H \succ 0$:

$$x^* = \arg \min_{x \in \mathbb{R}^p} \frac{1}{2} x^\top H x + \langle b, x \rangle$$

- ▶ The updates of coordinate descent write, for all $j \in 1, \dots, p$:

$$x_j \leftarrow x_j - (H_j \cdot x + b_j) / H_{jj}$$

- ▶ One pass on all the coordinates gives a **fixed point iteration**:

$$x^{(k+1)} = T x^{(k)} + v$$

$$T = \left(\text{Id}_p - e_p e_p^\top H / H_{pp} \right) \dots \left(\text{Id}_p - e_1 e_1^\top H / H_{11} \right)$$

nonsymmetric ✗

Theoretical properties

Weak theoretical properties for AA with non-symmetric T^7

Proposition (Non-symmetric T)

Let T be the iteration matrix of pseudo-symmetric coordinate descent: $T = H^{-1/2}SH^{1/2}$, with S the symmetric positive semidefinite matrix

$$S = \left(\text{Id}_p - H^{1/2} \frac{e_1 e_1^\top}{H_{11}} H^{1/2} \right) \times \cdots \times \left(\text{Id}_p - H^{1/2} \frac{e_p e_p^\top}{H_{pp}} H^{1/2} \right) \\ \times \left(\text{Id}_p - H^{1/2} \frac{e_p e_p^\top}{H_{pp}} H^{1/2} \right) \times \cdots \times \left(\text{Id}_p - H^{1/2} \frac{e_1 e_1^\top}{H_{11}} H^{1/2} \right) .$$

Let x^* be the limit of the sequence $(x^{(k)})$. Let

$\zeta = (1 - \sqrt{1 - \rho}) / (1 + \sqrt{1 - \rho})$. Then $\rho = \rho(T) = \rho(S) < 1$ and the iterates online extrapolation satisfy^a:

$$\|x_{e-on}^{(k)} - x^*\|_B \leq \left(\sqrt{\kappa(H)} \frac{2\zeta^{K-1}}{1+\zeta^{2(K-1)}} \right)^{k/K} \|x^{(0)} - x^*\|_B .$$

^aQ. Bertrand and M. Massias. "Anderson acceleration of coordinate descent". In: *AISTATS*. 2021.

⁷R. Bollapragada, D. Scieur, and A. d'Aspremont. "Nonlinear acceleration of momentum and primal-dual algorithms". In: *arXiv preprint arXiv:1810.04539* (2018).

Algorithm

Algorithm: Online Anderson PCD (proposed)

init: $x^{(0)} \in \mathbb{R}^p$

for $k = 1, \dots$ **do**

$x = x^{(k-1)}$

for $j = 1, \dots, p$ **do**

$\tilde{x}_j = x_j$

$x_j = \text{prox}_{\frac{\lambda}{L_j} g_j} (x_j - A_{:j}^\top \nabla f(Ax) / L_j)$

$Ax += (x_j - \tilde{x}_j) A_{:j}$

$x^{(k)} = x$ // regular iter. $\mathcal{O}(np)$

if $k = 0 \pmod K$ **then** // extrapol., $\mathcal{O}(K^3 + pK^2)$

$U = [x^{(k-K+1)} - x^{(k-K)}, \dots, x^{(k)} - x^{(k-1)}]$

$c = (U^\top U)^{-1} \mathbf{1}_K / \mathbf{1}_K^\top (U^\top U)^{-1} \mathbf{1}_K \in \mathbb{R}^K$

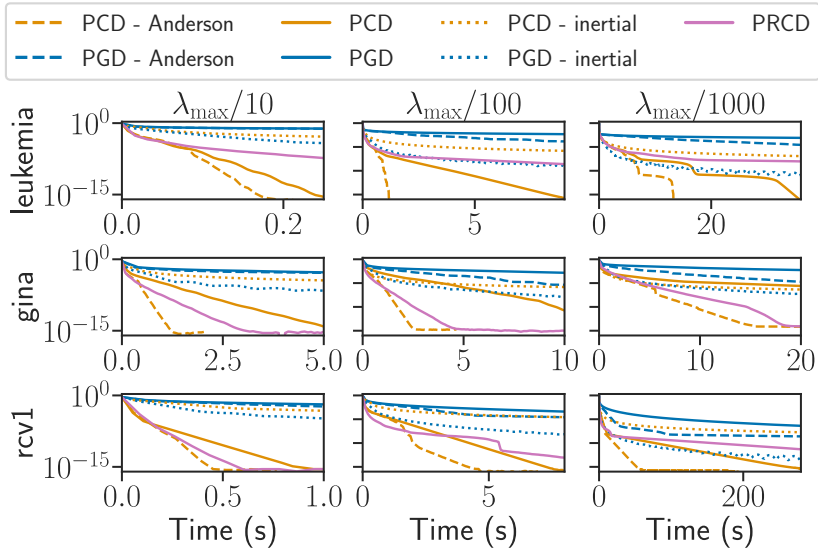
$x_e = \sum_{i=1}^K c_i x^{(k-K+i)}$

if $f(Ax_e) + \lambda g(x_e) \leq f(x^{(k)}) + \lambda g(x^{(k)})$ **then**

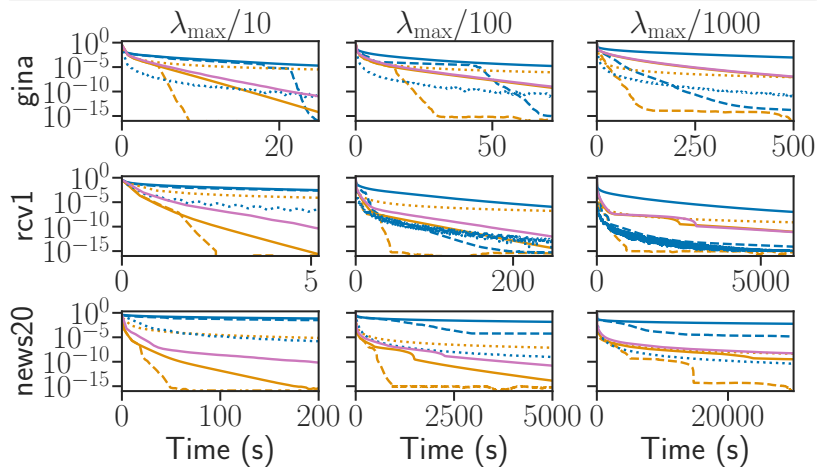
$x^{(k)} = x_e$

return $x^{(k)}$

Lasso



Sparse logistic regression



Conclusion and future work

- ▶ Accelerated proximal coordinate descent in practice
- ▶ Accepted paper⁸: <http://proceedings.mlr.press/v130/bertrand21a/bertrand21a.pdf>
- ▶ Open code: <https://github.com/mathurinm/andersoncd>

Future work:

- ▶ Working sets
- ▶ Non-convex penalties

⁸Q. Bertrand and M. Massias. "Anderson acceleration of coordinate descent". In: *AISTATS*. 2021.

Bibliographie

- ▶ Anderson, D. G. “Iterative procedures for nonlinear integral equations”. In: *Journal of the ACM* 12.4 (1965), pp. 547–560.
- ▶ Bertrand, Q. and M. Massias. “Anderson acceleration of coordinate descent”. In: *AISTATS*. 2021.
- ▶ Bollapragada, R., D. Scieur, and A. d’Aspremont. “Nonlinear acceleration of momentum and primal-dual algorithms”. In: *arXiv preprint arXiv:1810.04539* (2018).
- ▶ Fercoq, O. and P. Richtárik. “Accelerated, parallel, and proximal coordinate descent”. In: *SIAM Journal on Optimization* 25.4 (2015), pp. 1997–2023.
- ▶ Friedman, J. et al. “Pathwise coordinate optimization”. In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332.
- ▶ Lin, Q., Z. Lu, and L. Xiao. “An Accelerated Proximal Coordinate Gradient Method”. In: *NeurIPS*. 2014, pp. 3059–3067.
- ▶ Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python”. In: *JMLR* 12 (2011), pp. 2825–2830.
- ▶ Scieur, D. “Generalized Framework for Nonlinear Acceleration”. In: *arXiv preprint arXiv:1903.08764* (2019).