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:

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

Examples:

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

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

[2] J. 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:

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

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

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

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

CD on least squares

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

Algorithm: Gradient descent

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

for $k = 0, 1, \ldots,$ **do**

$\quad x \leftarrow x - \frac{A^\top(Ax-y)}{\|A\|_2^2}$

return x

Algorithm: CD

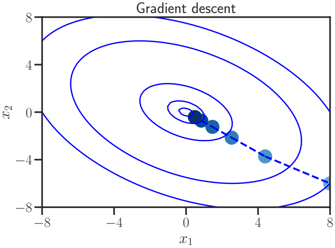
init : $x \in \mathbb{R}^p$

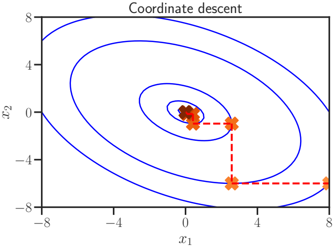
for $k = 0, 1, \ldots,$ **do**

$\quad$ Select $j \in [p]$

$\quad x_j \leftarrow x_j - \frac{A_{:j}^\top(Ax-y)}{\|A_j\|^2}$

return x





Why CD works well?

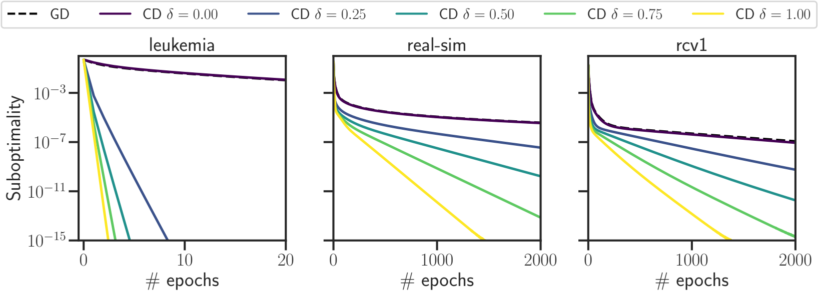
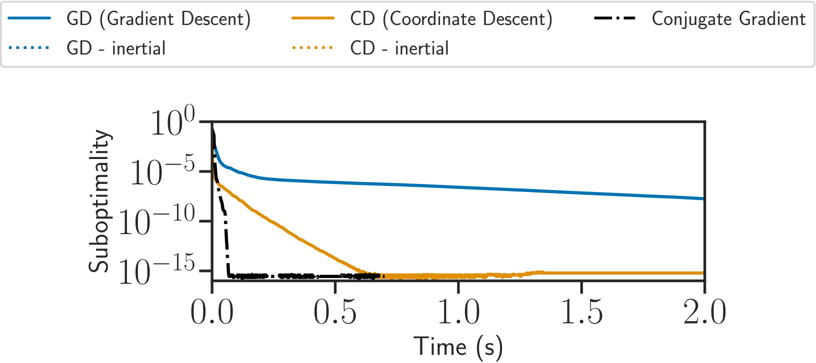


Figure: **Influence of the step size for coordinate descent, OLS.** Gradient descent is compared against coordinate descent with step sizes $\gamma_j = \delta / L_j + (1 - \delta)/L$, for multiple values of δ.

Large step sizes: better convergence

Acceleration of CD



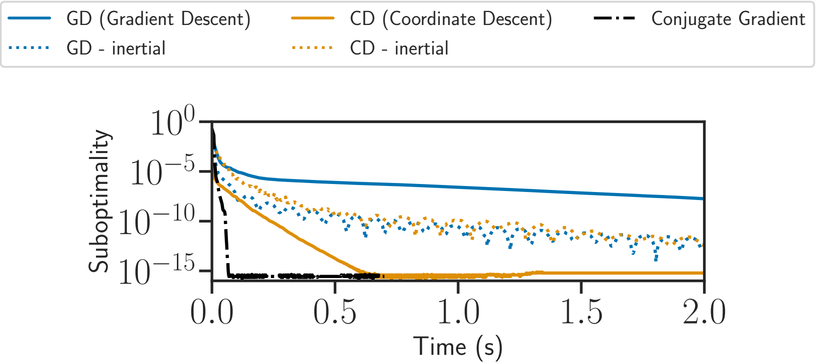
Least squares on *rcv1* ($n = p \approx 20$k)

Nesterov-like **inertial CD**[3,4] **slows down** convergence

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

[4] O. 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* ($n = p \approx 20\text{k}$)

Nesterov-like **inertial CD**[3,4] **slows down** convergence

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

[4] O. 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

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

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

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:

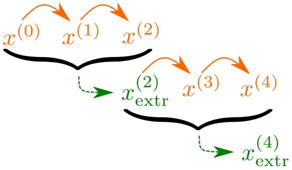
$$\sum_{i=1}^{k} c_i x^{(k-1)} \approx T \sum_{i=0}^{k-1} c_i x^{(k-1)} + b$$

Choose c_i such that

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

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

Anderson acceleration: algorithm[5,6]



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

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

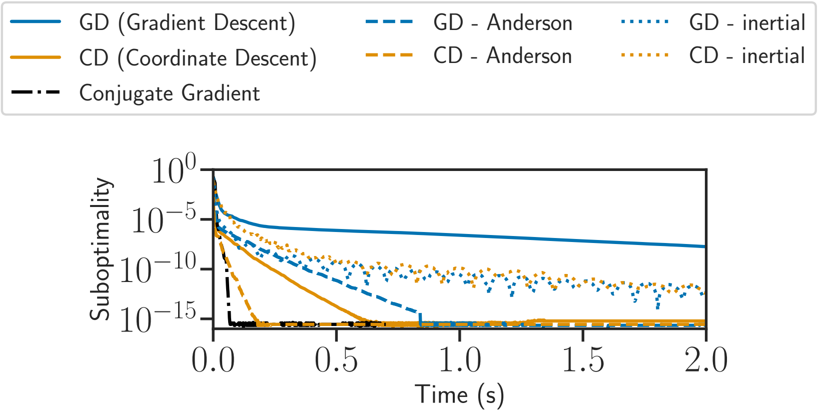
- $x^{(k)} = T x^{(k-1)} + b$ `// regular iter.`
- **if** $k = 0 \mod K$ **then**
  - $U = [x^{(k-K+1)} - x^{(k-K)}, \dots,]$
  - $c = (U^\top U)^{-1} \mathbf{1}_K$
  - $x^{(k)}_{\text{extr}} = \sum_i^K c_i x^{(k-K+i)} / \sum_i c_i$
  - $x^{(k)} = x^{(k)}_{\text{extr}}$ `// base sequence changes`

return $x^{(k)}$

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

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

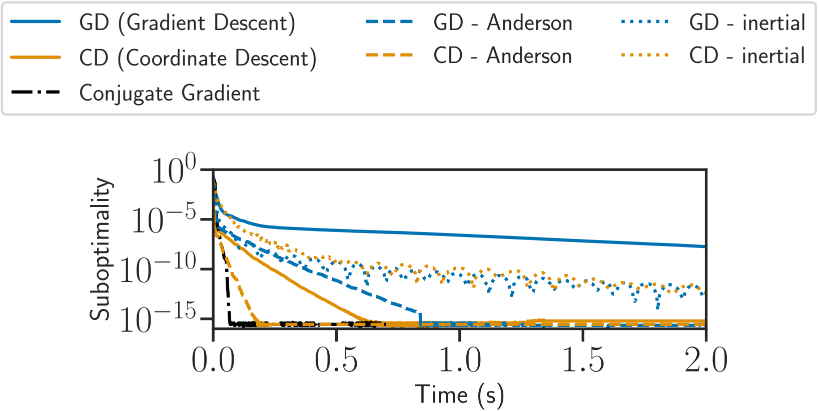
Acceleration of CD II



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

- Anderson acceleration provides speedups for CD

Acceleration of CD II



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

- 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 $B = (\mathrm{Id} - T)^2$:

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

[a] D. 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 $B = (\mathrm{Id} - T)^2$:

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

[a] D. 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 $b \in \mathbb{R}^p$, $H \in \mathbb{S}^p_{++}$, $H \succ 0$:

$$x^* = \mathop{\arg\min}_{x \in \mathbb{R}^p} \tfrac{1}{2} x^\top H x + \langle b, x \rangle$$

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

$$x_j \leftarrow x_j - (H_{j:} x + b_j) / H_{jj}$$

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

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

$$T = \Big(\mathrm{Id}_p - e_p e_p^\top H / H_{pp}\Big) \ldots \Big(\mathrm{Id}_p - e_1 e_1^\top H / H_{11}\Big)$$

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: $T = H^{-1/2} S H^{1/2}$, with S the symmetric positive semidefinite matrix

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

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]*:*

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

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

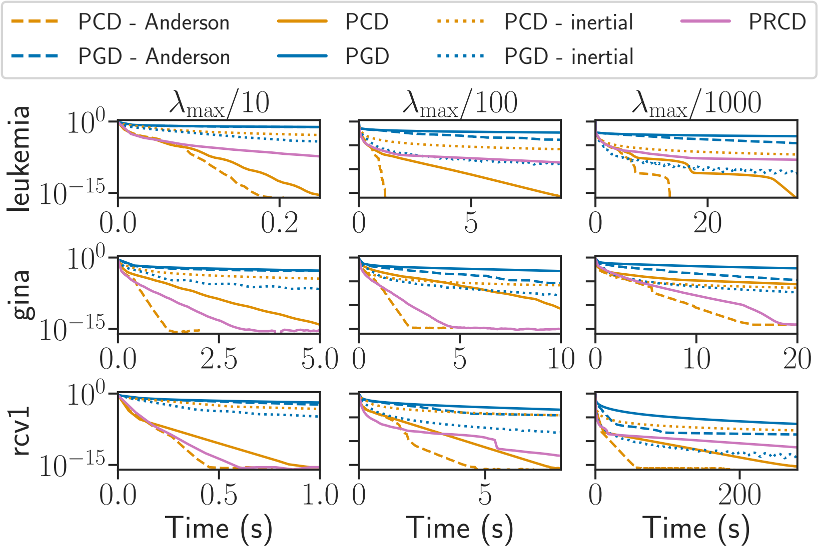
[7] R. Bollapragada, D. Scieur, and A. d'Aspremont. "Nonlinear acceleration of momentum and primal-dual algorithms". In: *arXiv preprint arXiv:1810.04539* (2018).

Algorithm

Algorithm: Online Anderson PCD (proposed)

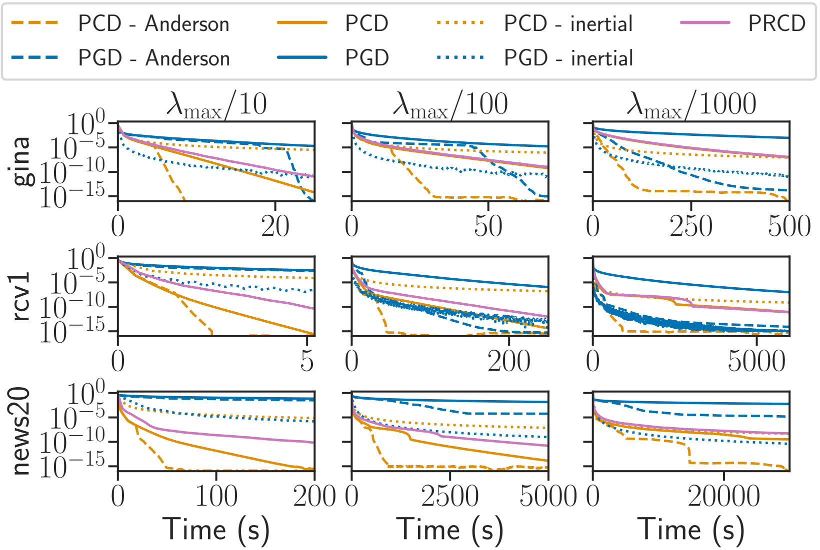
init: $x^{(0)} \in \mathbb{R}^p$
for $k = 1, \dots$ **do**
  $x = x^{(k-1)}$
  for $j = 1, \dots p$ **do**
    $\tilde{x}_j = x_j$
    $x_j = \mathrm{prox}_{\frac{\lambda}{L_j} g_j}(x_j - A_{:j}^\top \nabla f(Ax)/L_j)$
    $Ax \mathrel{+}= (x_j - \tilde{x}_j) A_{:j}$
  $x^{(k)} = x$ `// regular iter.` $\mathcal{O}(np)$
  if $k = 0 \mod K$ **then** `// extrapol.,` $\mathcal{O}(K^3 + pK^2)$
    $U = [x^{(k-K+1)} - x^{(k-K)}, \dots, x^{(k)} - x^{(k-1)}]$
    $c = (U^\top U)^{-1} \mathbf{1}_K / \mathbf{1}_K^\top (U^\top U)^{-1} \mathbf{1}_K \in \mathbb{R}^K$
    $x_{\mathrm{e}} = \sum_{i=1}^K c_i x^{(k-K+i)}$
    if $f(Ax_{\mathrm{e}}) + \lambda g(x_{\mathrm{e}}) \leq f(x^{(k)}) + \lambda g(x^{(k)})$ **then**
      $x^{(k)} = x_{\mathrm{e}}$
return $x^{(k)}$

Lasso



PCD - Anderson
PCD
PCD - inertial
PRCD
PGD - Anderson
PGD
PGD - inertial
$\lambda_{\max}/10$
$\lambda_{\max}/100$
$\lambda_{\max}/1000$
leukemia
gina
rcv1
10^{0}
10^{-15}
0.0
0.2
0
5
20
2.5
5.0
10
0.5
1.0
200
Time (s)

Sparse logistic regression



Conclusion and future work

- Accelerated proximal coordinate descent in practice
- Accepted paper[8]: http://proceedings.mlr.press/v130/bertrand21a/bertrand21a.pdf
- Open code: https://github.com/mathurinm/andersoncd

Future work:

- Working sets
- Non-convex penalties

[8] Q. 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).