Anderson Acceleration of Constrained Optimization Algorithms

Vien V. Mai and Mikael Johansson
KTH - Royal Institute of Technology

Introduction

Generic **unconstrained** convex optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(x)$$

Optimization algorithm produces **sequence** of iterates x_k converging to $x^\star$

Example: gradient descent:

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

Example: Newton's method:

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

Note. Algorithms only keep the last iterate for the next update

Multi-step methods

Use multiple past iterates to accelerate the convergence process

Example: Heavy-ball method [Polyak, 1964]:

$$x_{k+1} = x_k + \beta_k(x_k - x_{k-1}) - \gamma\nabla f(x_k))$$

Example: Accelerated gradient descent [Nesterov, 1983]:

$$x_{k+1} = x_k + \beta_k(x_k - x_{k-1}) - \gamma\nabla f(x_k + \beta_k(x_k - x_{k-1}))$$

Optimal **momentum** parameters β_k often depend on unknown constants

Anderson acceleration [Anderson, 1965]

Originally developed for solving nonlinear integral equations

Recently adapted for solving general fixed-point equations

$$g(x) = x$$

Key ideas. Make clever use of past iterates

- Keeps $m+1$ most recent iterates
- **Ideally**, forms $x_{\text{ext}} = \sum_{i=0}^{m} \alpha_i x_{k-i}$ such that

$$\alpha^\star = \underset{\alpha:\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \left\| g(x_{\text{ext}}) - x_{\text{ext}} \right\|.$$

- Finds coefficients α_i that minimize

$$\alpha \leftarrow \underset{\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \left\| \sum_{i=0}^{m} \alpha_i \left(g(x_{k-i}) - x_{k-i} \right) \right\|$$

- Sets $x_{k+1} = \sum_{i=0}^{m} \alpha_i x_{k-i}$

Optimization algorithms as fixed-point iterations

Many optimization methods can be written as fixed-point iterations

Gradient descent:

$$x_{k+1} = x_k - \gamma \nabla f(x_k) \quad \Leftrightarrow \quad x_{k+1} = g(x_k)$$

with $g(x) := x - \gamma \nabla f(x)$.

Finding an optimal solution reduces to finding a fixed-point of g:

$$g(x^\star) = x^\star \quad \Leftrightarrow \quad \nabla f(x^\star) = 0 \quad \Leftrightarrow \quad x^\star \in \operatorname{argmin} f(x)$$

Idea. Apply Anderson acceleration to speed-up optimization algorithms

[Scieur, d'Aspremont, Bach, 2016], [Zhang, O'Donoghue, and Boyd, 2018]

Anderson acceleration for gradient descent

Goal. Find a point $x^\star$ satisfying $\nabla f(x^\star) = 0$

AA-GD.

- Keeps $m+1$ most recent iterates
- Finds coefficients α_i such that

$$\alpha \leftarrow \underset{\alpha^\top \mathbf{1}=1}{\operatorname{argmin}} \left\| \sum_{i=0}^{m} \alpha_i \nabla f(x_{k-i}) \right\|$$

- Sets $x_{k+1} \leftarrow \sum_{i=0}^{m} \alpha_i x_{k-i}$

If f is convex quadratic and $m = k$, AA-GD is GMRES

Convergence rates for the quadratic objective

Accelerated gradient descent:

$$f(x_k) - f(x^\star) \leq O\left(1/k^2\right)\left(f(x_0) - f(x^\star)\right)$$

Strong **practical performance** requires local adaption

- efficient line-search procedures [Nesterov, 2007]
- adaptive restart techniques [O'Donoghue and Candes, 2014]

Anderson acceleration for GD:

$$f(x_k) - f(x^\star) \leq O\left(\min\left\{1/k^2, e^{-\frac{k}{\sqrt{\kappa}}}\right\}\right)\left(f(x_0) - f(x^\star)\right)$$

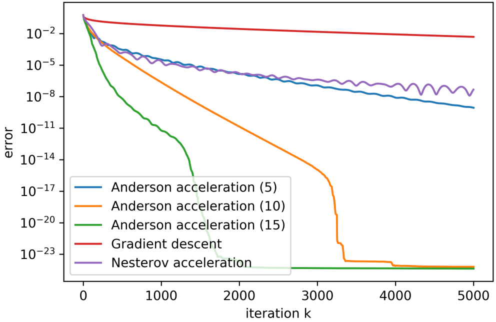
Very strong adaptive rate, remarkable property of Krylov subspace methods

Numerical example

Quadratic convex minimization

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}x^\top Ax + b^\top x$$

- $A \in \mathbb{R}^{200\times 200}$ with $\lambda_{\max}/\lambda_{\min} = 10^4$



Open problem: AA for constrained optimization

Consider **constrained** convex optimization problems

$$\underset{x \in \mathcal{C}}{\text{minimize}}\, f(x) \tag{1}$$

where $\mathcal{C}$ is a closed convex set.

Recall the Anderson acceleration update

$$x_{k+1} = \sum_{i=0}^{m} \alpha_i x_{k-i}, \quad \alpha_i \in \mathbb{R}$$

The extrapolated point x_{k+1} may become **infeasible**

Q. Can we use Anderson acceleration for constrained problems?

Outline

- Background and motivation
- Accelerating projected gradient descent
- Local convergence of AA for constrained problems
- Numerical examples
- Summary and conclusions

Projected gradient descent

$x^\star$ is an optimal solution to (1) if and only if

$$x^\star = \Pi_{\mathcal{C}}(x^\star - \gamma \nabla f(x^\star))$$

Projected gradient descent (PGD)

$$x_{k+1} = \Pi_{\mathcal{C}}(x_k - \gamma \nabla f(x_k))$$

Exactly the fixed-point iteration of the mapping

$$g(x) = \Pi_{\mathcal{C}}(x - \gamma \nabla f(x))$$

Naively using AA for g leads to iterates **infeasible**

Projected gradient descent

$x^\star$ is an optimal solution to (1) if and only if

$$x^\star = \Pi_{\mathcal{C}}(x^\star - \gamma \nabla f(x^\star))$$

An alternative representation of PGD:

$$\begin{aligned} y_{k+1} &= x_k - \gamma \nabla f(x_k) \\ x_{k+1} &= \Pi_{\mathcal{C}}(y_{k+1}) \end{aligned}$$

Can be seen as the fixed-point iteration of the mapping

$$g(y) = \Pi_{\mathcal{C}}(y) - \gamma \nabla f(\Pi_{\mathcal{C}}(y))$$

Observation. $x^\star$ is optimal if and only if

$$y^\star = g(y^\star) \quad \text{and} \quad x^\star = \Pi_{\mathcal{C}}(y^\star)$$

Anderson acceleration for PGD

Key idea. Use AA to speed-up fixed-point computations of $g(y)$

No iterates infeasibility since the sequence $\{y_k\}$ has no restriction

How to relate the convergence of $\{x_k\}$ and $\{y_k\}$:

$$\begin{aligned} \|x_{k+1} - x^\star\| &= \|\Pi_{\mathcal{C}}(y_{k+1}) - \Pi_{\mathcal{C}}(x^\star - \gamma \nabla f(x^\star))\| \\ &\leq \|y_{k+1} - y^\star\| \end{aligned}$$

- if AA quickly drives y_k to $y^\star$, so does x_k to $x^\star$

Convergence guarantee for general smooth mappings

Theorem 1 [Toth and Kelley, 2015]. Suppose that g is **differentiable** and contractive with constant ρ. If x_0 is sufficiently close to $x^\star$, then

$$\|x_k - x^\star\| \leq \rho^k \|x_0 - x^\star\|.$$

So far, all convergence guarantees for AA rely on linearizing g around $x^\star$:

$$g(x) = g(x^\star) + G'(x^\star)(x - x^\star) + o(\|x - x^\star\|)$$

Due to $\Pi_{\mathcal{C}}(\cdot)$, the mapping $g(y)$ defined above is **non-differentiable**

Extending the analysis to non-smooth mappings

Let $F(x) \triangleq x - g(x)$, Theorem 1 indeed only needs the bounds

$$\|F(x) - F'(x^\star)(x - x^\star)\| \leq \frac{c}{2} \|x - x^\star\|^2$$

for some constant $c > 0$ and for all x sufficiently close to $x^\star$.

Extending Theorem 1 to **non-smooth** case amounts to searching for such F'

Two key ingredients:

- Clarke's generalized Jacobian
- (strong) **semi-smoothness**

Performance guarantees of AA-PGD

Lemma 1. Projections onto the nonnegative orthant, second-order cone, positive semidefinite cone, and polyhedral set are all strongly semi-smooth.

Main result. Let f be a μ-strongly convex and L-smooth function. Suppose that $\Pi_{\mathcal{C}}(\cdot)$ is strongly semi-smooth and that x_0 is sufficiently close to $x^\star$. Then,

$$\|x_k - x^\star\| \leq \rho^k \|x_0 - x^\star\|,$$

where $\rho = \sqrt{1 - \gamma 2\mu L/(\mu + L)}$.

Numerical results

Constrained logistic regression.

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \ & \frac{1}{M}\sum_{i=1}^{M}\log(1+\exp(-y_i a_i^\top x)) + \mu\,\|x\|^2 \\ \text{subject to} \ & \|x\|_\infty \le 1, \end{aligned}$$

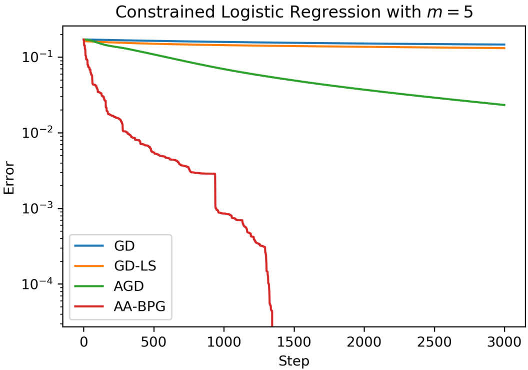
$a_i \in \mathbb{R}^n$ are training samples and $y \in \{-1, 1\}^n$ are the corresponding labels.

We use the UCI Madelon dataset with $M = 2000$ and $n = 500$

We set $\mu = 0.01$ and $m = 5$

- Extremely ill-conditioned problem with condition number $\kappa = 3 \times 10^9$

Numerical results



Numerical results

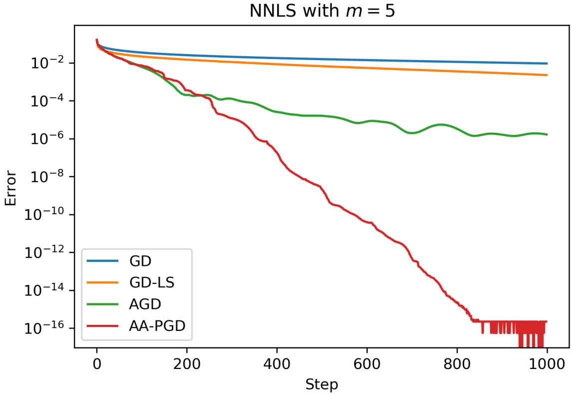


Figure: Nonnegative least-squares, $A \in \mathbb{R}^{200\times 200}$ with $\lambda_{\max}/\lambda_{\min} = 10^4$

Conclusion

Anderson acceleration

- dramatic speed-ups in local convergence, at small extra cost
- current theory only applies to unconstrained problems

Our contributions

- first convergence results for AA on constrained problems
- strong practical performance

Future work

- algorithms: primal-dual methods, ADMM
- applications: Sinkhorn-Knopp, generative adversarial network (GAN)