Anderson Acceleration of Proximal Gradient Methods

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

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

- Finds coefficients α_i that minimize

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

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

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

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 $x_{\mathrm{ext}} = \sum_{i=0}^{m} \alpha_i x_{k-i}$ such that

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

- Finds coefficients α_i that minimize

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

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

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

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.

- Finds coefficients α_i such that

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

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

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

Convergence rates for the quadratic objective

Anderson acceleration for GD (AA-GD):

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

Very strong **adaptive** rate, remarkable property of Krylov subspace methods

Accelerated gradient descent:

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

Strong **practical performance** requires local adaption

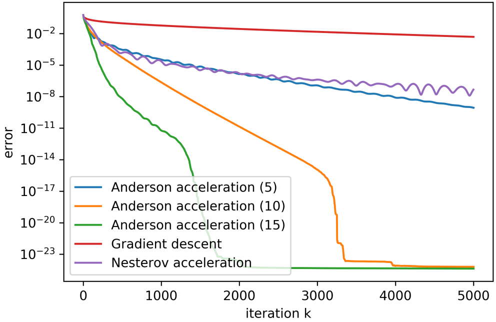
- adaptive restart techniques [O'Donoghue and Candes, 2014]

Numerical example

Quadratic convex minimization

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

- $A \in \mathbb{R}^{200\times 200}$ 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

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

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

Q. Can we use Anderson acceleration for constrained problems?

Outline

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

Composite minimization and proximal operator

Consider composite convex optimization problems

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

- f is smooth and convex; h is closed and convex

Proximal operator:

$$\mathrm{prox}_h(y) := \underset{x \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \frac{1}{2} \|x - y\|_2^2 + h(x) \right\}.$$

$x^\star$ is an optimal solution if and only if

$$x^\star = \mathrm{prox}_{\gamma h}(x^\star - \gamma \nabla f(x^\star))$$

Proximal gradient descent

Proximal gradient algorithm (PGA)

$$x_{k+1} = \mathrm{prox}_{\gamma h}\left(x_k - \gamma \nabla f(x_k)\right)$$

Exactly the fixed-point iteration of the mapping

$$g(x) = \mathrm{prox}_{\gamma h}\left(x - \gamma \nabla f(x)\right)$$

Finding $x^\star$ is equivalently to finding a fixed-point of g

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

Proximal gradient descent

An alternative representation of PGA:

$$y_{k+1} = x_k - \gamma \nabla f(x_k)$$
$$x_{k+1} = \mathrm{prox}_{\gamma h}(y_{k+1})$$

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

$$g(y) = \mathrm{prox}_{\gamma h}(y) - \gamma \nabla f(\mathrm{prox}_{\gamma h}(y))$$

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

$$y^\star = g(y^\star) \quad \text{and} \quad x^\star = \mathrm{prox}_{\gamma h}(y^\star)$$

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 $\{y_k\}$:

$$\begin{aligned}\|x_{k+1} - x^\star\| &= \left\|\mathrm{prox}_{\gamma h}\left(y_{k+1}\right) - \mathrm{prox}_{\gamma h}\left(x^\star - \gamma \nabla f(x^\star)\right)\right\| \\ &\le \|y_{k+1} - y^\star\|\end{aligned}$$

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

Convergence guarantee

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

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

Note. The mapping g is non-differentiable

Assumption. The operator $\operatorname{prox}_h(\cdot)$ is **strongly semi-smooth** [1]

Main result. Let f be a μ-strongly convex and L-smooth function. Suppose that x_0 is sufficiently close to $x^\star$. Then,

$$\|x_k - x^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^\star\|^2.$$

[1] [Mifflin, 1977], [Qi and Sun, 1993]

Numerical results

Constrained logistic regression.

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

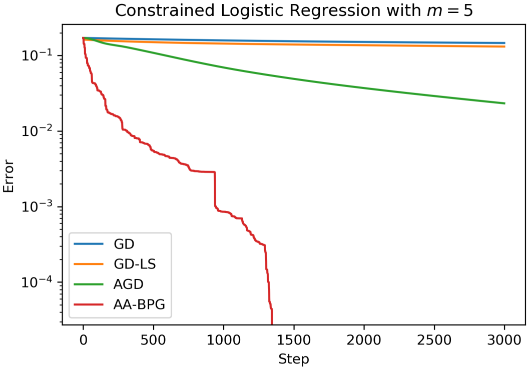
$a_i \in \mathbb{R}^n$ are training samples and $y \in \{-1,1\}^n$ are the corresponding labels.

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

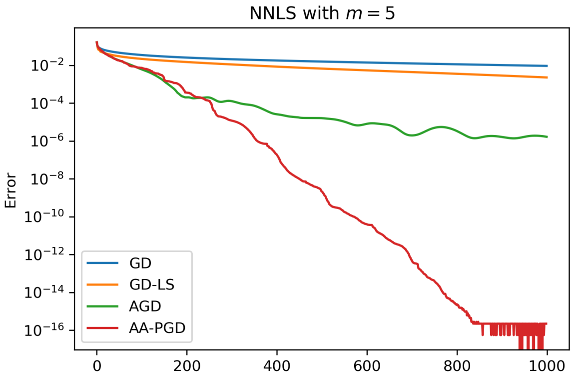


Numerical results

Nonnegative least squares.

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2m} \|Ax - b\|^2 \quad \text{subject to} \ \ x \geq 0,$$

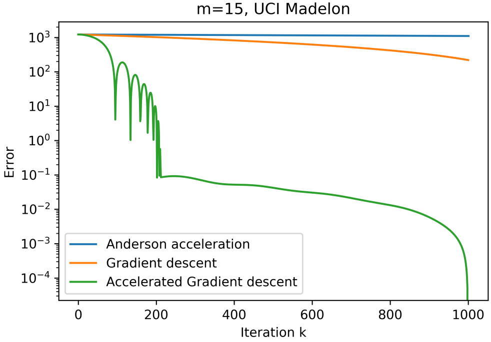
with $A \in \mathbb{R}^{1000 \times 5000}$ and $b \in \mathbb{R}^{1000}$.



Challenges with global convergence

Even for smooth problems, AA may not converge!

Example. AA-GD gets stuck in unconstrained logistic regression



The real algorithm (Guarded-AA)

1. Computes

$$g_k = x_k - \gamma \nabla f(x_k) \qquad \text{(gradient step)}$$

2. Applies AA-PGA

$$y_{\text{ext}} = \sum_{i=0}^{m_k} \alpha_i^k g_{k-i} \quad \text{and} \quad x_{\text{test}} = \operatorname{prox}_{\gamma h}(y_{\text{ext}})$$

3. Guarded step

If $f(x_{\text{test}}) \leq f(x_k) - \frac{\gamma}{2} \|\nabla f(x_k)\|_2^2$

$\quad x_{k+1} = x_{\text{test}}, \quad y_{k+1} = y_{\text{ext}}$

else

$\quad x_{k+1} = \operatorname{prox}_{\gamma h}(g_k), \quad y_{k+1} = g_k$

end

Numerical results

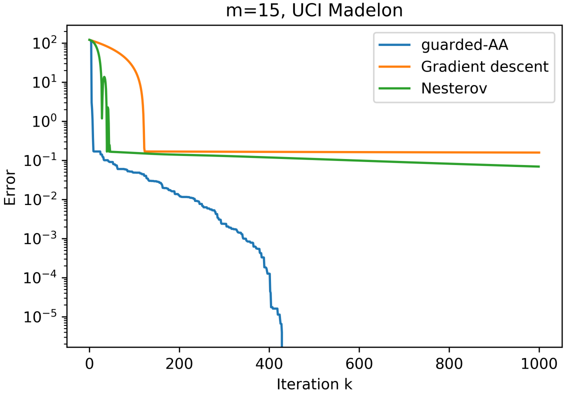


Figure: Bounded logistic regression

Extending to Non-Euclidean Geometry

Consider optimization problems of the form

$$\underset{x \in \mathcal{D}}{\text{minimize}}\ f(x) + h(x).$$

Problem geometry is exploited by a **kernel** function φ

Example. Energy function $\varphi(x) = (1/2)\,\|x\|_2^2$

Example. Shannon entropy $\varphi(x) = \sum_{i=1}^{n} x_i \log x_i$, $\operatorname{dom}\varphi = \mathbb{R}_+^n$

Bregman proximal operator:

$$\operatorname{prox}_h^{\varphi}(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ h(x) + D_{\varphi}(x, y) \right\}, \quad y \in \operatorname{int}\operatorname{dom}\varphi.$$

Bregman proximal gradient (BPG)

Bregman proximal gradient (BPG)

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), x - x_k \rangle + \gamma^{-1} D_\varphi(x, x_k) + h(x) \right\}.$$

Can be expressed as

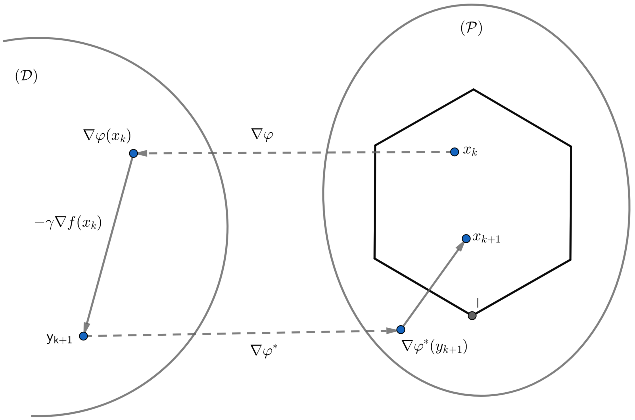
$$x_{k+1} = \mathrm{prox}_{\gamma h}^{\varphi}\left(\nabla \varphi^*\left(\nabla \varphi(x_k) - \gamma \nabla f(x_k)\right)\right)$$

Equivalent form

$$\begin{aligned} y_{k+1} &= \nabla \varphi(x_k) - \gamma \nabla f(x_k) \\ x_{k+1} &= \mathrm{prox}_{\gamma h}^{\varphi}\left(\nabla \varphi^*\left(y_{k+1}\right)\right). \end{aligned}$$

Mirror Descent is a special instance

Illustration of Bregman Proximal gradient



Anderson acceleration of BPG

Strategy: Extrapolate the **dual** sequence $\{y_k\}$:

$$g(y) = \nabla\varphi(\mathrm{prox}_{\gamma h}^{\varphi} \circ \nabla\varphi^*(y)) - \gamma\nabla f(\mathrm{prox}_{\gamma h}^{\varphi} \circ \nabla\varphi^*(y)).$$

Assumption. The conjugate of φ has full domain, i.e., $\mathrm{dom}\,\varphi^* = \mathbb{R}^n$.

Reduced to AA-PGA if $\varphi(\cdot) = (1/2)\left\|\cdot\right\|_2^2$

A similar guarded step as in AA-PGA guarantees global convergence

Relative-entropy nonnegative regression

The task is to reconstruct the signal $x \in \mathbb{R}^n_+$ by solving

$$\underset{x}{\text{minimize}}\, D_{\text{KL}}\left(Ax, b\right) + \lambda \left\|x\right\|_1 \quad \text{subject to} \;\; x \geq 0,$$

- $A \in \mathbb{R}^{m\times n}_+$ is given nonnegative observation matrix
- $b \in \mathbb{R}^m_{++}$ is a noisy measurement vector

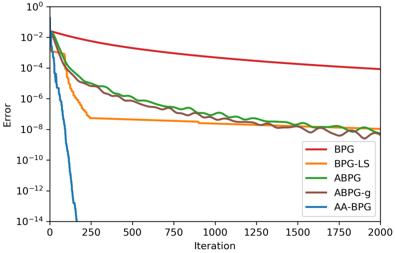
We adapt the family of BPG methods with:

- $\mathcal{D} = \mathbb{R}^n_+$
- φ is the Shannon entropy, $f(x) = D_{\text{KL}}\left(Ax, b\right)$
- $h(x) = \lambda \left\|x\right\|_1$ with $\lambda = 0.001$

Compare with BPG, BPG-LS, ABPG, ABPG-LS

[Bauschke, Bolte, Teboulle, 2016], [Hanzely, Richtárik, Xiao, 2018]

Numerical results



(a) (m,n) = (1000,100)

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
- first application to non-Euclidean geometry

Future work

- algorithms: primal-dual methods, ADMM
- applications: Sinkhorn-Knopp, optimal transport