Optimization for Deep Learning

Lecture 3: Accelerated Gradient Descent

Kun Yuan

Peking University

Main contents in this lecture

- Gradient descent with Polyak's momentum
- Gradient descent with Nesterov's momentum
- Anderson acceleration
- Lower bound and optimal algorithms

Gradient descent

- Recall the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x)$$

- Gradient descent recursion:

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

- When $f(x)$ is convex and L-smooth, GD converges at rate $O(L/k)$

- When $f(x)$ is further μ-strongly convex, GD converges at $O((1 - \frac{\mu}{L})^k)$

Gradient descent can be slow

- Gradient descent can be very slow for ill-conditioned problems
- For example, GD converges very slow when μ/L is sufficiently small[1]

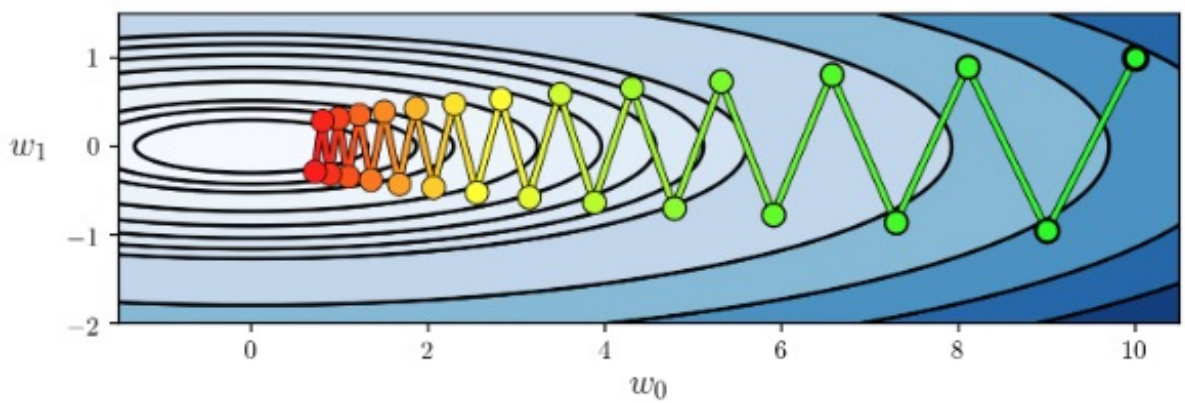

Figure: GD converges slow for ill-conditioned problem

[1] Image is from `https://github.com/jermwatt/machine_learning_refined`

Gradient descent with Polyak's momentum

- We have to alleviate the "Zig-Zag" to accelerate the algorithm

- **Polyak's momentum** method, a.k.a, **heavy-ball** gradient method

$$x_k = x_{k-1} - \gamma \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$$

where $\beta \in (0, 1)$ is the momentum parameter

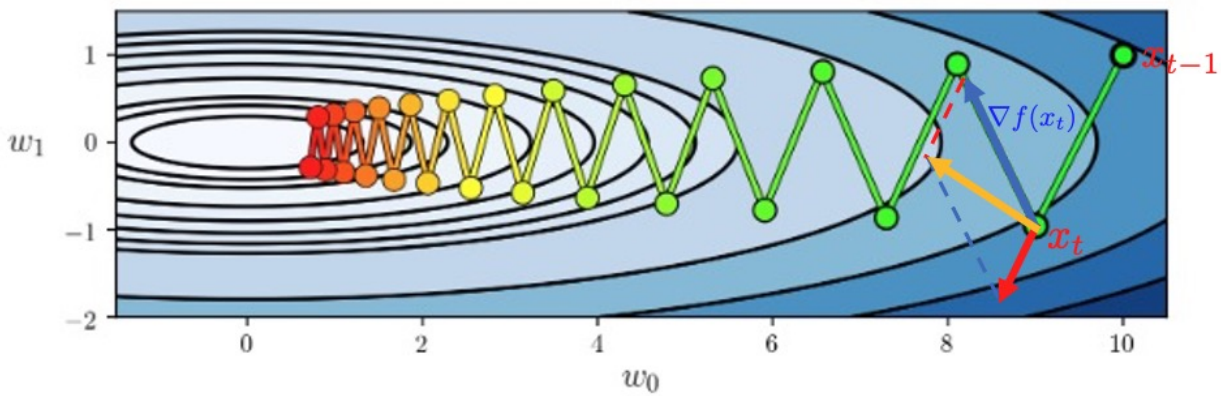

Figure: Momentum can alleviate the "Zig-Zag"

Gradient descent with Polyak's momentum

- We have to alleviate the "Zig-Zag" to accelerate the algorithm

- **Polyak's momentum** method, a.k.a, **heavy-ball** gradient method

$$x_k = x_{k-1} - \gamma \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$$

where $\beta \in (0, 1)$ is the momentum parameter

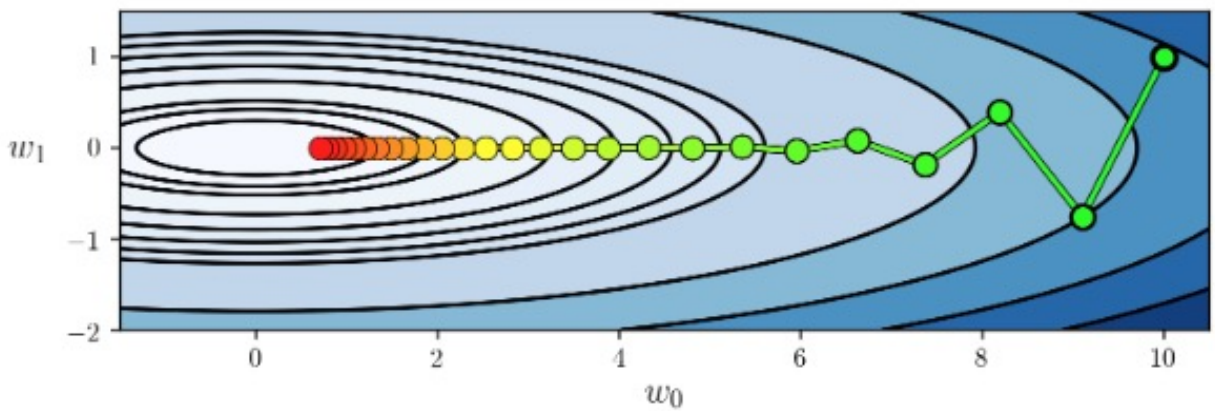

Figure: Momentum can alleviate the "Zig-Zag"

Boris Polyak

Boris Polyak (1935—2023)

Russian mathematician

Polayk's momentum; stochstic gradient descent; stochastic averaging

Gradient descent with Polyak's momentum

$$\min_x \quad f(x) = \frac{1}{2} x^\top A x \quad \text{where} \quad A = \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}$$

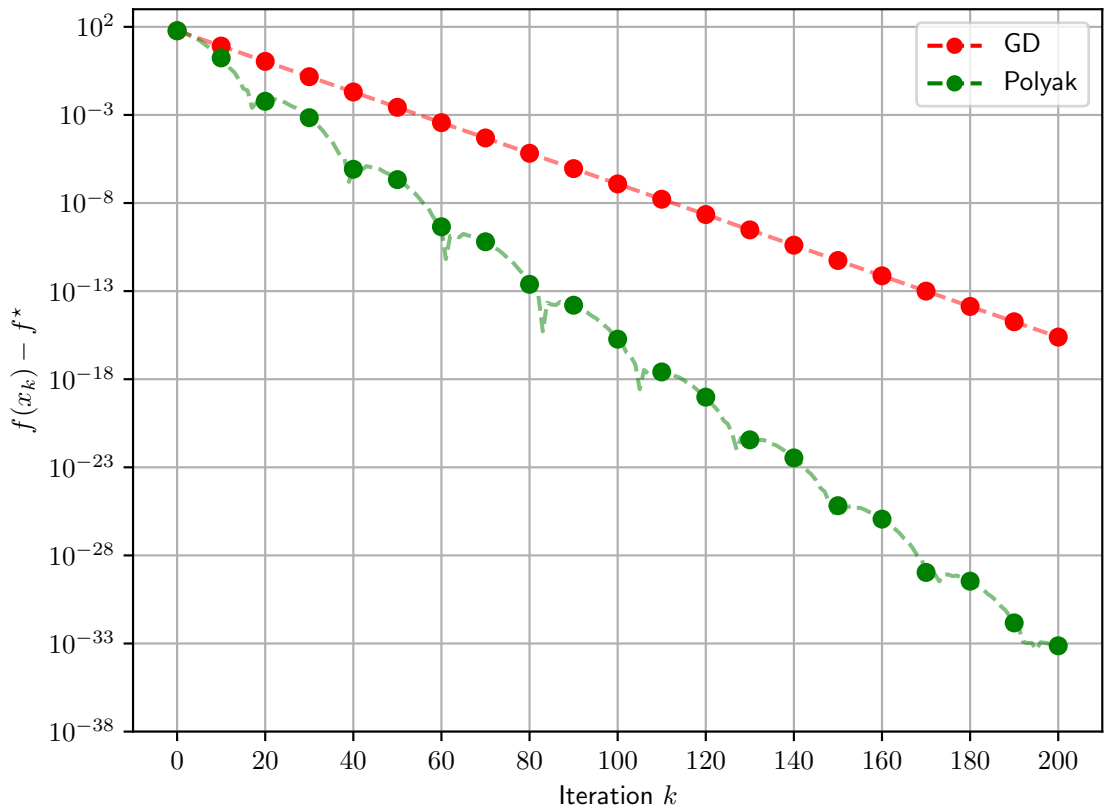

Figure: GD v.s. Polayk momentum

Gradient descent with Polyak's momentum

- Very intutive that Polyak's momentum can work
- But can Polyak's momentum **theoretically** improve the convergence? Yes!

Polyak's momentum leads to theoretically faster convergence

- Consider the following toy example:

$$\min_x \quad f(x) = \frac{1}{2} x^\top A x \quad \text{where} \quad A = \begin{bmatrix} L & 0 \\ 0 & \mu \end{bmatrix} \tag{1}$$

- Apparently, $f(x)$ is L-smooth and μ-strongly convex

- Gradient descent solves the above problem at rate $O((1-\mu/L)^k)$

- Very slow when $\mu/L \approx 0$, say $\mu/L = 0.01$.

Polyak's momentum leads to theoretically faster convergence

- Gradient descent with Polyak's momentum:

$$x_{k+1} = x_k - \gamma A x_k + \beta(x_k - x_{k-1})$$

- With an additional equality $x_k = x_k$, we can rewrite the recursion as

$$\underbrace{\begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix}}_{\mathbf{x}_{k+1}} = \underbrace{\begin{bmatrix} (1+\beta)I - \gamma A & -\beta I \\ I & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}}_{\mathbf{x}_k}. \tag{2}$$

Therefore, we rewrite the heavy ball gradient method into a linear system.

Polyak's momentum leads to theoretically faster convergence

- Given the above matrix $\mathbf{A}$, there exists a permutation matrix P such that

$$\mathbf{A} = P \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} P^T \tag{3}$$

where $PP^T = I$ and

$$A_1 = \begin{bmatrix} 1+\beta-\gamma L & -\beta \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1+\beta-\gamma \mu & -\beta \\ 1 & 0 \end{bmatrix}$$

- A_1 and A_2 can be eigen-decomposed as follows

$$A_1 = U_1 \Lambda_1 U_1^{-1} \quad \text{where} \quad \Lambda_1 = \text{diag}\{\lambda_{11}, \lambda_{12}\}$$
$$A_2 = U_2 \Lambda_2 U_2^{-1} \quad \text{where} \quad \Lambda_2 = \text{diag}\{\lambda_{21}, \lambda_{22}\}$$

where λ_{11} and λ_{12} are the eigenvalues of A_1, and λ_{21} and λ_{22} are the eigenvalues of A_2

Polyak's momentum leads to theoretically faster convergence

- Substituting the above eigen-decomposition into (3), we get

$$\mathbf{A} = \underbrace{P \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} U_1^{-1} & 0 \\ 0 & U_2^{-1} \end{bmatrix} P^\top}_{\mathbf{U}^{-1}}$$

- Substituting the above relation into (2), heavy-ball GD becomes

$$\mathbf{x}_{k+1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}\mathbf{x}_k$$

  where $\mathbf{U}$ is an invertible matrix and $\mathbf{\Lambda}$ is a diagonal matrix

- Left-multiplying $\mathbf{U}^{-1}$ to both sides of the above recursion, we have

$$\mathbf{y}_{k+1} = \mathbf{\Lambda}\mathbf{y}_k$$

  where $\mathbf{y}_k = \mathbf{U}^{-1}\mathbf{x}_k$.

Polyak's momentum leads to theoretically faster convergence

- In summary, the heavy-ball GD method (2) becomes

$$\mathbf{y}_{k+1} = \mathbf{\Lambda}\mathbf{y}_k = \mathbf{\Lambda}^k\mathbf{y}_1$$

  which further implies that

$$\|\mathbf{y}_{k+1}\| \leq \|\mathbf{\Lambda}^k\|\|\mathbf{y}_1\| = \max\{|\lambda_{11}|^k, |\lambda_{12}|^k, |\lambda_{21}|^k, |\lambda_{22}|^k\}\|\mathbf{y}_1\|$$

- Due to the definition of λ, we have

$$\lambda_{1i} = \frac{(1+\beta-\gamma L) \pm \sqrt{(1+\beta-\gamma L)^2 - 4\beta}}{2}, i = 1, 2$$
$$\lambda_{2i} = \frac{(1+\beta-\gamma \mu) \pm \sqrt{(1+\beta-\gamma \mu)^2 - 4\beta}}{2}, i = 1, 2$$

Polyak's momentum leads to theoretically faster convergence

- When $(1+\beta-\gamma L)^2-4\beta \le 0$, we have $|\lambda_{11}| = |\lambda_{12}| = \sqrt{\beta}$, i.e.,

$$\frac{(1-\sqrt{\beta})^2}{L} \le \gamma \le \frac{(1+\sqrt{\beta})^2}{L} \quad \Longrightarrow \quad |\lambda_{11}| = |\lambda_{12}| = \sqrt{\beta}$$

- Similarly, when $(1+\beta-\mu L)^2-4\beta \le 0$, we have $|\lambda_{21}| = |\lambda_{22}| = \sqrt{\beta}$, i.e.,

$$\frac{(1-\sqrt{\beta})^2}{\mu} \le \gamma \le \frac{(1+\sqrt{\beta})^2}{\mu} \quad \Longrightarrow \quad |\lambda_{21}| = |\lambda_{22}| = \sqrt{\beta}$$

- With the above relations, we have

$$\begin{cases} \frac{(1-\sqrt{\beta})^2}{\mu} \le \frac{(1+\sqrt{\beta})^2}{L} \\ \gamma \in \left[\frac{(1-\sqrt{\beta})^2}{\mu}, \frac{(1+\sqrt{\beta})^2}{L}\right] \end{cases} \quad \Longrightarrow \quad \|\mathbf{y}_{k+1}\| \le (\sqrt{\beta})^k \|\mathbf{y}_k\|$$

Polyak's momentum leads to theoretically faster convergence

- Due to $\frac{(1-\sqrt{\beta})^2}{\mu} \leq \frac{(1+\sqrt{\beta})^2}{L}$, we have $\sqrt{\beta} \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

- When $\sqrt{\beta} = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ and $\gamma = \frac{(1-\sqrt{\beta})^2}{\mu} = \frac{(1+\sqrt{\beta})^2}{L}$, it holds that

$$\|\mathbf{y}_{k+1}\| \leq \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \|\mathbf{y}_k\|$$

- Since $\|\mathbf{x}_k\| = \|\mathbf{U}\mathbf{y}_k\| \leq \|\mathbf{U}\|\|\mathbf{y}_k\|$, we finally achieve the convergence rate of the heavy-ball gradient method when solving (1):

$$\|x_k\| \leq \|\mathbf{x}_{k+1}\| \leq \|\mathbf{U}\|\|\mathbf{U}^{-1}\|\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \|\mathbf{x}_1\| = O\left((1-\sqrt{\frac{\mu}{L}})^k\right)$$

where $\|\mathbf{U}\|, \|\mathbf{U}^{-1}\|$ and $\|\mathbf{x}_1\|$ can be regarded as constants.

Gradient descent v.s. heavy-ball method

- Consider the optimization problem

$$\min_x \quad f(x) = \frac{1}{2} x^\top A x \quad \text{where} \quad A = \begin{bmatrix} L & 0 \\ 0 & \mu \end{bmatrix}$$

- Comparison between gradient descent and heavy-ball method

	Convergence rate	Iteration complexity
Gradient descent	$O\big((1-\frac{\mu}{L})^k\big)$	$O\big(\frac{L}{\mu}\log(1/\epsilon)\big)$
Polyak's momentum	$O\big((1-\sqrt{\frac{\mu}{L}})^k\big)$	$O\big(\sqrt{\frac{L}{\mu}}\log(1/\epsilon)\big)$

- Heavy ball is significantly faster than gradient descent for ill-conditioned problem with large L/μ; **momentum accelerates gradient descent**!

Drawbacks in heavy-ball method

- The accelerated rate of heavy-ball method can only be theoretically established for quadratic optimization algorithms
- It is unknown whether heavy-ball can theoretically outperform gradient descent in problems other than quadratic optimization problems
- In practice, heavy ball is always faster than gradient descent

Gradient descent with Nesterov's momentum

- Gradient descent with **Nesterov's momentum**, a.k.a, **Nesterov accelerated gradient (NAG)** method

$$y_{k-1} = x_{k-1} + \beta(x_{k-1} - x_{k-2})$$
$$x_k = y_{k-1} - \gamma \nabla f(y_{k-1})$$

where $\beta \in (0, 1)$ is the momentum parameter

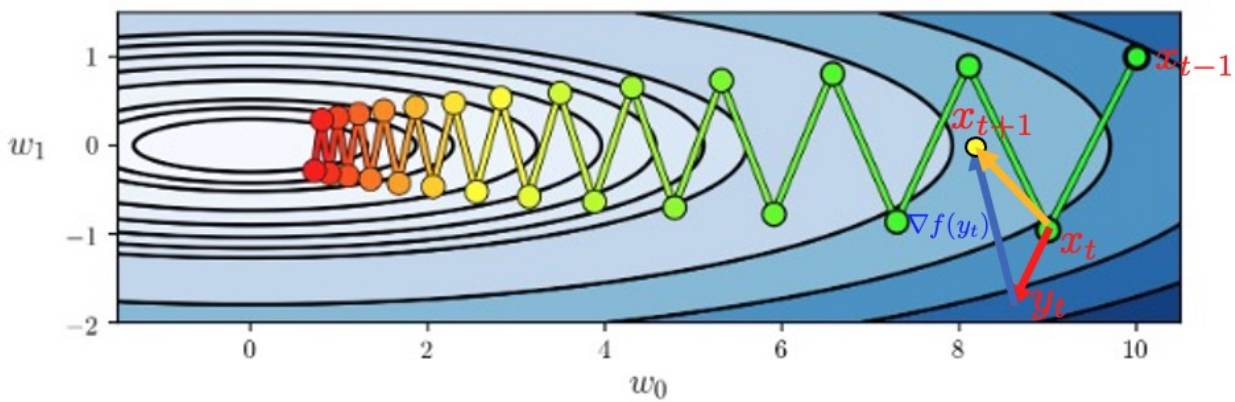

Figure: Nesterov method can alleviate the "Zig-Zag"

Yurii Nesterov

Yurii Nesterov (1956—)

Russian mathematician

Boris Polyak's student

Nesterov acceleration

Gradient descent with Nesterov's momentum

$$\min_x \quad f(x) = \frac{1}{2} x^\top A x \quad \text{where} \quad A = \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}$$

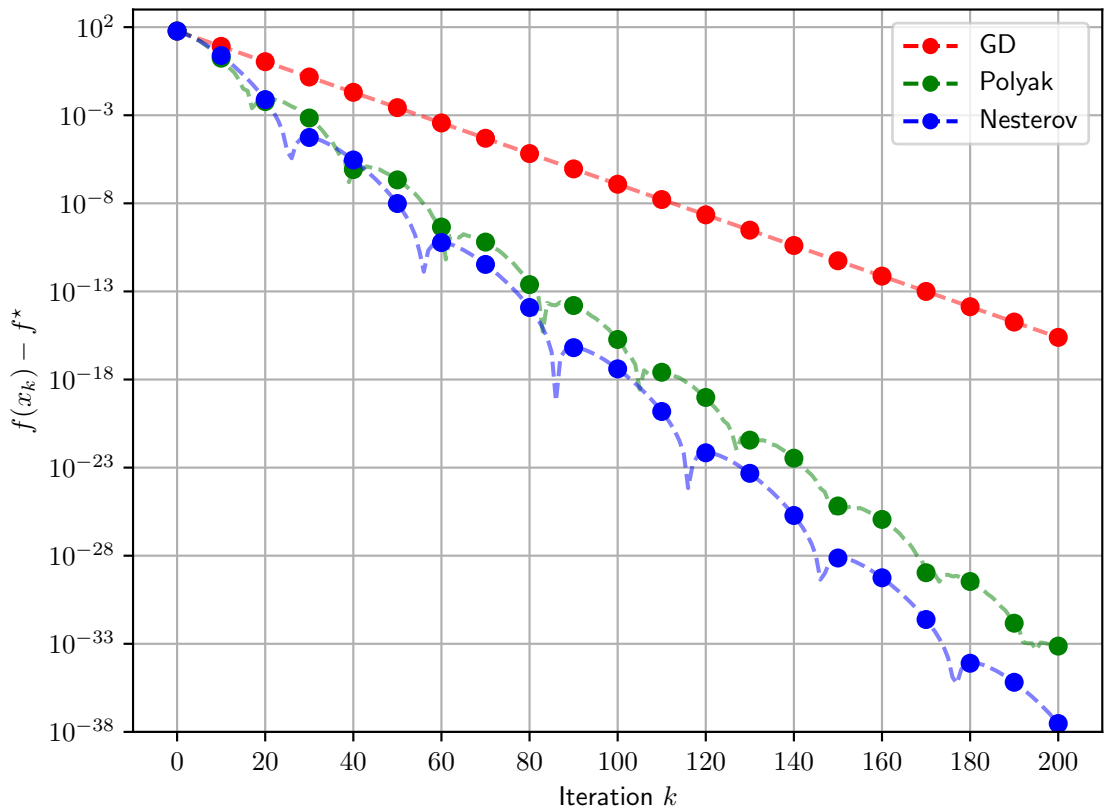

Figure: GD v.s. various momentums

NAG Convergence under smooth and convex setting

Theorem

When $f(x)$ is convex and L-smooth, if $\beta = (k-2)/(k+1)$ and $\gamma = 1/L$, it holds that

$$f(x_T) - f^\star = O\left(\frac{L}{k^2}\right)$$

Recall that vanilla gradient descent converges at rate $O(L/k)$

It is observed that NAG is theoretically faster in convergence rate

NAG Convergence under smooth and strongly-convex setting

Theorem

When $f(x)$ is μ-strongly convex and L-smooth, if $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ and $\gamma = 1/L$, it holds that

$$f(x_k) - f^\star = O\left((1 - \sqrt{\frac{\mu}{L}})^k\right)$$

Recall that vanilla gradient descent converges at rate $O\left((1 - \frac{\mu}{L})^k\right)$

It is observed that NAG is theoretically faster in convergence rate

Gradient descent v.s. NAG

Method	Convexity	Rate	Complexity
GD	Convex	$O(L/k)$	$O(L/\epsilon)$
	Strongly convex	$O((1-\frac{\mu}{L})^k)$	$O(\frac{L}{\mu}\log(1/\epsilon))$
NAG	Convex	$O(L/k^2)$	$O(L/\sqrt{\epsilon})$
	Strongly convex	$O((1-\sqrt{\frac{\mu}{L}})^k)$	$O(\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$

NAG is **theoretically faster** than GD in **convex** scenarios, especially in ill-conditioned scenarios.

Is there an even faster algorithm than NAG?

- Is there any algorithm that can converge theoretically even better than NAG?
- Is it possible to use more historical gradients and variables to produce a better algorithm than NAG?

Many researchers have thought about it

- **Anderson acceleration** updates x_{k+1} using $m+1$ historical variables:

$$x_{k+1} = \sum_{i=0}^{m} a_i^k x_{k-i} = \sum_{i=0}^{m} a_i^k \left(x_{k-i-1} - \gamma \nabla f(x_{k-i-1}) \right)$$

  where $\{a_i^k\}_{i=0}^m$ satisfies $\sum_{i=0}^m a_i^k = 1$.

- To choose proper $\{a_i^k\}_{i=0}^m$, Anderson acceleration uses the following metric:

$$\min_{\{a_i^k\}_{i=0}^m} \|\nabla f(x_{k+1})\|^2 = \|\nabla f\Big(\sum_{i=0}^{m} a_i^k x_{k-i} \Big)\|^2, \quad \text{s.t.} \quad \sum_{i=0}^{m} a_i^k = 1.$$

  which is typically difficult to solve.

- Instead, Anderson acceleration proposes an easy approximate problem

$$\min_{\{a_i^k\}_{i=0}^m} \| \sum_{i=0}^{m} a_i^k \nabla f(x_{k-i})\|^2, \quad \text{s.t.} \quad \sum_{i=0}^{m} a_i^k = 1.$$

Anderson acceleration: algorithm

- Let $G^k = [\nabla f(x_k), \cdots, \nabla f(x_{k-m})] \in \mathbb{R}^{d\times(m+1)}$, Anderson acceleration is

$$\boldsymbol{\alpha}^k = \underset{\boldsymbol{\alpha}:1^\top\boldsymbol{\alpha}=1}{\arg\min}\{\|G^k\boldsymbol{\alpha}\|^2\} \tag{4}$$

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

- Problem (4) has a closed-form solution if G^k has full column rank, we leave it as an exercise.

- Compared to NAG, Anderson acceleration uses more historical variables. It is supposed to be faster than NAG.

Anderson acceleration: experiment

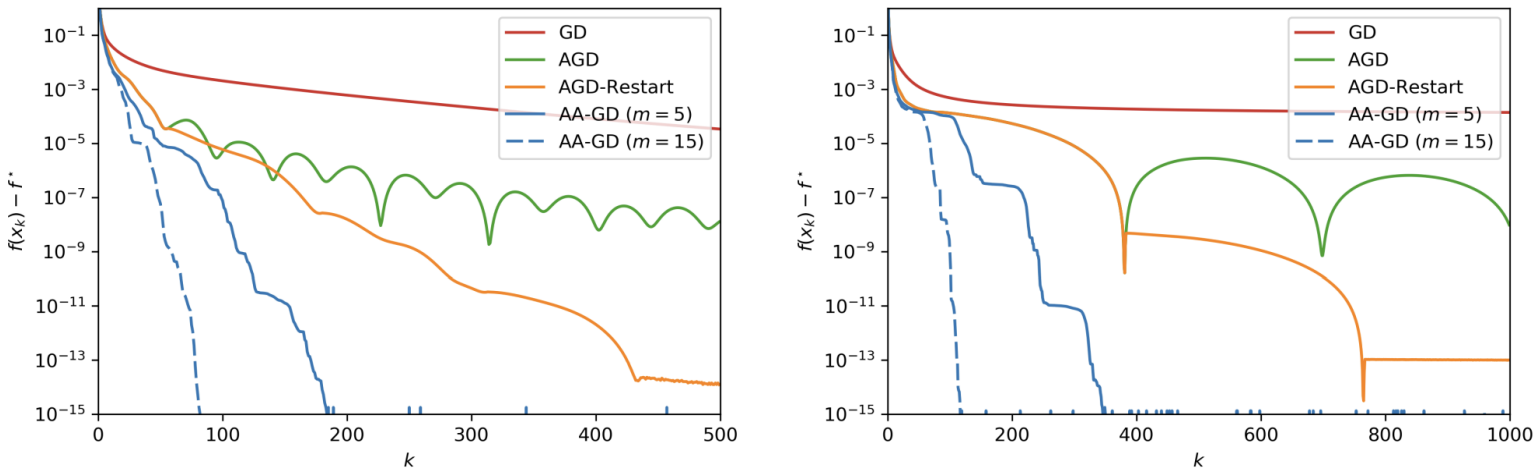

Figure: The figure is from (Mai and Johansson, 2020). The left figure is for a quadratic problem with $L/\mu = 10^3$, and the right one is with $L/\mu = 10^4$.

Anderson acceleration: experiment

$$\min_x \quad f(x) = \frac{1}{2} x^\top A x \quad \text{where} \quad A = \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}$$

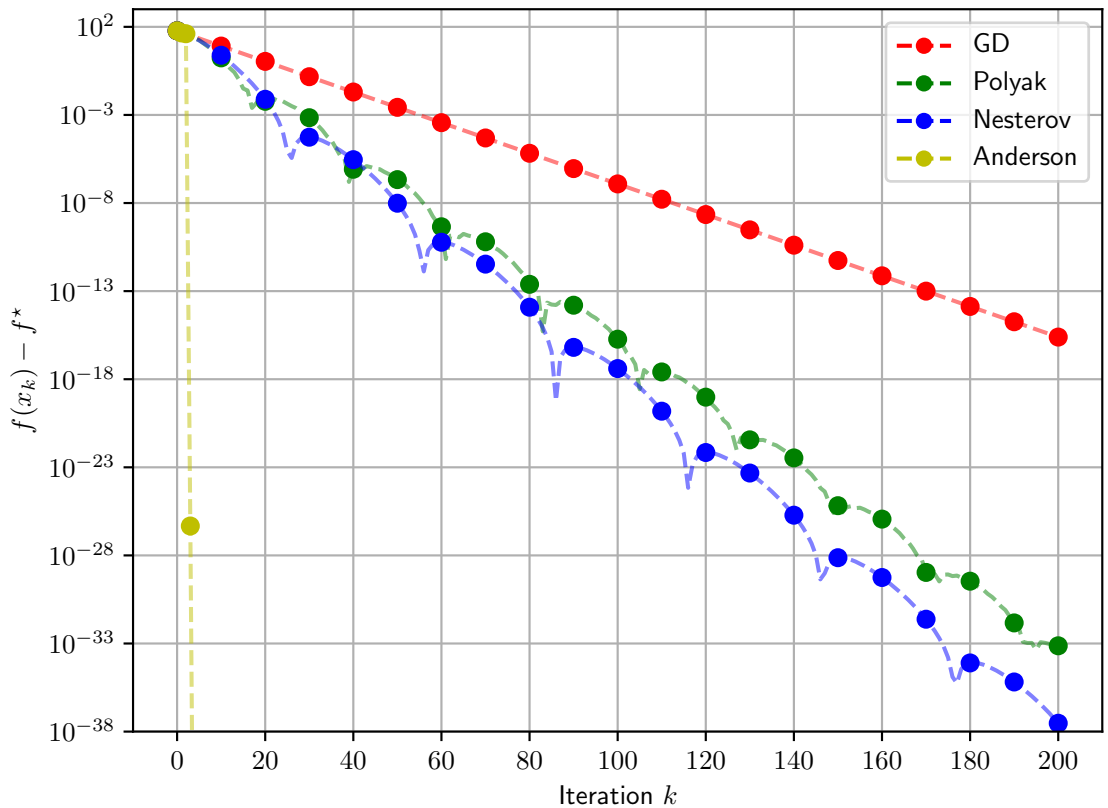

Figure: Convergence comparison

Lower bound

- Is Anderson theoretically better than NAG? **No, it is not!**

- Is there any algorithm theoretically better than NAG? **No, there is not!**

Theorem (Lower bound (Nesterov, 2003))

For any first-order algorithm satisfying

$$x_t \in x_0 + \textit{span}\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_{t-1})\},$$

there always exists some convex and L-smooth $f(x)$ such that

$$f(x_k) - f^\star \geq \frac{3L\|x_0 - x^\star\|^2}{32(k+1)^2}$$

It implies the rate $O(L/k^2)$ cannot be improved. Similarly, the rate $O((1-\sqrt{\mu/L})^k)$ in the strongly convex scenario cannot be improved

NAG has optimal convergence rate

Method	Convexity	Rate	Complexity
GD	Convex	$O(L/k)$	$O(L/\epsilon)$
	Strongly convex	$O((1-\frac{\mu}{L})^k)$	$O(\frac{L}{\mu}\log(1/\epsilon))$
NAG	Convex	$O(L/k^2)$	$O(L/\sqrt{\epsilon})$
	Strongly convex	$O((1-\sqrt{\frac{\mu}{L}})^k)$	$O(\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$
Lower bound	Convex	$\Omega(L/k^2)$	$\Omega(L/\sqrt{\epsilon})$
	Strongly convex	$\Omega((1-\sqrt{\frac{\mu}{L}})^k)$	$\Omega(\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$

NAG has achieved the optimal rate and compelxity; **it cannot be improved**

Anderson typically outperforms NAG in most practical problems, but, for the worst-case problem, it cannot be faster than NAG

Wait! What about non-convex problem?

Method	Convexity	Rate	Complexity
GD	Non-convex	$O(L/k)$	$O(L/\epsilon)$
	Convex	$O(L/k)$	$O(L/\epsilon)$
	Strongly convex	$O((1-\frac{\mu}{L})^k)$	$O(\frac{L}{\mu}\log(1/\epsilon))$
NAG	Non-convex	$O(L/k)$	$O(L/\epsilon)$
	Convex	$O(L/k^2)$	$O(L/\sqrt{\epsilon})$
	Strongly convex	$O((1-\sqrt{\frac{\mu}{L}})^k)$	$O(\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$
Lower bound	Non-convex	$\Omega(L/k)$	$\Omega(L/\epsilon)$
	Convex	$\Omega(L/k^2)$	$\Omega(L/\sqrt{\epsilon})$
	Strongly convex	$\Omega((1-\sqrt{\frac{\mu}{L}})^k)$	$\Omega(\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$

NAG and GD has the **same** rate and complexity in non-convex scenarios (Carmon et al., 2020); **GD is optimal and cannot be improved!**

Summary

- Polyak's momentum and Nesterov's momentum can accelerate gradient descent for convex problems

- Nesterov's accelerated gradient method is the optimal first-order algorithm; it reaches the fastest theoretical rate

- Anderson acceleration uses more historical states and performs well in practice; in theorty, it cannot outperform NAG in the worst-case scenario

References I

V. Mai and M. Johansson, "Anderson acceleration of proximal gradient methods," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6620–6629.

Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003, vol. 87.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," *Mathematical Programming*, vol. 184, no. 1-2, pp. 71–120, 2020.