AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

February 4, 2026

Outline

Polyak Heavy Ball Method

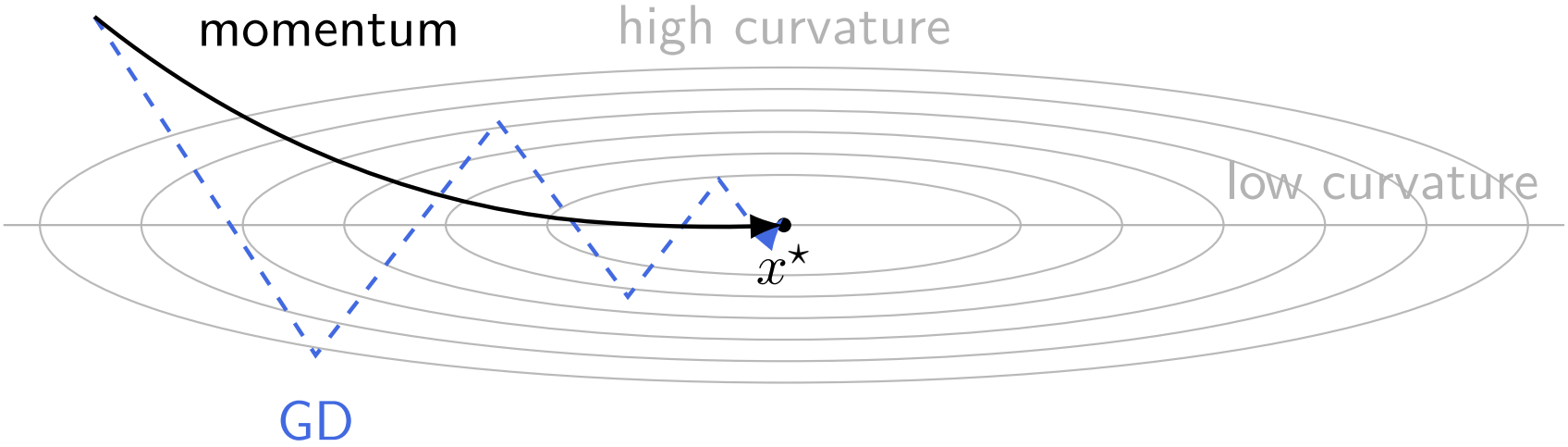
Nesterov Accelerated Gradient Method

Ill-Conditioning Makes GD Zig-Zag

Key picture: when $\kappa = \ell/\mu$ is large, level sets is a long narrow valley.

The gradient points mostly toward the valley walls, not the valley floor.

- Step size is limited by the steep direction: $\alpha \lesssim 1/\ell$.
- Progress along the flat direction is tiny: $\approx \alpha\mu \approx 1/\kappa$ fraction of the remaining distance.



Momentum Helps

Damped running velocity:

$$v_{k+1} = \beta v_k - \alpha \nabla f(x_k), \quad x_{k+1} = x_k + v_{k+1}.$$

- **Steep direction:** GD tends to overshoot, so that component of $\nabla f(x_k)$ flips sign from one step to the next.
  $\Longrightarrow$ velocity v_k cancels these back-and-forth oscillations.
- **Flat direction:** the gradient component changes slowly and keeps a consistent sign for many steps.
  $\Longrightarrow$ v_k accumulates this consistent drift.
- **Intuition:** momentum acts like a filter:

  damp fast oscillations + reinforce slow drift.

Connection to conditioning: GD is limited by ℓ everywhere, while momentum makes it behave as if it can take effectively larger steps along directions governed by μ without blowing up the ℓ-directions.

Polyak Heavy Ball Method

Polyak's idea (1964):

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}). \qquad \text{(HB)}$$

Convergence Rate of HB (Strongly Convex Quadratics)

Suppose $f(x) := \frac{1}{2}(x - x^*)^\top A(x - x^\top)$, where $\lambda_{\min}(A) = \mu$, $\lambda_{\max}(A) = \ell$ and $0 < \mu \leq \ell$. Let $\alpha := \frac{4}{(\sqrt{\mu}+\sqrt{\ell})^2}$ and $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ where $\kappa = \ell/\mu$, then for any $\epsilon > 0$, HB satisfies:

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} \right\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \epsilon \right)^k \left\| \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|$$

for some norm $\|\cdot\|$.

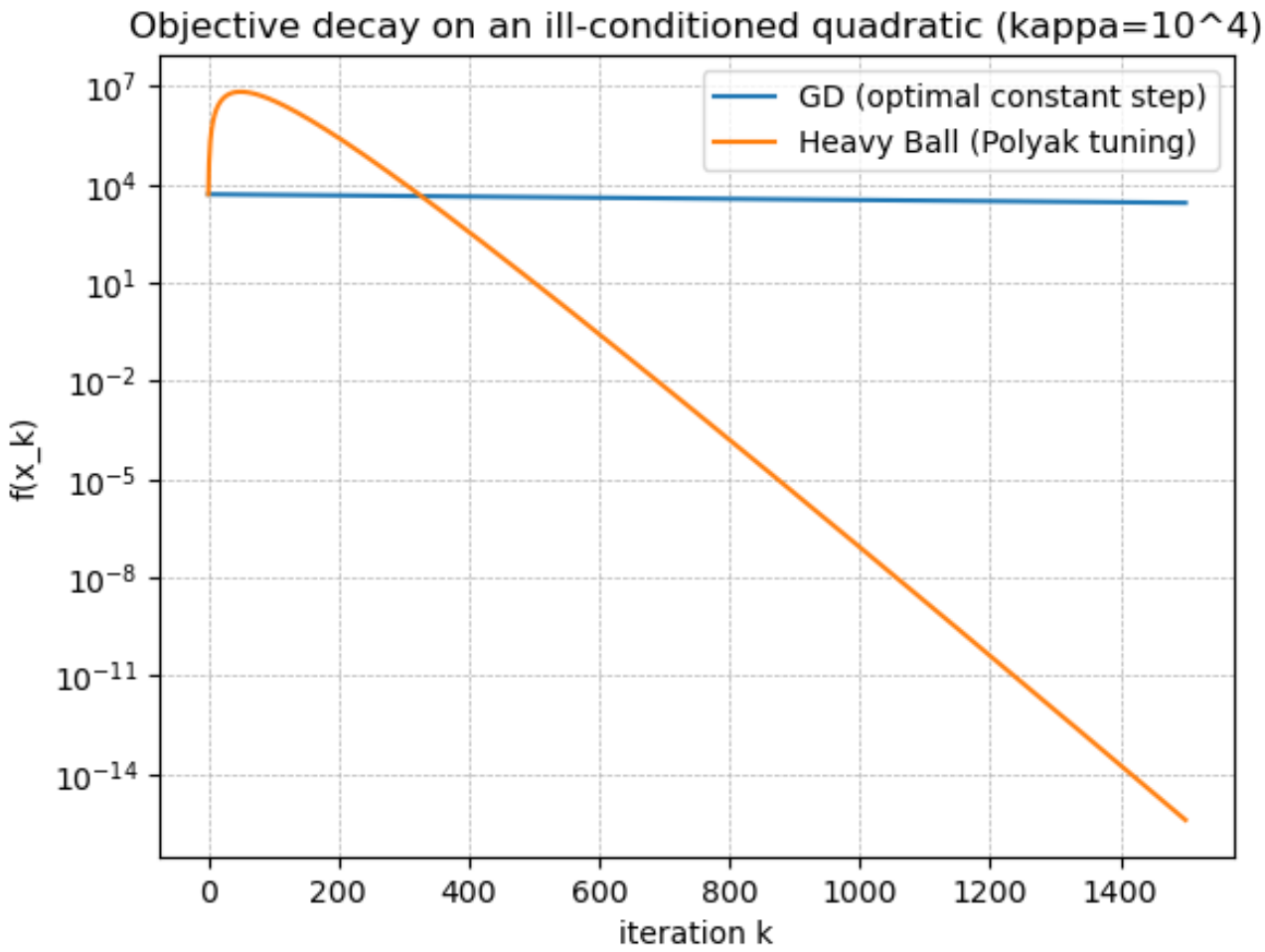
Acceleration: Pick $\epsilon = \frac{1}{1+\sqrt{\kappa}}$, so $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + \epsilon = 1 - \frac{1}{1+\sqrt{\kappa}} \leq e^{-1/(1+\sqrt{\kappa})}$.

Then $e^{-k/(1+\sqrt{\kappa})} \leq \varepsilon \implies k \geq (1+\sqrt{\kappa}) \log \frac{1}{\varepsilon} = O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$.

Numerical Performance

Problem: $f(x) = \frac{1}{2}x^\top Ax$, $A = \mathrm{diag}(\ell, \mu)$ with $\ell = 100$ and $\mu = 0.01$.

Methods: GD with $\eta = \frac{2}{\ell+\mu}$ vs. HB with optimal parameters.



Theory of HB

HB update rule: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$.

Quadratics: $f(x) = \frac{1}{2}(x - x^\star)^\top A(x - x^\star)$, so $\nabla f(x) = A(x - x^\star)$.

Stacked iterates:

$$z_k := \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \in \mathbb{R}^{2n}.$$

Then, for $k \geq 1$, we have $z_{k+1} = M z_k$:

$$\begin{aligned}
\begin{bmatrix} x_{k+1} - x^\star \\ x_k - x^\star \end{bmatrix} &= \begin{bmatrix} x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) - x^\star \\ x_k - x^\star \end{bmatrix} && \text{(a)} \\
&= \begin{bmatrix} x_k - \alpha A(x_k - x^\star) + \beta(x_k - x_{k-1}) - x^\star \\ x_k - x^\star \end{bmatrix} && \text{(b)} \\
&= \underbrace{\begin{bmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{bmatrix}}_{=:M} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} && \text{(c)}
\end{aligned}$$

Theory of HB

Iterating $z_{k+1} = M z_k$ gives

$$z_{k+1} = M^k z_1.$$

By submultiplicativity of the induced matrix norm,

$$\|z_{k+1}\| \leq \|M^k\| \, \|z_1\| \leq \|M\|^k \, \|z_1\|.$$

Key lemma

For any matrix M and any $\varepsilon > 0$, there exists a vector norm $\|\cdot\|$ with induced matrix norm such that

$$\|M\| \leq \rho(M) + \varepsilon.$$

With that norm choice,

$$\|z_{k+1}\| \leq (\rho(M) + \varepsilon)^k \|z_1\|.$$

Theory of HB

Key observation: notice that we can write M as

$$M = \underbrace{\begin{bmatrix} 1+\beta & -\beta \\ 1 & 0 \end{bmatrix}}_{=:C} \otimes I_n \; - \; \alpha \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}}_{=:E} \otimes A.$$

Diagonalization: let $A = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then,

$$S^{-1}MS = C \otimes I_n - \alpha E \otimes \Lambda = \mathrm{diag}(B(\lambda_1), \ldots, B(\lambda_n)),$$

where $S := I_2 \otimes U$ and

$$B(\lambda) := \begin{bmatrix} 1+\beta-\alpha\lambda & -\beta \\ 1 & 0 \end{bmatrix}.$$

Spectrum radius:

$$\rho(M) = \max_{i=1,\ldots,n} \rho(B(\lambda_i)).$$

Theory of HB

The characteristic polynomial of $B(\lambda)$ is

$$r^2 - (1 + \beta - \alpha\lambda)r + \beta = 0,$$

The discriminant is

$$\Delta(\lambda) = (1 + \beta - \alpha\lambda)^2 - 4\beta.$$

Under the parameters $\alpha = \left(\frac{2}{\sqrt{\mu}+\sqrt{\ell}}\right)^2$ and $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$:

$$\Delta(\lambda) = \frac{16(\mu - \lambda)(\ell - \lambda)}{(\sqrt{\ell} + \sqrt{\mu})^4} \leq 0.$$

Hence the two roots are complex conjugates $(z, \bar{z})$ with $z\bar{z} = \beta$. Therefore

$$|z| = \sqrt{z\bar{z}} = \sqrt{\beta} \implies \rho(M) = \rho(B(\lambda)) = \sqrt{\beta}, \quad \forall\, \lambda \in [\mu, \ell].$$

More Discussions

Norm Issues: Can we get rid of this non-standard norm which can be very far from ℓ_2-norm?

- Yes, but you need to sacrifice a little, either make $\rho(M)$ a little smaller than $\sqrt{\beta}$ or obtain

$$k = O\left(\sqrt{\kappa}\left(\log\tfrac{1}{\delta} + \log\sqrt{\kappa} + \log\log\tfrac{1}{\delta}\right)\right)$$

- HB is has nearly optimal global convergence complexity.

Beyond Quadratics: assume f is ℓ-smooth and μ-strongly convex

- In non-quadratic problems, the same parameters can produce divergence for some initializations [Goujaud et al., 2025].
- For carefully chosen (smaller) parameters, HB converges globally with the same rate as GD for convex ℓ-smooth and strongly convex ℓ-smooth [Ghadimi et al., 2015].

Outline

Polyak Heavy Ball Method

Nesterov Accelerated Gradient Method

A Simple Fix to HB

Question: How to get a global convergence with acceleration like in HB beyond strongly convex quadratics?

In 1983, Yurii Nesterov proposed the famous Nesterov's accelerated gradient method (NAG): for μ-strongly convex ℓ-smooth f,

$$\begin{cases} y_k = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_k - \alpha \nabla f(y_k) \end{cases}$$

Compact form:

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \beta(x_k - x_{k-1})) \qquad \text{(NAG)}$$

Instead of directly adding momentum using the previous update, NAG uses the estimated future position to compute the gradient.

Prof. Yurii Nesterov

香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen | SCHOOL OF DATA SCIENCE 數據科學學院

SDS Portal | SEARCH | CUHK-Shenzhen | 简体中文

About SDS | Programmes | Faculty | Students | News & Announcements | School Events | Research | Jobs | Career

HOME

NESTEROV, Yurii

Professor (Fractional)

EDUCATION BACKGROUND

Ph.D. Institute of Control Problems, Moscow (1980-1984)

M.S. Lomonosov Moscow State University (1972-1977)

ACADEMIC AREA

Artificial Intelligence, Computer Science, Machine Learning and Artificial Intelligence, Computer Vision and Computer Graphics, Operations Research, Optimization, Transportation Management, Financial Engineering

RESEARCH FIELD

Theory and Methods of Nonlinear Optimization, Models of Operations Research, Applications of

Convergence of NAG

Assumptions: the function f is μ-strongly convex and ℓ-smooth.

Update: $x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \beta(x_k - x_{k-1}))$.

Convergence of NAG for Strongly Convex Smooth Functions

Let f be μ-strongly convex ℓ-smooth. Let $\alpha = 1/\ell$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ where $\kappa = \ell/\mu$, and $x_0 = x_{-1}$, then

$$f(x_k) - f(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \frac{(\ell+\mu)\|x_0 - x^*\|^2}{2}$$

Acceleration: to satisfy $\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \frac{(\ell+\mu)\|x_0-x^*\|^2}{2} \leq \varepsilon$, it suffices to have $k \geq \sqrt{\kappa} \log\left(\frac{(\ell+\mu)\|x_0-x^*\|^2}{2\varepsilon}\right) = O(\sqrt{\kappa} \log \frac{1}{\varepsilon})$.

Theory of NAG

Update: $x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \beta(x_k - x_{k-1}))$.

Lyapunov Function:

$$\mathcal{E}_k(z) = A_k(f(x_k) - f(z)) + \frac{1}{2}\|w_k - z\|^2,$$

with a suitable sequence A_k and an auxiliary iterate w_k.

If we can show $\mathcal{E}_{k+1}(z) \leq \mathcal{E}_k(z)$ for all z, then:

- Plug in $z = x^*$ to get a rate for $f(x_k) - f^*$.
- The norm term simultaneously yields a uniform control of w_k.

Theory of NAG

Define: $\rho^2 := 1 - \frac{1}{\sqrt{\kappa}} \in (0,1)$; $x^* = 0$ (WLOG).

Lyapunov function:

$$V_k := f(x_k) - f^* + \frac{\ell}{2}\|x_k - \rho^2 x_{k-1}\|^2.$$

One-step contraction: $V_{k+1} \leq \rho^2 V_k$.

Notation: $u_k := \frac{1}{\ell}\nabla f(y_k)$, so $x_{k+1} = y_k - u_k$.

Theory of NAG

Step 1: smoothness at y_k

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{\ell}{2}\|x_{k+1} - y_k\|^2 = f(y_k) - \frac{\ell}{2}\|u_k\|^2.$$

Step 2: expand V_{k+1}

$$V_{k+1} = f(x_{k+1}) - f^* + \frac{\ell}{2}\|x_{k+1} - \rho^2 x_k\|^2.$$

Substitute $x_{k+1} = y_k - u_k$ and regroup terms to get

$$V_{k+1} \leq \rho^2 V_k + R_k,$$

where R_k is an explicit quadratic form in $(x_k - y_k)$, y_k, and u_k.

Step 3: strong convexity makes $R_k \leq 0$

$$f(x_k) - f^* - \langle \nabla f(y_k), x_k - y_k \rangle \geq \frac{m}{2}\|x_k - y_k\|^2, \quad f(y_k) - f^* \geq \frac{m}{2}\|y_k\|^2,$$

and plug $\rho^2 = 1 - \frac{1}{\sqrt{\kappa}}$ to verify $R_k \leq 0$.

Conclusion: $V_{k+1} \leq \rho^2 V_k$.

Theory of NAG

From $V_{k+1} \leq \rho^2 V_k$ we get

$$f(x_k) - f^* \leq V_k \leq \rho^{2k} V_0.$$

With $x_{-1} = x_0$,

$$\begin{aligned} V_0 &= f(x_0) - f^* + \frac{\ell}{2}\|x_0 - \rho^2 x_0\|^2 \\ &\leq \frac{\ell}{2}\|x_0\|^2 + \frac{m}{2}\|x_0\|^2 = \frac{\ell + m}{2}\|x_0\|^2. \end{aligned}$$

Convergence of iterates: by strong convexity,

$$\|x_k - x^*\|^2 \leq \frac{2}{m}(f(x_k) - f^*) \leq \frac{\ell + m}{m}\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \|x_0 - x^*\|^2.$$

Convergence of NAG

Assumptions: the function f is convex and ℓ-smooth.

Update: $x_{k+1} = x_k + \beta_k(x_k - x_{k-1}) - \alpha_k \nabla f(x_k + \beta_k(x_k - x_{k-1}))$.

Convergence of NAG for Convex Smooth Functions

Let f be convex ℓ-smooth. Let $\alpha = 1/\ell$,

$$\beta_k = \frac{t_k - 1}{t_{k+1}}, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

and $t_0 = 1$, $\beta_0 = 0$, then

$$f(x_k) - f(x^*) \leq \frac{2\ell \|x_0 - x^*\|^2}{k^2}$$

Acceleration: recall that for GD, $f(x_k) - f^* = O(\frac{1}{k})$.

Theory of NAG

Auxiliary iterate:

$$z_k := x_k + (t_k - 1)(x_k - x_{k-1}).$$

Lyapunov function:

$$\mathcal{E}_k(z) := t_k^2(f(x_k) - f(z)) + \frac{\ell}{2}\|z_k - z\|^2.$$

Monotonicity: $\mathcal{E}_{k+1}(z) \leq \mathcal{E}_k(z)$ for all z.

Notation: $u_k := \frac{1}{\ell}\nabla f(y_k)$, so $x_{k+1} = y_k - u_k$.

Theory of NAG

Key identities:

- y_k is a convex combination of x_k and z_k:
$$y_k = x_k + \tfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) = \left(1 - \tfrac{1}{t_{k+1}}\right) x_k + \tfrac{1}{t_{k+1}} z_k.$$
- The auxiliary iterate evolves by a GD step:
$$z_{k+1} = z_k - t_{k+1} u_k.$$

One-step descent: for any z, by convexity and smoothness
$$f(x_{k+1}) - f(z) \le \ell \langle u_k, y_k - z \rangle - \tfrac{\ell}{2} \|u_k\|^2.$$

Algebraic manipulation: multiply by t_{k+1}^2 and use
$$t_{k+1}^2 - t_{k+1} = t_k^2, \quad \text{plus } z_{k+1} = z_k - t_{k+1} u_k,$$
to obtain
$$\mathcal{E}_{k+1}(z) - \mathcal{E}_k(z) \le \ell t_k^2 \langle u_k, x_k - z \rangle - t_k^2 (f(x_k) - f(z)).$$
which by convexity implies $\mathcal{E}_{k+1}(z) \le \mathcal{E}_k(z)$.

Theory of NAG

Set $z = x^*$ in $\mathcal{E}_k(z)$ and use $\mathcal{E}_k(x^*) \le \mathcal{E}_0(x^*)$:

$$t_k^2(f(x_k) - f^*) \le \mathcal{E}_k(x^*) \le \mathcal{E}_0(x^*) = \frac{\ell}{2}\|x_0 - x^*\|^2.$$

Lower bound of t_k: using induction, we can show $t_k^2 \ge \frac{k+1}{2}$, thus

$$f(x_k) - f^* \le \frac{\ell}{2t_k^2}\|x_0 - x^*\|^2 \le \frac{2\ell}{k^2}\|x_0 - x^*\|^2.$$

Boundedness of iterates:

$$\|z_k - x^*\|^2 \le \frac{2}{\ell}\mathcal{E}_k(x^*) \le \|x_0 - x^*\|^2,$$

so $\{z_k\}$ and $\{x_k\}$ is bounded.

Convergence of iterates: proved with the assistance of AI recently [Bot et al., 2025, Jang and Ryu, 2025].

More Discussions

Lower bound: Is NAG optimal in terms of convergence rate?

- optimal in terms of the rate.
- not optimal in terms of the constant.

Beyond smooth convex: Do we have acceleration if f is only ℓ-smooth? only C^2 convex?

- nonconvex ℓ-smooth: still active research
  [Gupta and Wojtowytsch, 2024]
- C^2 convex: what is the rate on x^4? what about x^{100}?

Wrap-up and Next Week

Week 3: Key takeaways

- Ill-conditioning slows GD.
- Heavy Ball (HB): strongly convex quadratics, accelerated factor $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \approx 1 - \frac{2}{\sqrt{\kappa}}$.
- Nesterov Acceleration (NAG): a Lyapunov function yields clean worst-case guarantees:
  - convex ℓ-smooth: $f(x_k) - f^* = O(1/k^2)$;
  - m-strongly convex ℓ-smooth: $f(x_k) - f^* \leq (1 - 1/\sqrt{\kappa})^k \cdot C_0$.

Thank you for your attendance!

Questions?

Next week (Week 4): Subgradient Methods and Mirror Descent

References


Bot, R. I., Fadili, J., and Nguyen, D.-K. (2025).
The iterates of nesterov's accelerated algorithm converge in the critical regimes.
arXiv preprint arXiv:2510.22715.


Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M. (2015).
Global convergence of the heavy-ball method for convex optimization.
In *2015 European control conference (ECC)*, pages 310–315. IEEE.


Goujaud, B., Taylor, A., and Dieuleveut, A. (2025).
Provable non-accelerations of the heavy-ball method: B. goujaud et al.
Mathematical Programming, pages 1–59.


Gupta, K. and Wojtowytsch, S. (2024).
Nesterov acceleration in benignly non-convex landscapes.
arXiv preprint arXiv:2410.08395.


Jang, U. and Ryu, E. K. (2025).
Point convergence of nesterov's accelerated gradient method: An ai-assisted proof.
arXiv preprint arXiv:2510.23513.