CHAPTER 2. ACCELERATED GRADIENT DESCENT

Yutong He Kun Yuan

September 26, 2023

1 Problem formulation

This chapter considers the following unconstrained convex problem

$$\min_{x\in\mathbb{R}^d} \quad f(x) \tag{1}$$

where $f(x)$ is a differentiable convex objective function.

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal solution to problem (1).
- Let $f^\star := \min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal function value.

2 Polyak's momentum gradient descent

2.1 Polyak's algorithm

Gradient descent with Polyak's momentum (or heavy-ball) method iterates as follows:

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

where $\gamma > 0$ is the learning rate (or step-size), and $\beta \in (0, 1)$ is the momentum parameter.

2.2 Convergence analysis of Polyak's method

Theorem 2.1 (Convergence rate of Polyak's momentum method in quadratic scenario)**.** Let $f(x) = \frac{1}{2}x^\top Ax$, where $x \in \mathbb{R}^2$, $A = \begin{pmatrix} L & 0 \\ 0 & \mu \end{pmatrix}$ with $L > \mu > 0$, if $\gamma = 1/L$ and $\beta = L/(\sqrt{L}+\sqrt{\mu})^2$, the Polyak's momentum method (with initialization $x_{-1} = x_0$) converges as

$$\|x_k\|_2 \le c(L,\mu)\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \|x_0\|_2, \quad k = 1, 2, \cdots, \tag{3}$$

where $c(L,\mu) < +\infty$ is a positive constant that only depends on L and μ.

Proof. Denote $y_k := (x_k^\top, x_{k-1}^\top)^\top$, the Polyak's iteration in (2) can be rewritten as

$$y_{k+1} = By_k, \tag{4}$$

where

$$B = \begin{pmatrix} 1+\beta-\gamma L & 0 & -\beta & 0 \\ 0 & 1+\beta-\gamma\mu & 0 & -\beta \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Let

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1+\beta-\gamma L & -\beta \\ 1 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1+\beta-\gamma\mu & -\beta \\ 1 & 0 \end{pmatrix},$$

it holds that $P = P^{-1} = P^T$ and

$$B = P \underbrace{\begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}}_{\tilde{B}} P^{-1},$$

When $(1-\sqrt{\beta})^2/\mu < \gamma < (1+\sqrt{\beta})^2/L$, both B_1 and B_2 have a pair of conjugate complex eigenvalues with magnitude $\sqrt{\beta}$. Consider the eigen decomposition as follows:

$$B_1 = P_1\Lambda_1 P_1^{-1}, \quad B_2 = P_2\Lambda_2 P_2^{-1},$$

where

$$P_1 := \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ 1 & 1 \end{pmatrix}, \quad \Lambda_1 := \begin{pmatrix} \lambda_{11} & 0 \\ 0 & \lambda_{12} \end{pmatrix}, \quad P_2 := \begin{pmatrix} \lambda_{21} & \lambda_{22} \\ 1 & 1 \end{pmatrix}, \quad \Lambda_2 := \begin{pmatrix} \lambda_{21} & 0 \\ 0 & \lambda_{22} \end{pmatrix},$$

and

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

Denote

$$z_k := \underbrace{\begin{pmatrix} P_1^{-1} & 0 \\ 0 & P_2^{-1} \end{pmatrix}}_{\tilde{P}^{-1}} P^{-1} y_k,$$

then (4) can be rewritten as

$$z_{k+1} = \underbrace{\begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}}_{\tilde{\Lambda}} z_k.$$

Note that $|\lambda_{11}| = |\lambda_{12}| = |\lambda_{21}| = |\lambda_{22}| = \sqrt{\beta}$, we obtain $\|z_k\|_2 \le \beta^{k/2}\|z_0\|_2$, which implies

$$\begin{aligned}\|x_k\|_2 \le & \|y_k\|_2 = \|P\tilde{P}z_k\|_2 = \|\tilde{P}z_k\|_2 \le \|\tilde{P}\|_2\|z_k\|_2 \le \|\tilde{P}\|_2\beta^{k/2}\|z_0\|_2 \\ = & \|\tilde{P}\|_2\beta^{k/2}\|(\tilde{P})^{-1}P^{-1}y_0\|_2 \le \|\tilde{P}\|_2\beta^{k/2}\|\tilde{P}^{-1}\|_2\|P^{-1}y_0\|_2 \\ = & \|\tilde{P}\|_2\beta^{k/2}\|\tilde{P}^{-1}\|_2(\sqrt{2}\|x_0\|_2).\end{aligned}$$

It can be verified that

$$\begin{aligned}\|\tilde{P}\|_2^2 = & 1+\beta+\sqrt{(1-\beta)^2+\max\{|1+\beta-\gamma L|, |1+\beta-\gamma\mu|\}^2}, \\ \|\tilde{P}^{-1}\|_2^2 = & \max\left\{\frac{1+\beta+\sqrt{(1-\beta)^2+(1+\beta-\gamma L)^2}}{4\beta-(1+\beta-\gamma L)^2}, \frac{1+\beta+\sqrt{(1-\beta)^2+(1+\beta-\gamma\mu)^2}}{4\beta-(1+\beta-\gamma\mu)^2}\right\}\end{aligned}$$

Specifically, if we set the parameters as

$$\beta = \frac{L}{(\sqrt{L}+\sqrt{\mu})^2}, \quad \gamma = \frac{1}{L},$$

there exists positive constant $c(L,\mu) < +\infty$ such that

$$\|x_k\|_2 \le c(L,\mu)\left(1-\frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \|x_0\|_2, \quad k=1,2,\cdots.$$

□

Remark. When the condition number $\kappa := L/\mu \gg 1$, the term $1-\frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ can be approximated by $1-\frac{1}{\sqrt{\kappa}}$.

3 Nesterov's momentum gradient descent

3.1 Nesterov's algorithm

Gradient descent with Nesterov's momentum (or Nesterov's accelerated gradient) method iterates as follows:

$$y_{k-1} = x_{k-1} + \beta(x_{k-1} - x_{k-2}), \tag{5}$$
$$x_k = y_{k-1} - \gamma \nabla f(y_{k-1}), \tag{6}$$

where $\gamma > 0$ is the learning rate (or step-size), and $\beta \in (0, 1)$ is the momentum parameter.

Remark. Although we do not specify subscripts for γ and β, they do not have to be fixed as constants and may depend on the step k.

3.2 Convergence analysis of Nesterov's method

Theorem 3.1 (Convergence rate of Nesterov's momentum method in generally–convex scenario)**.** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth. If $\gamma = 1/L$ and we specify $\beta = (k-2)/(k+1)$ in (5), the Nesterov's momentum method (with initialization $x_{-1} = x_0$) converges as

$$f(x_k) - f^\star \leq \frac{2L\|x_0 - x^\star\|_2^2}{(k+1)^2}. \tag{7}$$

Proof. Defining $\theta_k = 2/(k+1)$, $v_0 = x_0$, the update rule of Nesterov's momentum method can be rewritten as

$$\begin{aligned} y_{k-1} &= (1-\theta_k)x_{k-1} + \theta_k v_{k-1}, \\ x_k &= y_{k-1} - \frac{1}{L}\nabla f(y_{k-1}), \\ v_k &= x_{k-1} + \frac{1}{\theta_k}(x_k - x_{k-1}). \end{aligned}$$

By L-smooth property, we have

$$\begin{aligned} f(x_k) &\leq f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1}\rangle + \frac{L}{2}\|x_k - y_{k-1}\|_2^2 \\ &= f(y_{k-1}) - \frac{L}{2}\|x_k - y_{k-1}\|_2^2. \end{aligned} \tag{8}$$

By convex property, we have for any $x \in \mathbb{R}^d$ that

$$\begin{aligned} f(y_{k-1}) - f(x) &\leq \langle \nabla f(y_{k-1}), y_{k-1} - x\rangle \\ &= L\langle y_{k-1} - x_k, y_{k-1} - x\rangle \\ &= \frac{L}{2}\left(\|y_{k-1} - x_k\|_2^2 + \|y_{k-1} - x\|_2^2 - \|x_k - x\|_2^2\right). \end{aligned} \tag{9}$$

Combining (8)(9), we obtain

$$f(x_k) - f(x) \leq \frac{L}{2}\left(\|y_{k-1} - x\|_2^2 - \|x_k - x\|_2^2\right), \quad \forall\, x \in \mathbb{R}^d. \tag{10}$$

Consequently,

$$\begin{aligned}
&f(x_k) - f^\star - (1-\theta_k)(f(x_{k-1}) - f^\star) \\
=&f(x_k) - (1-\theta_k)f(x_{k-1}) - \theta_k f^\star \\
\leq&f(x_k) - f((1-\theta_k)x_{k-1} + \theta_k x^\star) \\
\overset{(10)}{\leq}&\frac{L}{2}\left(\|y_{k-1} - (1-\theta_k)x_{k-1} - \theta_k x^\star\|_2^2 - \|x_k - (1-\theta_k)x_{k-1} - \theta_k x^\star\|_2^2\right) \\
=&\frac{L\theta_k^2}{2}\left(\|v_{k-1} - x^\star\|_2^2 - \|v_k - x^\star\|_2^2\right).
\end{aligned} \tag{11}$$

By rearranging (11) we obtain

$$\begin{aligned}
\frac{1}{L\theta_k^2}(f(x_k) - f^\star) + \frac{1}{2}\|v_k - x^\star\|_2^2 \leq&\frac{1-\theta_k}{L\theta_k^2}(f(x_{k-1}) - f^\star) + \frac{1}{2}\|v_{k-1} - x^\star\|_2^2, \quad \forall\, k \geq 1, \\
\leq&\frac{1}{L\theta_{k-1}^2}(f(x_{k-1}) - f^\star) + \frac{1}{2}\|v_{k-1} - x^\star\|_2^2, \quad \forall\, k \geq 2.
\end{aligned} \tag{12}$$

(12) implies that

$$\begin{aligned}
\frac{1}{L\theta_k^2}(f(x_k) - f^\star) + \frac{1}{2}\|v_k - x^\star\|_2^2 \leq&\frac{1}{L\theta_1^2}(f(x_1) - f^\star) + \frac{1}{2}\|v_1 - x^\star\|_2^2 \\
\leq&\frac{1-\theta_1}{L\theta_1^2}(f(x_0) - f^\star) + \frac{1}{2}\|v_0 - x^\star\|_2^2 \\
=&\frac{1}{2}\|x_0 - x^\star\|_2^2,
\end{aligned}$$

which further implies

$$f(x_k) - f^\star \leq \frac{L\theta_k^2}{2}\|x_0 - x^\star\|_2^2 = \frac{2L\|x_0 - x^\star\|_2^2}{(k+1)^2} = \mathcal{O}\left(\frac{L}{k^2}\right).$$

□

Theorem 3.2 (Convergence rate of Nesterov's momentum method in strongly-convex scenario)**.** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth. If $\gamma = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$, the Nesterov's momentum method (with initialization $x_{-1} = x_0$) converges as

$$f(x_k) - f^\star \leq \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)^k \left(f(x_0) - f^\star + \frac{\mu}{2}\|x_0 - x^\star\|_2^2\right) \tag{13}$$

Proof. It can be verified that when choosing $\gamma = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$,

Nesterov's momentum method is equivalent to the following update rules:

$$y_{k-1} = \frac{\sqrt{L}}{\sqrt{L}+\sqrt{\mu}} x_{k-1} + \frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} v_{k-1}, \tag{14}$$

$$x_k = y_{k-1} - \frac{1}{L}\nabla f(y_{k-1}), \tag{15}$$

$$v_k = \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) v_{k-1} + \frac{\sqrt{L}}{\sqrt{\mu}} x_k + \left(\frac{\sqrt{\mu}}{\sqrt{L}} - \frac{\sqrt{L}}{\sqrt{\mu}}\right) y_{k-1}, \tag{16}$$

with initialization $v_0 = x_0$. By the new update rules, we have

$$\begin{aligned}
&\|v_k - x^\star\|_2^2 \\
=&\left\| \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)(v_{k-1} - x^\star) + \frac{\sqrt{\mu}}{\sqrt{L}}(y_{k-1} - x^\star) - \frac{1}{\sqrt{\mu L}}\nabla f(y_{k-1}) \right\|_2^2 \\
=&\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|v_{k-1} - x^\star\|_2^2 + \frac{\sqrt{\mu}}{\sqrt{L}}\|y_{k-1} - x^\star\|_2^2 - \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\frac{\sqrt{\mu}}{\sqrt{L}}\|v_{k-1} - y_{k-1}\|_2^2 \\
&+ \frac{1}{\mu L}\|\nabla f(y_{k-1})\|_2^2 - \frac{2}{\sqrt{\mu L}}\left\langle \nabla f(y_{k-1}), \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) v_{k-1} + \frac{\sqrt{\mu}}{\sqrt{L}} y_{k-1} - x^\star \right\rangle \\
=&\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|v_{k-1} - x^\star\|_2^2 + \frac{\sqrt{\mu}}{\sqrt{L}}\|y_{k-1} - x^\star\|_2^2 - \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\frac{\sqrt{\mu}}{\sqrt{L}}\|v_{k-1} - y_{k-1}\|_2^2 \\
&+ \frac{1}{\mu L}\|\nabla f(y_{k-1})\|_2^2 - \frac{2}{\sqrt{\mu L}}\left\langle \nabla f(y_{k-1}), \frac{\sqrt{L}}{\sqrt{\mu}} y_{k-1} - \left(\frac{\sqrt{L}}{\sqrt{\mu}} - 1\right) x_{k-1} - x^\star \right\rangle.
\end{aligned} \tag{17}$$

Rearranging (17) we have

$$\begin{aligned}
&\left\langle \nabla f(y_{k-1}), \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) x_{k-1} + \frac{\sqrt{\mu}}{\sqrt{L}} x^\star - y_{k-1} \right\rangle \\
=&\frac{\mu}{2}\|v_k - x^\star\|_2^2 - \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|v_{k-1} - x^\star\|_2^2 - \frac{\mu\sqrt{\mu}}{2\sqrt{L}}\|y_{k-1} - x^\star\|_2^2 - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2 \\
&+ \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\frac{\sqrt{\mu}}{\sqrt{L}}\|v_{k-1} - y_{k-1}\|_2^2.
\end{aligned} \tag{18}$$

By L-smooth and μ-strongly-convex property, we have

$$\begin{aligned}
f(x_k) \le& f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1}\rangle + \frac{L}{2}\|x_k - y_{k-1}\|_2^2 \\
=& f(y_{k-1}) - \frac{1}{L}\|\nabla f(y_{k-1})\|_2^2 + \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2 \\
=& f(y_{k-1}) - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2 \\
\le& f(u) - \langle \nabla f(y_{k-1}), u - y_{k-1}\rangle - \frac{\mu}{2}\|u - y_{k-1}\|_2^2 - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2, \quad \forall\, u \in \mathbb{R}^d.
\end{aligned} \tag{19}$$

Consider $(1 - \sqrt{\mu}/\sqrt{L}) \times$ (19) (where $u = x_{k-1}$) + $(\sqrt{\mu}/\sqrt{L}) \times$ (19) (where $u = x^\star$), we

obtain

$$
\begin{aligned}
f(x_k) - f^\star \leq & \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)(f(x_{k-1}) - f^\star) - \left\langle \nabla f(y_{k-1}), \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) x_{k-1} + \frac{\sqrt{\mu}}{\sqrt{L}} x^\star - y_{k-1} \right\rangle \\
& - \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|x_{k-1} - y_{k-1}\|_2^2 - \frac{\mu\sqrt{\mu}}{2\sqrt{L}}\|x^\star - y_{k-1}\|_2^2 - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2 \\
\overset{(18)}{=} & \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)(f(x_{k-1}) - f^\star) - \frac{\mu}{2}\|v_k - x^\star\|_2^2 + \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|v_{k-1} - x^\star\|_2^2 \\
& - \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\frac{\sqrt{\mu}}{\sqrt{L}}\|v_{k-1} - y_{k-1}\|_2^2 - \frac{\mu}{2}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\|x_{k-1} - y_{k-1}\|_2^2. \qquad (20)
\end{aligned}
$$

Denoting $\Phi_k = f(x_k) - f^\star + \frac{\mu}{2}\|v_k - x^\star\|_2^2$, (20) implies

$$
\Phi_k \leq \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)\Phi_{k-1},
$$

and further

$$
f(x_k) - f^\star \leq \Phi_k \leq \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)^k \Phi_0 = \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)^k \left(f(x_0) - f^\star + \frac{\mu}{2}\|x_0 - x^\star\|_2^2\right).
$$

□

4 Anderson's acceleration method

Let $g(x) = x - \gamma\nabla f(x)$ ($\gamma > 0$), the fixed points of g are zero points of ∇f. Anderson acceleration method aims to minimize $\|g(x) - x\|_2$, which iterates as follows:

$$
\begin{aligned}
m_k =& \min\{m, k\}, \qquad (21)\\
R_k =& (r_k, r_{k-1}, \cdots, r_{k-m_k}), \text{where } r_i = g(x_i) - x_i, \qquad (22)\\
\alpha_k =& \arg\min_{\alpha^\top \mathbf{1} = 1} \|R_k \alpha\|_2, \qquad (23)\\
x_{k+1} =& \sum_{m=0}^{m_k} \alpha_i^k g(x_{k-i}), \qquad (24)
\end{aligned}
$$

with initialization $x_1 = g(x_0)$ and m an integer typically set between 1 and 10 in practice.

Proposition 4.1. If $R_k^\top R_k$ is non-singular, the iteration step (23) is equivalent to

$$
\alpha_k = \frac{(R_k^\top R_k)^{-1}\mathbf{1}}{\mathbf{1}^\top (R_k^\top R_k)^{-1}\mathbf{1}}. \qquad (25)
$$

Convergence rate. Suppose f is μ-strongly convex and L-smooth, and assume $\|\alpha_k\|_1$ be bounded by a constant M_α, if we let $\gamma \in (0, 1/L]$, Anderson's acceleration converges to $x^\star$ with averaged rate asymptotically the same order as that of GD's. However, in practice, Anderson's acceleration can be even faster than Nesterov's acceleration. Reasons behind the clear gap between theoretical analysis and practical performance remains an open question. We refer the specific convergence results to Theorem 1 of [1]

5 Optimal convergence rate

In this section we present the lower complexity bounds for first-order algorithms to solve smooth convex objective functions. Compared with the convergence results of Nesterov's method, we come to the conclusion of the optimal convergence rates.

Theorem 5.1 (Convergence lower bound in the generally-convex scenario)**.** For any first-order algorithm satisfying

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_{k-1})\},$$

there always exists some convex and L-smooth function f such that

$$f(x_k) - f^\star \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(k+1)^2}. \tag{26}$$

Proof. We refer the proof to Theorem 2.1.7 in [2]. □

Theorem 5.2 (Convergence lower bound in the strongly-convex scenario)**.** For any first-order algorithm satisfying

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_{k-1})\},$$

there always exists some μ-strongly convex and L-smooth function f such that

$$f(x_k) - f^\star \geq \frac{\mu}{2}\left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \|x_0 - x^\star\|_2^2. \tag{27}$$

Proof. We refer the proof to Theorem 2.1.13 in [2]. □

Comparison of convergence complexity. We list the convergence complexity (number of iterations) for reaching an ϵ-optimal solution such that $f(x_k) - f^\star \leq \epsilon$:

Algorithm	generally-convex	strongly-convex
GD	$\mathcal{O}\left(\frac{L}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\ln\left(\frac{1}{\epsilon}\right)\right)$
Nesterov	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$
Lower Bound	$\Omega\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\Omega}\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$

where $\tilde{\mathcal{O}}$, $\tilde{\Omega}$ hides logarithm terms of L, μ.

6 Simulation

In this section we examine GD (gradient descent), Nesterov's and Polyak's momentum gradient descent algorithm with a quadratic objective function. Define $f(x) := \frac{1}{2}x^\top Ax$

where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 20 \end{pmatrix},$$

it can be verified that f is 1-strongly convex and 20-smooth.

Figure 1 shows iteration points of each algorithm starting at point $(10, 1)^\top$ within 20 steps, where the step-sizes are chosen as $\gamma = 0.1$ for GD and $\gamma = 0.05$ for both momentum methods, and the momentum parameter is set as $\beta = \frac{20}{(\sqrt{20}+1)^2}$. For Anderson acceleration, we set $\gamma = 0.05$ and $m = 5$, which nearly achieves the optimal point within 3 iteration steps. Figure 2 shows loss curves of the four methods within 200 steps.

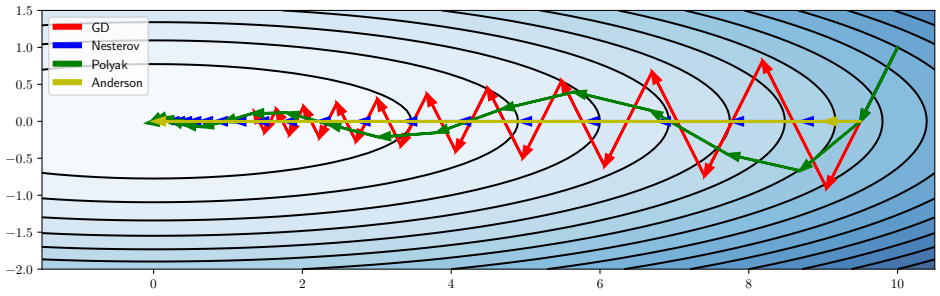

Figure 1: Iteration points

7 Exercises

1. Consider quadratic function $f(x) := \frac{1}{2}(x - x^\star)Q(x - x^\star)$, where $Q = Q^\top$ and $LI \succeq Q \succeq \mu I$. Prove that

   (i) Nesterov iterations (5)(6) are equivalent to the following iterations:

   $$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1},$$

   $$\text{where } \boldsymbol{x}_k := \begin{pmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{pmatrix} \text{ and } \boldsymbol{A} = \begin{pmatrix} (1+\beta)(I - \gamma Q) & -\beta(I - \gamma Q) \\ I & 0 \end{pmatrix}.$$

   (ii) If we choose $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ and $\gamma = \frac{1}{L}$, it holds that $\|\boldsymbol{A}\|_2 \geq 1$.

2. Prove the equivalence of iterations (5)(6) and (14)(15)(16).

3. Prove proposition 4.1 by considering the following constrained convex optimization problem:

   $$\min_{\alpha \in \mathbb{R}^{m_k}} \frac{1}{2}\|R_k \alpha\|_2^2, \quad \text{s.t. } \alpha^\top \mathbf{1} = 1.$$

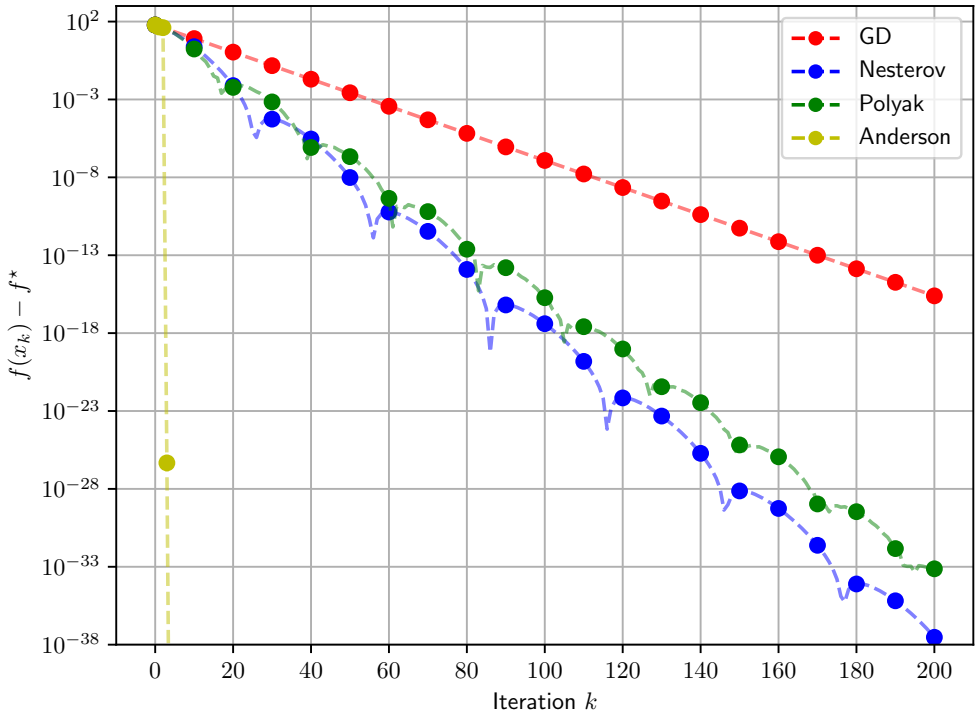

Figure 2: Loss curves

References

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

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