

NJUAI 南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

Lecture 3. Gradient Descent Method

Advanced Optimization (Fall 2026)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

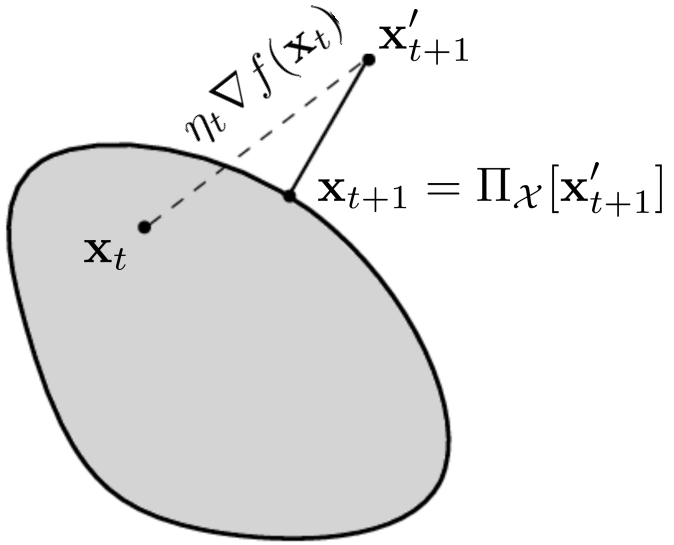
Outline

- Gradient Descent
- Convex and Lipschitz
  - Polyak Step Size
  - Convergence without Optimal Value
  - Optimal Time-Varying Step Sizes
- Strongly Convex and Lipschitz

Gradient Descent

- GD Template:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)\right]$$



- $\mathbf{x}_1$ can be an arbitrary point inside the domain.
- $\eta_t > 0$ is the potentially time-varying ***step size*** (or called ***learning rate***).
- Projection $\Pi_{\mathcal{X}}[\mathbf{y}] = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$ ensures the feasibility.

GD Convergence

- Lipschitz (non-smooth) optimization
  - Convex
  - Strongly convex
- Smooth optimization
  - Convex
  - Strongly convex

This lecture will focus on GD analysis for ***Lipschitz*** functions, and next lecture will discuss ***smooth*** functions.

The First Gradient Descent Lemma

Lemma 1. *Suppose that f is proper, closed and convex; the feasible domain $\mathcal{X}$ is nonempty, closed and convex. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method, $\mathcal{X}^\star$ be the optimal set of the optimization problem and $f^\star$ be the optimal value. Then for any $\mathbf{x}^\star \in \mathcal{X}^\star$ and $t \geq 0$,*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2.$$

Proof:

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &= \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \quad \text{(GD)} \\
&\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)} \\
&= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^\star) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\
&\quad \text{(convexity: } f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle\text{)}
\end{aligned}$$

□

Part 1. Polyak Step Size

- Polyak Step Size
- Convergence
- Convergence Rate

Polyak Step Size

- GD method satisfies the following inequality:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \le \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2$$

$$h(\eta) \triangleq -2\eta(f(\mathbf{x}_t) - f^\star) + \eta^2\|\nabla f(\mathbf{x}_t)\|^2$$

A natural idea:

minimizing the right-hand side of the inequality

$$\Rightarrow \quad \eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$$

assume known $f^\star$ for a moment

Polyak Step Size

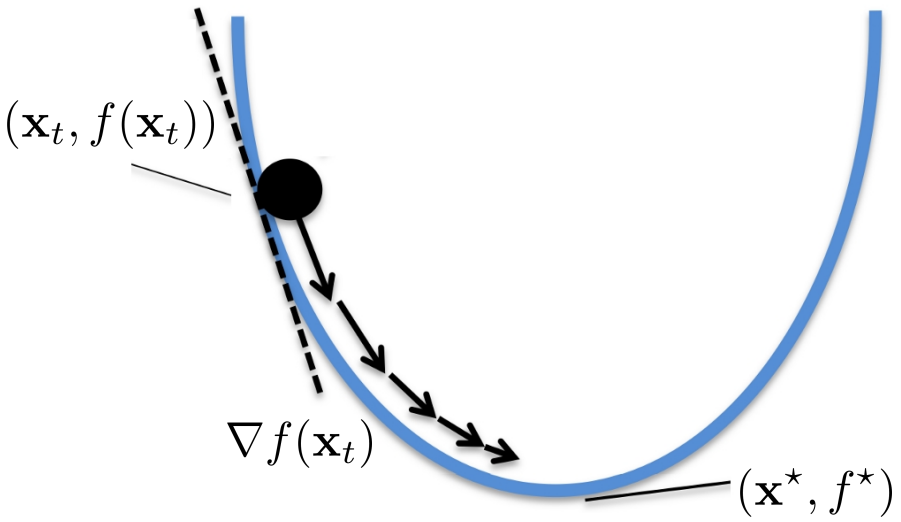
$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \underbrace{2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2}_{h(\eta) \triangleq -2\eta(f(\mathbf{x}_t) - f^\star) + \eta^2\|\nabla f(\mathbf{x}_t)\|^2}$$

Cornercase: when $\nabla f(\mathbf{x}_t) = \mathbf{0}$

$\Rightarrow$ actually a good news owing to convexity, $\nabla f(\mathbf{x}_t) = \mathbf{0}$ implies *optimality*

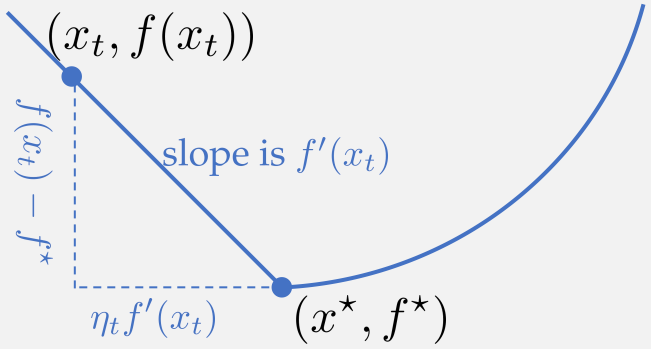
Polyak step size: $$\eta_t = \begin{cases} \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$$

Polyak Step Size: A Geometric View



Q: if we have known $f^\star$ already, how would we set $\mathbf{x}_{t+1}$?

Geometric way to "optimize" (consider the 1-dim function)



Geometrically, the best way of iterates

$$x_{t+1} = x_t - \eta_t f'(x_t)$$

would satisfy that (given known $f^\star$)

$$\eta_t f'(x_t) \cdot f'(x_t) = f(x_t) - f^\star$$

$$\Rightarrow \quad \eta_t = \frac{f(x_t) - f^\star}{f'(x_t)^2}$$

(Unconstrained) GD with Polyak Step Size

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t), \quad \eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$$

Convergence

- With Polyak step size, we obtain the convergence results.

Theorem 1. *Under the same assumptions with Lemma 1, assume the gradient of f is bounded by G, i.e., $\|\nabla f(\cdot)\| \leq G$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method with Polyak step size and $f^\star$ be the optimal value. Then,*

(i) $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2$.

(ii) $f(\mathbf{x}_t) \to f^\star$ *as* $t \to \infty$.

Note: recall that ***bounded gradients*** condition implies ***Lipschitz continuity***.

Convergence

Proof: $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2$

(the first GD lemma)

- **Case 1:** $\nabla f(\mathbf{x}_t) = \mathbf{0}$. By convexity, $f(\mathbf{x}_t) = f^\star \Rightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 = \|\mathbf{x}_t - \mathbf{x}^\star\|^2$.
- **Case 2:** $\nabla f(\mathbf{x}_t) \neq \mathbf{0}$. Polyak's step size $\eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$

$$\Longrightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{(f(\mathbf{x}_t) - f^\star)^2}{\|\nabla f(\mathbf{x}_t)\|^2} \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

(*i*) is proved.

Convergence

Proof: we can simply focus on the case of $\nabla f(\mathbf{x}_t) \neq \mathbf{0}$

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{(f(\mathbf{x}_t) - f^\star)^2}{\|\nabla f(\mathbf{x}_t)\|^2} \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{(f(\mathbf{x}_t) - f^\star)^2}{G^2}$$

$$(\|\nabla f(\cdot)\| \leq G)$$

$$\Longrightarrow \quad \frac{1}{G^2}\sum_{t=1}^{T}(f(\mathbf{x}_t) - f^\star)^2 \leq \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{T+1} - \mathbf{x}^\star\|^2$$

$$\Longrightarrow \quad \sum_{t=1}^{T}(f(\mathbf{x}_t) - f^\star)^2 \leq G^2\|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

Infinite summation is bounded by constants → **convergent** series.

(*ii*) is proved. □

Convergence Rate

Polyak step size: $\eta_t = \begin{cases} \frac{f(\mathbf{x}_t)-f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$

- Theorem 1 proves the convergence, and now we give the ***rate***.

Theorem 2. *Assume the gradient of f is bounded by G, i.e., $\|\nabla f(\cdot)\| \leq G$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the GD method with Polyak step size and $f^\star$ be the optimal value. Define $\bar{\mathbf{x}}_T = \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$, we have*

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{G\|\mathbf{x}_1 - \mathbf{x}^\star\|}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

best-iterated guarantee

Proof:

$$\left.\begin{aligned} & f(\bar{\mathbf{x}}_T) = \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t) \leq f(\mathbf{x}_t) \\ & \sum_{t=1}^T (f(\mathbf{x}_t) - f^\star)^2 \leq G^2\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 \end{aligned}\right\} T(f(\bar{\mathbf{x}}_T) - f^\star)^2 \leq G^2\|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

□

Optimality

- It can be proved that $O\left(1/\sqrt{T}\right)$ is minimax optimal for first-order methods when optimizing **convex and Lipschitz functions**.

Theorem 3 (Bubeck, 2015, Theorem 3.13). *Let $t \leq n$ and $L, R > 0$. We consider the class of first-order (black-box) procedures satisfying $\mathbf{x}_1 = \mathbf{0}$, and for any $t \geq 1$, $\mathbf{x}_{t+1} \in \mathrm{Span}(\mathbf{g}_1, \ldots, \mathbf{g}_t)$, where $\mathbf{g}_s = \nabla f(\mathbf{x}_s)$ is the gradient (or subgradient) queried from the oracle at $\mathbf{x}_s$.*

Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be the canonical basis of $\mathbb{R}^n$, and denote by $B_2(R) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq R\}$. Then there exists a convex and L-Lipschitz function $f : B_2(R) \to \mathbb{R}$ such that for any black-box procedure satisfying the above,

$$\min_{1 \leq s \leq t} f(\mathbf{x}_s) - \min_{\mathbf{x} \in B_2(R)} f(\mathbf{x}) \geq \frac{RL}{2(1+\sqrt{t})}.$$

Why Polyak Step Size Matters

- Although Polyak step size [Polyak, 1969] looks simple and even requires knowing $f^\star$, it has a profound ***enlightening*** value.
- It established the foundation for the study of step-size tuning and convergence-rate analysis in modern optimization.

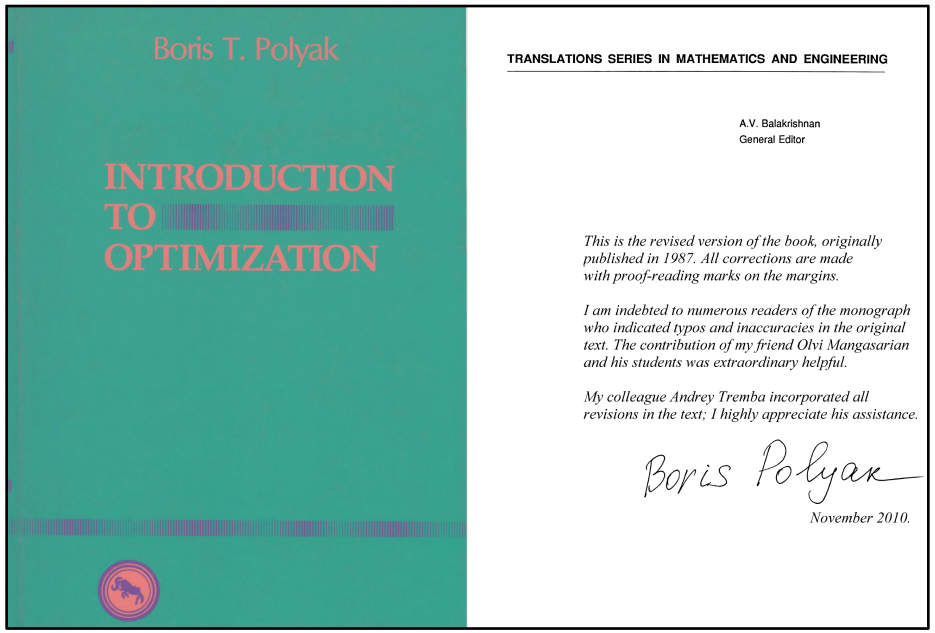
- Technically,
  - Optimal theoretical convergence with minimal assumptions
  - Connect the step size tuning with the gradient norms
  - Lays the conceptual ground for future developments: adaptive tuning, universal optimization, parameter-free optimization, etc.

Polyak Step Size

Polyak step size:

$$\eta_t = \begin{cases} \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$$







Boris T. Polyak
1935-2023

Introduction to optimization

Boris T. Polyak

Optimization Software, Inc., 1987

Part 2. Convergence without Optimal Value

- The Second Gradient Descent Lemma
- Convergent Step Size
- Convergence without Optimal Value

Step Size without Optimal Value

- Note that Polyak step size requires the optimal value $f^\star$

Polyak step size: $$\eta_t = \begin{cases} \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$$

assume known $f^\star$ for a moment

From now on, we try to design step sizes ***without*** the optimal value $f^\star$.

The Second Gradient Descent Lemma

- A second version of gradient descent lemma.

Lemma 2. *Under the same assumptions as Theorem 1. Let* $\{\mathbf{x}_t\}_{t=1}^T$ *be the sequence generated by GD. Then we have*

$$\sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f^\star) \leq \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2.$$

Proof: The statement can be derived directly from the gradient descent lemma:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^\star) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

$$\Rightarrow \quad \eta_t (f(\mathbf{x}_t) - f^\star) \leq \frac{1}{2}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) + \frac{1}{2}\eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \qquad \square$$

Convergence Result

- GD lemma implies the following convergence result.

Lemma 3. *Under the same assumptions as Theorem 1. Let* $\{\mathbf{x}_t\}_{t=1}^T$ *be the sequence generated by GD. Define*

- *best iterate:* $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$;
- *average iterate:* $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$.

Then, both iterates guarantee that

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t}.$$

Convergence Result

Proof:

- **Case 1:** best iterate $\bar{\mathbf{x}}_T = \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$.

$$\sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f^\star) \geq \left(\sum_{t=1}^{T} \eta_t \right) (f(\bar{\mathbf{x}}_T) - f^\star). \qquad (f(\bar{\mathbf{x}}_T) = \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t) \leq f(\mathbf{x}_t))$$

Combining the above inequality with Lemma 2 (as restated below),

$$\sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f^\star) \leq \frac{1}{2} \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{1}{2} \sum_{t=1}^{T} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2,$$

we have completed the proof of the desired result:

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2 \sum_{t=1}^{T} \eta_t} + \frac{\sum_{t=1}^{T} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2 \sum_{t=1}^{T} \eta_t}.$$

Convergence Result

Proof:

- **Case 2:** average iterate $\bar{\mathbf{x}}_T = \sum_{t=1}^{T} \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^{T} \eta_t}$.

$$\sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f^\star) = \left(\sum_{t=1}^{T} \eta_t \right) \left(\sum_{t=1}^{T} \frac{\eta_t}{\sum_{t=1}^{T} \eta_t} f(\mathbf{x}_t) - f^\star \right)$$

(distribution)

$$\geq \left(\sum_{t=1}^{T} \eta_t \right) \left(f\left(\sum_{t=1}^{T} \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^{T} \eta_t} \right) - f^\star \right)$$

(Jensen's inequality)

Thus, we achieve the desired result:

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2 \sum_{t=1}^{T} \eta_t} + \frac{\sum_{t=1}^{T} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2 \sum_{t=1}^{T} \eta_t}. \qquad \square$$

Convergent Step Size

Theorem 3. *Under the assumption that $\|\nabla f(\cdot)\| \leq G$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method (note that the step size setting cannot use knowledge of T ahead of time). If*

$$\frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t} \to 0 \text{ as } T \to \infty,$$

then we have $f(\bar{\mathbf{x}}_T) \to f^\star$ as $T \to \infty$ for both best and average iterates.

Proof: From the second gradient descent lemma, we know that

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t} \leq G^2$$

The condition $\frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t} \to 0$ implies the convergence of the second term.

Moreover, this condition implies $\sum_{t=1}^T \eta_t \to \infty$ (think why?). □

Convergent Step Size

Theorem 3. *Under the assumption that $\|\nabla f(\cdot)\| \leq G$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method (note that the step size setting cannot use knowledge of T ahead of time). If*

$$\frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t} \to 0 \text{ as } T \to \infty,$$

then we have $f(\bar{\mathbf{x}}_T) \to f^\star$ as $T \to \infty$ for both best and average iterates.

Example:

a typical ***time-varying*** (in fact, decreasing) step sizes:

$$\eta_t = \frac{1}{\sqrt{t}} \Rightarrow \frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t} \approx \frac{\log T}{\sqrt{T}} \to 0.$$

Convergence without Optimal Value

Theorem 4. *Under the same assumptions with Theorem 3. Let* $\{\mathbf{x}_t\}_{t=1}^T$ *be the sequence generated by GD with step size*

$$\eta_t = \frac{1}{\|\nabla f(\mathbf{x}_t)\|\sqrt{t}}.$$

Then

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{G\left(\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \log T + 1\right)}{2\sqrt{T}} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right),$$

where $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ *or* $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$.

This tuning does not require the knowledge of $f^\star$, but the convergence rate exhibits a $\log T$ factor gap.

Convergence without Optimal Value

Proof:

$$
\begin{aligned}
f(\bar{\mathbf{x}}_T) - f^\star &\le \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t} && \text{(Lemma 3)} \\
&\le \frac{G\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t \|\nabla f(\mathbf{x}_t)\|} + \frac{G\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t \|\nabla f(\mathbf{x}_t)\|} && (\|\nabla f(\cdot)\| \le G) \\
&\le \frac{G\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \frac{1}{\sqrt{t}}} + \frac{G\sum_{t=1}^T \frac{1}{t}}{2\sum_{t=1}^T \frac{1}{\sqrt{t}}} && (\textstyle\sum_{t=1}^T \frac{1}{t} \le \log T + 1) \\
& && (\sqrt{T} \le \textstyle\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 2\sqrt{T})
\end{aligned}
$$

Thus,

$$
f(\bar{\mathbf{x}}_T) - f^\star \le \frac{G\left(\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \log T + 1\right)}{2\sqrt{T}} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right). \qquad \square
$$

Part 3. Optimal Rates

- Optimal Result with Known T

- Optimal Result with Unknown T

Towards Optimal Resolutions

Theorem 4. *Under the same assumptions with Theorem 1. Let* $\{\mathbf{x}_t\}_{t=1}^T$ *be the sequence generated by GD with step size*

$$\eta_t = \frac{1}{\|\nabla f(\mathbf{x}_t)\|\sqrt{t}}.$$

Then

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{G\left(\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \log T + 1\right)}{2\sqrt{T}} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right),$$

where $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ *or* $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$.

Theorem 2. *Under the same assumptions with Theorem 1. Let* $\{\mathbf{x}_t\}_{t=1}^T$ *be the sequence generated by the gradient descent method with Polyak step size and* $f^\star$ *be the optimal value. Define* $\bar{\mathbf{x}}_T = \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$*, we have*

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{G\|\mathbf{x}_1 - \mathbf{x}^\star\|}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

with Polyak's step size (known $f^\star$)

Now, we will improve this to optimality with an additional ***bounded domain*** assumption.

i.e., $\|\mathbf{x} - \mathbf{y}\| \leq D$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

Optimal Result with Known T

Theorem 5. *Under the same assumptions with Theorem 1, assume the feasible domain $\mathcal{X}$ is bounded and convex with a diameter $D > 0$, that is, $\|\mathbf{x} - \mathbf{y}\|_2 \leq D$ holds for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by GD with step size*

$$\eta_t = \frac{D}{G\sqrt{T}}.$$

Then

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),$$

where $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ or $\bar{\mathbf{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$.

Optimal Result with Known T

step size $\eta_t = \frac{D}{G\sqrt{T}}$ $\Rightarrow$ $f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ or $\bar{\mathbf{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$

Proof: Plugging $\eta_t = \frac{D}{G\sqrt{T}}$ into

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t} \quad (\|\mathbf{x}_1 - \mathbf{x}^\star\| \leq D)$$

$$(\|\nabla f(\cdot)\| \leq G)$$

Notice that $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t} = \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$. $\square$

Optimal Result with Known T

step size $\eta_t = \frac{D}{G\sqrt{T}}$ $\Rightarrow$ $f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ or $\bar{\mathbf{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$

- This convergence rate is equivalent to the sample complexity bound $T = \frac{D^2G^2}{\varepsilon^2} = \mathcal{O}(\frac{1}{\varepsilon^2})$ to achieve $f(\bar{\mathbf{x}}_T) - f^\star \leq \varepsilon$.
- $\frac{DG}{\sqrt{T}}$ is already minimax optimal for convex and Lispchitz functions.
- This result needs to know the total round number T in advance.

The last characteristics could be undesirable in practice.

Optimal Result with Unknown T

Theorem 6. *Under the same assumptions with Theorem 1, assume the feasible domain $\mathcal{X}$ is bounded and convex with a diameter $D > 0$, that is, $\|\mathbf{x} - \mathbf{y}\|_2 \leq D$ holds for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by GD with step size*

$$\eta_t = \frac{D}{G\sqrt{t}}. \qquad \text{“anytime” algorithm}$$

Then

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),$$

where $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=\lceil T/2 \rceil}^T} f(\mathbf{x}_t)$ *or* $\bar{\mathbf{x}}_T \triangleq \sum_{t=\lceil T/2 \rceil}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=\lceil T/2 \rceil}^T \eta_t}$.

Intuition: bounded domain assumption ensures $\|\mathbf{x}_t - \mathbf{x}^\star\|$ (not just $\|\mathbf{x}_1 - \mathbf{x}^\star\|$) to be bounded so that we can avoid the $\mathcal{O}(\log T)$ factor in the analysis.

Optimal Result with Unknown T

Proof: It is easy to extend the second GD lemma from $t = 1, \ldots, T$ to $t = \lceil \frac{T}{2} \rceil, \ldots, T$:

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t}$$

$$\Longrightarrow \quad f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{\|\mathbf{x}_{\lceil \frac{T}{2} \rceil} - \mathbf{x}^\star\|^2}{2\left(\sum_{t=\lceil \frac{T}{2} \rceil}^T \eta_t\right)} + \frac{G^2}{2} \frac{\sum_{t=\lceil \frac{T}{2} \rceil}^T \eta_t^2}{\sum_{t=\lceil \frac{T}{2} \rceil}^T \eta_t}$$

$$\leq \frac{DG}{2} \frac{1}{\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}}} + \frac{DG}{2} \frac{\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{t}}{\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}}}$$

$(\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}} \geq \frac{T}{2} \cdot \frac{1}{\sqrt{T}} = \frac{\sqrt{T}}{2})$

$(\sum_{\lceil T/2 \rceil}^T \frac{1}{t} \leq \log(T+1) - \log(\lceil T/2 \rceil) \leq \log(3))$

$\approx \sqrt{T}$

$$\Longrightarrow \quad f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \qquad \square$$

"Parameter-Free" Extension

- Previous discussions focus on the ***bounded*** domain.
- Step size tuning in ***unbounded*** domain is always nontrivial.

Algorithm 1 DoG with SGD [Ivgi et al., 2023]

Input: feasible domain $\mathcal{X}$ (which can be unbounded); initial point $\widehat{\mathbf{x}}_0 \in \mathcal{X}$; step size $\{\eta_t\}_{t=1}^T$; a small constant $r_\varepsilon > 0$.

1: Set $\eta_0 = \frac{r_\varepsilon}{\|\mathbf{g}_0\|}$

2: **for** $t = 1$ **to** $\cdots$ (maybe T) **do**

3: Perform the SGD update

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \mathbf{g}_t], \tag{4}$$

where $\mathbf{g}_t$ is the stochastic gradient of f at $\mathbf{x}_t$ and the step size is set as

$$\eta_t = \frac{\bar{r}_t}{\sqrt{\sum_{s=1}^t \|\mathbf{g}_s\|^2}}, \text{ where } \bar{r}_t \triangleq \max_{s\in[t]} \max\{\|\mathbf{x}_s - \mathbf{x}_0\|, r_\varepsilon\} \tag{5}$$

4: **end for**

Output: weighted average $\bar{\mathbf{x}}_t = \frac{1}{\sum_{s=0}^{t-1} \bar{r}_s} \cdot \sum_{s=0}^{t-1} \bar{r}_s \mathbf{x}_s$.

Ivgi, Maor, Oliver Hinder, and Yair Carmon. DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule. ICML 2023.

“Parameter-Free” Extension

Assumption 1 (convexity)**.** The function $f : \mathcal{X} \mapsto \mathbb{R}$ is convex.

Assumption 2 (domain boundedness)**.** The feasible domain $\mathcal{X}$ is convex and bounded by D, that is, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have $\|\mathbf{x} - \mathbf{y}\| \leq D$.

Assumption 3 (boundedness of gradient estimates)**.** The norm of gradient estimates is bounded by G, that is, for any $\mathbf{x} \in \mathcal{X}$, we have $\|\widetilde{\nabla} f(\mathbf{x})\|_* \leq G$.

Theorem 1. *Under Assumptions 1–3, the* DoG *algorithm (Algorithm 1) achieves the following convergence guarantee:*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_t) - f_*\right] \leq \mathcal{O}\left(\frac{DG}{\sqrt{T}} \log_+\left(\frac{D}{r_\varepsilon}\right)\right), \tag{6}$$

where D and G are the upper bounds of the domain diameter and the stochastic gradient norm, as defined in Assumptions 2 and 3, respectively. Notably, those constants (D and G) are not required as the algorithmic input.

Results so far

- For convex and G-Lipschitz functions,

  GD with different settings (step size, output sequence) achieve

	Step Size	Output Sequence	Convergence Rate	Remark
(1)	$\eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$	$\mathcal{O}(1/\sqrt{T})$	optimal Polyak's step size require $f^\star$
(2)	$\eta_t = \frac{1}{\|\nabla f(\mathbf{x}_t)\|\sqrt{t}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$	$\mathcal{O}(\log T/\sqrt{T})$	suboptimal
(3)	$\eta_t = \frac{D}{G\sqrt{T}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$	$\mathcal{O}(1/\sqrt{T})$	bounded domain require T
(4)	$\eta_t = \frac{D}{G\sqrt{t}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=\lceil T/2 \rceil}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=\lceil T/2 \rceil}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=\lceil T/2 \rceil}^T \eta_t}$	$\mathcal{O}(1/\sqrt{T})$	bounded domain

Revisit the Value of Polyak Step Size

For ***unconstrained*** optimization, Polyak step size is very simple and enjoys many good properties (e.g., "universality"); though estimating $f^\star$ is a problem to explore.

> In this landscape of stepsize selection strategies, the Polyak stepsize, proposed by Polyak [1969] stands out for its theoretical elegance and convergence properties. Starting from an arbitrary $\boldsymbol{x}_1 \in \mathbb{R}^d$, the Polyak stepsize is defined as
>
> $$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{f(\boldsymbol{x}_t) - f^\star}{\|\boldsymbol{g}_t\|^2}\boldsymbol{g}_t, \tag{1}$$
>
> where $\mathbf{0} \neq \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$ and $f^\star = \min_{\boldsymbol{x}} f(\boldsymbol{x})$. If $\boldsymbol{g}_t = \mathbf{0}$, then $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$. This update rule can achieve linear convergence for strongly convex and smooth functions, $\mathcal{O}(1/T)$ rate for convex smooth functions, and $\mathcal{O}(1/\sqrt{T})$ rate for non-smooth convex ones. This is particularly interesting because all of these rates are achieved with a unique stepsize and without knowledge of smoothness or curvature constants. In other words, this update rule is *adaptive* to the geometry of the functions to optimize.

F. Orabona, R. D'Orazio. New Perspectives on the Polyak Stepsize: Surrogate Functions and Negative Results. NeurIPS 2025.

Revisit the Value of Polyak Step Size

For ***unconstrained*** optimization, Polyak step size is very simple and enjoys many good properties (e.g., "universality"); though estimating $f^\star$ is a problem to explore.

2 [50pts] Universality of GD with the Polyak Step Size

The Polyak step size is a fundamental result in optimization theory and is also the first step size introduced in this course. While it requires the optimal function value $f^\star$ as input that is undesired practice, it exhibits two notable properties that merit further exploration.

(i) **Universality.** GD with the Polyak step size is *"universal"*: it can achieve *near-optimal* convergence rates across all four settings *simultaneously* — {convex, strongly convex} × {Lipschitz, smooth} — without knowing the specific function class.

(ii) **Robustness.** Certain relaxations can be made to reduce the dependency on exact knowledge of $f^\star$, with minimal distortion to the algorithm's performance.

In this problem, we will investigate these above two aspects.

GD with the Polyak step size. We focus on unconstrained convex optimization, i.e., $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, and consider the following gradient descent update with the Polyak step size (assuming $\|\nabla f(\mathbf{x}_t)\| \neq 0$):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t), \quad \eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, \qquad t \in [T]. \tag{2.1}$$

(1) **[10pts]** Let's first review the strong convexity and smoothness. If function $f(\mathbf{x})$ is σ-strongly convex, please prove that (**Hint**: $h(\mathbf{y}) \triangleq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}_\star \rangle - \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}_\star\|^2$):

$$\frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_\star\|^2 \leq f(\mathbf{x}) - f^\star \leq \frac{1}{2\sigma}\|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, \tag{2.2}$$

where $\mathbf{x}_\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ is the optimum, and $f^\star = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.
If function $f(\mathbf{x})$ is L-smooth, please prove that (**Hint**: $f^\star \leq f(\mathbf{y}), \forall \mathbf{y} \in \mathbb{R}^d$):

$$\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f^\star \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}_\star\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, \tag{2.3}$$

(2) **[15pts]** In this subproblem, we demonstrate that gradient descent with the Polyak step size (2.1) achieves universality. Specifically, we will prove that:

$$\min_{t\in[T+1]} f(\mathbf{x}_t) - f^\star \leq R_T = \mathcal{O}\left(\min\left\{\frac{Gd_0}{\sqrt{T}}, \frac{G^2}{\sigma T}, \frac{Ld_0^2}{T}, Ld_0^2 \exp\left(\frac{-\sigma T}{4L}\right)\right\}\right) \tag{2.4}$$

holds for some R_T, where $d_0 \triangleq \|\mathbf{x}_1 - \mathbf{x}_\star\|$, G is the Lipschitz constant, i.e., $\|\nabla f(\mathbf{x}_t)\| \leq G$ for all $t \in [T]$, and we define $\sigma = 0$ if f is not strongly convex, and $L = +\infty$ if f is non-smooth. Now, please follow the steps below to prove (2.4).

AOPT 2025, HW1; Problem 2

Revisit the Value of Polyak Step Size

For ***unconstrained*** optimization, Polyak step size is very simple and enjoys many good properties (e.g., "universality"); though estimating $f^\star$ is a problem to explore.

2 [50pts] Universality of GD with the Polyak Step Size

Now, we consider the *robustness* aspect of the Polyak step size. To reduce the dependence on the knowledge of $f^\star$, we explore a scenario where only a *lower bound* $\bar{f}$ of $f^\star$ is required instead of its precise value. The lower bound $\bar{f}$ is generally easier to obtain. For instance, in the case of non-negative loss functions, a reasonable choice would be $\bar{f} = 0$.

Our goal remains to achieve the similar universality as $\mathcal{O}(R_T)$ from (2.4) with T *pre-specified*. Below, we will prove that this can be achieved within no more than T multiplied by a logarithmic factor iterations.

(3) **[5pts]** Please prove that in (2.1), the step size η_t, instead of being restricted to the exact Polyak step size value, only needs to lie within a *relaxed* range $\left[\frac{f(\mathbf{x}_t)-f^\star}{2\|\nabla f(\mathbf{x}_t)\|^2}, \frac{f(\mathbf{x}_t)-f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}\right]$ to achieve a best-iterate convergence rate of R'_T, specified as:

$$R'_T = \mathcal{O}\left(\min\left\{\frac{Gd_0}{\sqrt{T}}, \frac{G^2}{\sigma T}, \frac{Ld_0^2}{T}, Ld_0^2 \exp\left(\frac{-3\sigma T}{16L}\right)\right\}\right). \tag{2.7}$$

This relaxation leads to the following algorithm, where, at each epoch, the lower bound is updated, and gradient descent with the Polyak step size is restarted for T iterations.

Algorithm 1 Epoch Polyak Gradient Descent

Input: Iteration T, lower bound $\bar{f}$ of $f^\star$, initial point $\mathbf{x}_1$.

Initialize lower bound $\bar{f}^1 \leftarrow \bar{f}$
for Epoch $k = 1, \ldots,$ **do**
  $\mathbf{x}_1^k \leftarrow \mathbf{x}_1$
  for $t = 1, \ldots, T$ **do**
    $\mathbf{x}_{t+1}^k = \mathbf{x}_t^k - \eta_t^k \nabla f(\mathbf{x}_t^k)$ with $\eta_t^k = \frac{f(\mathbf{x}_t^k)-\bar{f}^k}{2\|\nabla f(\mathbf{x}_t^k)\|^2}$
  end for
  Update lower bound $\bar{f}^{k+1} \leftarrow \frac{1}{2}\left(\bar{f}^k + f(\bar{\mathbf{x}}^k)\right)$ with $\bar{\mathbf{x}}^k = \arg\min_{\mathbf{x}\in\{\mathbf{x}_t^k\}_{t\in[T+1]}} f(\mathbf{x})$
end for

Output: $\arg\min_{\bar{\mathbf{x}}\in\{\bar{\mathbf{x}}^k\}_{k\in[K+1]}} f(\bar{\mathbf{x}})$ with $(K+1)$ being the total epochs.

(4) **[20pts]** For Algorithm 1, assume that in each of the first K epochs, there exist some iterations where the relaxed condition in subproblem (3) is not satisfied, but it is eventually satisfied throughout the $(K+1)$-th epoch, thereby achieving the desired convergence rate of $\mathcal{O}(R'_T)$. Then what is the maximum possible value of K?

AOPT 2025, HW1; Problem 2

Part 4. Strongly Convex and Lipschitz

- Strong Convexity

- Convergence Result

Strongly Convex and Lipschitz

Theorem 7. *Under the same assumptions with Theorem 1, except that f is σ-strongly-convex. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by GD with step size*

$$\eta_t = \frac{2}{\sigma(t+1)}.$$

Then (i)

$$f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{2G^2}{\sigma(T+1)} = \mathcal{O}\left(\frac{1}{T}\right),$$

where $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ or $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{2t}{T(T+1)}\mathbf{x}_t$.

And (ii)

$$\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\| \leq \frac{2G}{\sigma\sqrt{T+1}}.$$

Strongly Convex and Lipschitz

Proof: we start by extending ***the first GD lemma*** to strongly convex case.

Strongly convex case:

$$\begin{aligned}
\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 &\le \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - 2\eta_t\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^\star\rangle + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2 \\
&\le \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - 2\eta_t\left(f(\mathbf{x}_t)-f^\star+\frac{\sigma}{2}\|\mathbf{x}_t-\mathbf{x}^\star\|^2\right) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2 \\
&\qquad \text{(strong convexity: } f(\mathbf{x}_t)-f(\mathbf{x}^\star)+\tfrac{\sigma}{2}\|\mathbf{x}_t-\mathbf{x}^\star\|^2 \le \langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^\star\rangle\text{)} \\
&\le (1-\sigma\eta_t)\|\mathbf{x}_t-\mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t)-f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

$$\Longrightarrow f(\mathbf{x}_t)-f^\star \le \frac{\eta_t^{-1}-\sigma}{2}\|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \frac{\eta_t^{-1}}{2}\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 + \frac{\eta_t G^2}{2} \quad \text{(rearranging)}$$

Strongly Convex and Lipschitz

$$f(\mathbf{x}_t) - f^\star \le \frac{\eta_t^{-1} - \sigma}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\eta_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 + \frac{\eta_t G^2}{2}$$

$$= \frac{\sigma}{4}\left((t-1)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - (t+1)\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) + \frac{G^2}{\sigma(t+1)}$$

$$\Longrightarrow \quad t(f(\mathbf{x}_t) - f^\star) \le \frac{\sigma}{4}\left((t-1)t\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - t(t+1)\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) + \frac{G^2}{\sigma}$$

telescope now

$$\Longrightarrow \quad \sum_{t=1}^{T} t(f(\mathbf{x}_t) - f^\star) \le \frac{\sigma}{4}\left(0 \cdot 1 \cdot \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 - T(T+1)\,\|\mathbf{x}_{T+1} - \mathbf{x}^\star\|^2\right) + \frac{G^2 T}{\sigma} \le \frac{G^2 T}{\sigma}$$

Next step: relating $\sum_{t=1}^{T} t(f(\mathbf{x}_t) - f(\mathbf{x}^\star))$ to $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star)$.

Strongly Convex and Lipschitz

Recall that the output sequence is $\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ or $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{2t}{T(T+1)}\mathbf{x}_t$.

Case 1: $$\sum_{t=1}^T t(f(\mathbf{x}_t) - f^\star) \geq \left(\sum_{t=1}^T t\right)(f(\bar{\mathbf{x}}_T) - f^\star) = \frac{T(T+1)}{2}(f(\bar{\mathbf{x}}_T) - f^\star)$$

Case 2: $$\sum_{t=1}^T t(f(\mathbf{x}_t) - f^\star) = \sum_{t=1}^T t f(\mathbf{x}_t) - \frac{T(T+1)}{2}f^\star = \frac{T(T+1)}{2}\left(\sum_{t=1}^T \underbrace{\frac{2t}{T(T+1)}}_{\text{(distribution)}} f(\mathbf{x}_t) - f^\star\right)$$

$$\geq \frac{T(T+1)}{2}(f(\bar{\mathbf{x}}_T) - f^\star)$$

(Jensen's inequality)

(*i*) is proved. □

Strongly Convex and Lipschitz

Proof: *(ii)* can be derived directly from *(i)* and strong convexity.

$$\frac{\sigma}{2}\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\|^2 \leq \langle \nabla f(\mathbf{x}^\star), \bar{\mathbf{x}}_T - \mathbf{x}^\star \rangle + \frac{\sigma}{2}\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\|^2 \leq f(\bar{\mathbf{x}}_T) - f^\star \leq \frac{2G^2}{\sigma(T+1)}$$

(strong convexity) — *(i)*

(first-order optimality condition: $\langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle \geq 0$)

Thus, we prove that no matter for which constructions of $\bar{\mathbf{x}}_T$, it holds that

$$\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\| \leq \frac{2G}{\sigma\sqrt{T+1}}.$$

(*ii*) is proved. □

Summary

Table 1: A summary of convergence rates of GD method.

Function Family	Step Size	Output Sequence	Convergence Rate	Remark
convex and G-Lipschitz	$\eta_t = \frac{f(\mathbf{x}_t)-f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$	$\mathcal{O}(1/\sqrt{T})$	optimal Polyak's step size require $f^\star$
	$\eta_t = \frac{1}{\|\nabla f(\mathbf{x}_t)\|\sqrt{t}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$	$\mathcal{O}(\log T/\sqrt{T})$	suboptimal
	$\eta_t = \frac{D}{G\sqrt{T}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$	$\mathcal{O}(1/\sqrt{T})$	bounded domain require T
	$\eta_t = \frac{D}{G\sqrt{t}}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=\lceil T/2\rceil}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=\lceil T/2\rceil}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=\lceil T/2\rceil}^T \eta_t}$	$\mathcal{O}(1/\sqrt{T})$	bounded domain
σ-strongly convex and G-Lipschitz	$\eta_t = \frac{2}{\sigma(t+1)}$	$\bar{\mathbf{x}}_T \triangleq \arg\min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$ $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$	$\mathcal{O}(1/T)$	$\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\|$ is bounded

Summary

GRADIENT DESCENT
- Convex Optimization Problem
- Gradient Descent
- The First Gradient Descent Lemma

OPTIMAL IN CONVEX AND LIPSCHITZ CASE
- Optimal Result with Known T
- Optimal Result with Unknown T

POLYAK'S STEP SIZE
- Polyak's Step Size
- Convergence
- Convergence Rate

STRONGLY CONVEX AND LIPSCHITZ
- Strong Convexity
- Convergence Result

CONVERGENCE WITHOUT OPTIMAL VALUE
- The Second Gradient Descent Lemma
- Convergent Step Size
- Convergence without Optimal Value

Q & A

Thanks!