Lecture 3. Gradient Descent Method

Advanced Optimization (Fall 2024)

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

Part 1. Gradient Descent

- Convex Optimization Problem

- Gradient Descent

- Performance Measure

- The First Gradient Descent Lemma

Convex Optimization Problem

- We adopt a minimization language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned}$$

- optimization variable $\mathbf{x} \in \mathbb{R}^d$
- objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$: convex and continuously differentiable
- feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$: convex

Goal

To output a sequence $\{\bar{\mathbf{x}}_t\}_{t=1}^T$ such that $\bar{\mathbf{x}}_t$ approximates $\mathbf{x}^\star$ when t goes larger.

- Function-value level: $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \leq \varepsilon(T)$
- Optimizer-value level: $\|\bar{\mathbf{x}}_T - \mathbf{x}^\star\| \leq \varepsilon(T)$

where $\{\bar{\mathbf{x}}_t\}_{t=1}^T$ can be ***statistics*** of the original sequence $\{\mathbf{x}_t\}_{t=1}^T$,

and $\varepsilon(T)$ is the *approximation error* and is a function of iterations T.

Goal

- In general, there are two performance measures (essentially same).

Convergence: $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \varepsilon(T),$

- **Qualitatively:** $\varepsilon(T) \to 0$ when $T \to \infty$
- **Quantitatively:** $\mathcal{O}(\frac{1}{\sqrt{T}})$ / $\mathcal{O}(\frac{1}{T})$ / $\mathcal{O}(\frac{1}{T^2})$ / $\mathcal{O}(\frac{1}{e^T})$ / $\ldots$

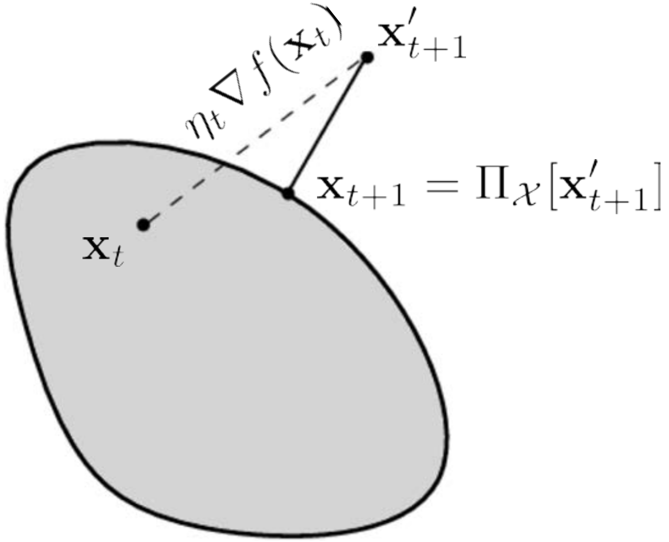
Complexity:

- **Definition:** number of iterations required to achieve $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \varepsilon$.
- **Quantitatively:** $\mathcal{O}(\frac{1}{\varepsilon^2})$ / $\mathcal{O}(\frac{1}{\varepsilon})$ / $\mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ / $\mathcal{O}(\ln(\frac{1}{\varepsilon}))$ / $\ldots$

corresponds to $\mathcal{O}(\frac{1}{\sqrt{T}})$ / $\mathcal{O}(\frac{1}{T})$ / $\mathcal{O}(\frac{1}{T^2})$ / $\mathcal{O}(\frac{1}{e^T})$ / $\ldots$

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.

Why Gradient Descent?

- For simplicity, we consider the ***unconstrained*** setting.

- **A General Idea:** ***Surrogate Optimization***

We aim to find a sequence of ***local upper bounds*** $U_1, \cdots, U_T$, where the surrogate function $U_t : \mathbb{R}^d \mapsto \mathbb{R}$ may depend on $\mathbf{x}_t$ such that

(i) $f(\mathbf{x}_t) = U_t(\mathbf{x}_t)$;

(ii) $f(\mathbf{x}) \leq U_t(\mathbf{x})$ holds for all $\mathbf{x} \in \mathbb{R}^d$;

(iii) $U_t(\mathbf{x})$ should be simple enough to minimize.

⇨ Then, our proposed algorithm would be $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} U_t(\mathbf{x})$

Why Gradient Descent?

- Following the ***surrogate optimization*** principle, let's invent GD for convex and ***smooth*** functions.

Proposition 1. *Suppose that f is convex and differentiable. Moreover, suppose that f is L-smooth with respect to ℓ_2-norm. Define the surrogate $U_t : \mathbb{R}^d \mapsto \mathbb{R}$ as*

$$U_t(\mathbf{x}) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 .$$

Then, we have

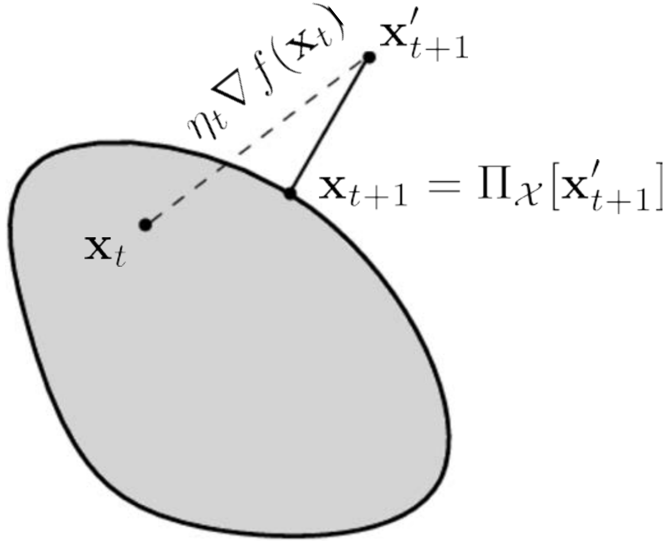
(i) $f(\mathbf{x}_t) = U_t(\mathbf{x}_t)$;

(ii) $f(\mathbf{x}) \leq U_t(\mathbf{x})$ holds for all $\mathbf{x} \in \mathbb{R}^d$;

(iii) $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} U_t(\mathbf{x})$ is equivalent to $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$.

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.

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

GD Convergence Analysis

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:

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

(convexity: $f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle$) □

Part 2. Polyak Step Size

- Polyak Step Size

- Convergence

- Convergence Rate

Polyak Step Size

- GD method satisfies the following inequality:

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

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

A natural idea:

minimizing the right-hand side of the inequality

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

assume known $f^\star$ for a moment

Polyak Step Size

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

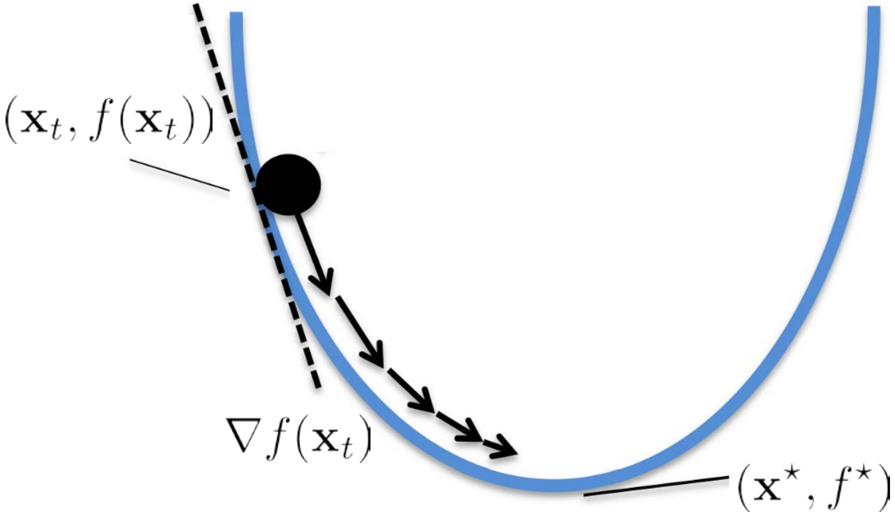
$$h(\eta) \triangleq -2\eta(f(\mathbf{x}_t) - f^\star) + \eta^2\|\nabla f(\mathbf{x}_t)\|^2$$

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

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

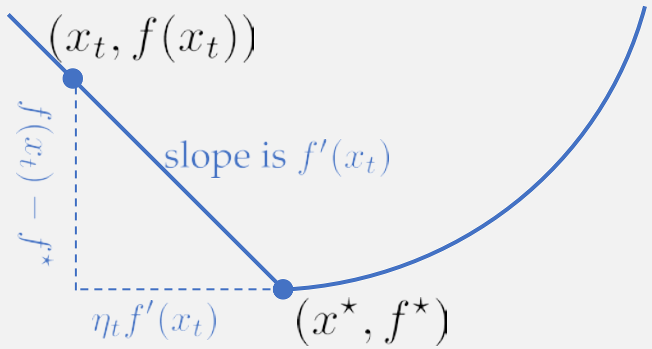
Polyak step size: $$\eta_t = \begin{cases} \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$$

A Geometric View of Polyak Step Size



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

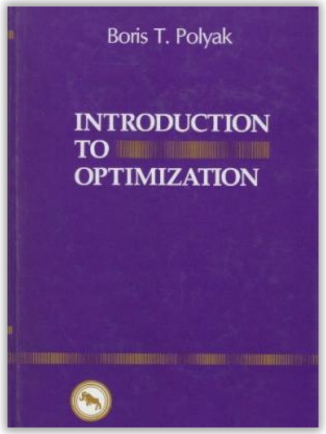
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t), \quad \eta_t = \frac{f(\mathbf{x}_t) - f^\star}{\|\nabla f(\mathbf{x}_t)\|^2}$$

Polyak Step Size

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

assume known $f^\star$ for a moment.

Boris T. Polyak
1935-2023



Introduction to optimization

Boris T. Polyak

Optimization Software, Inc., 1987

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

- We can also derive the convergence rate.

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*

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

Proof:

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

□

Part 3. 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:

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

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

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

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

Convergence Result

Proof:

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

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

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

(distribution)

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

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

(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 same assumptions with Theorem 1. 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 $f(\bar{\mathbf{x}}_T) \to f^\star$ as $T \to \infty$.

Proof: Indeed, this structure appears in the second gradient descent lemma.

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

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 same assumptions with Theorem 1. 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 $f(\bar{\mathbf{x}}_T) \to f^\star$ as $T \to \infty$.

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 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

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

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

Convergence without Optimal Value

Proof:

$$f(\bar{\mathbf{x}}_T) - f^\star \le \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2\sum_{t=1}^T \eta_t} \qquad \text{(the second GD lemma)}$$

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

$$\le \frac{G\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\sum_{t=1}^T \frac{1}{\sqrt{t}}} + \frac{G\sum_{t=1}^T \frac{1}{t}}{2\sum_{t=1}^T \frac{1}{\sqrt{t}}} \qquad (\textstyle\sum_{t=1}^T \frac{1}{t} \le \log T + 1) \quad (\sqrt{T} \le \textstyle\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 2\sqrt{T})$$

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 4. Optimal in Convex and Lipschitz Case

- 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 $f^\star$)

Remark: The last theorem gives an $\mathcal{O}(\log T/\sqrt{T})$ convergence rate. However, this rate is ***worse*** than the $\mathcal{O}(1/\sqrt{T})$ with Polyak step size.

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

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

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

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

$(\|\mathbf{x}_1 - \mathbf{x}^\star\| \leq D)$

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

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

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

- $\frac{DG}{\sqrt{T}}$ convergence rate is equivalent to $T = \frac{D^2G^2}{\varepsilon^2}$ complexity result 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*

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

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 $t = \lceil \frac{T}{2} \rceil, \ldots, T$:

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

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

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

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

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

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

□

Parameter-Free Extension

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

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.

Part 5. 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:

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

$$\le \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \left(f(\mathbf{x}_t) - f^\star + \frac{\sigma}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \right) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

(strong convexity: $f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\sigma}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle$)

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

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

(rearranging)

Strongly Convex and Lipschitz

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

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

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

telescope now

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

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

(distribution)

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

(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
  - Performance Measure
  - 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!