



Lecture 3. Gradient Descent Method

Advanced Optimization (Fall 2023)

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}^* when t goes larger.

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

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: $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon(T),$

- **Qualitatively:** $\varepsilon(T) \rightarrow 0$ when $T \rightarrow \infty$

- **Quantitatively:** $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) / \mathcal{O}\left(\frac{1}{T}\right) / \mathcal{O}\left(\frac{1}{T^2}\right) / \mathcal{O}\left(\frac{1}{e^T}\right) / \dots$

Complexity:

- **Definition:** number of iterations required to achieve $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon.$

- **Quantitatively:** $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right) / \mathcal{O}\left(\frac{1}{\varepsilon}\right) / \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right) / \mathcal{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right) / \dots$

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

Gradient Descent

- GD Template:

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

- \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?

Let's simply focus on the *unconstrained* setting.

- **Idea: surrogate optimization**

We aim to find a sequence of *local upper bounds* U_1, \dots, 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:

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

- \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}^* be the optimal set of the optimization problem and f^* be the optimal value. Then for any $\mathbf{x}^* \in \mathcal{X}^*$ and $t \geq 0$,*

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

Proof: $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 = \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^*\|^2$ (GD)
 $\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^*\|^2$ (Pythagoras Theorem)
 $= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$
 $\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^*) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$
(convexity: $f(\mathbf{x}_t) - f^* = f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle$) \square

Part 2. Polyak Step Size

- Polyak Step Size
- Convergence
- Convergence Rate

Polyak Step Size

- GD method satisfies the following inequality:

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

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

A natural idea:

minimizing the right-hand side of the inequality

$$\Rightarrow \eta_t = \frac{f(\mathbf{x}_t) - f^*}{\|\nabla f(\mathbf{x}_t)\|^2} \quad \text{assume known } f^* \text{ for a moment}$$

Polyak Step Size

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

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

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

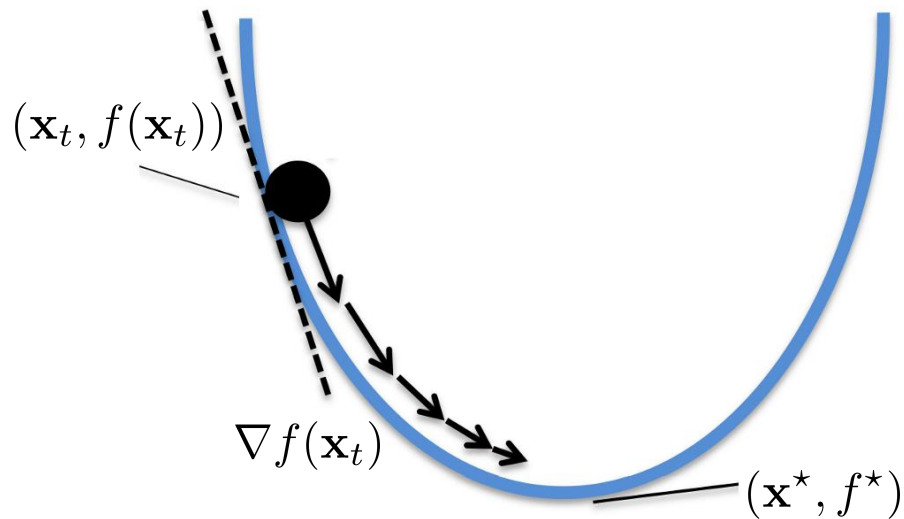
Cornercase: when $\nabla f(\mathbf{x}_t) = \mathbf{0}$

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

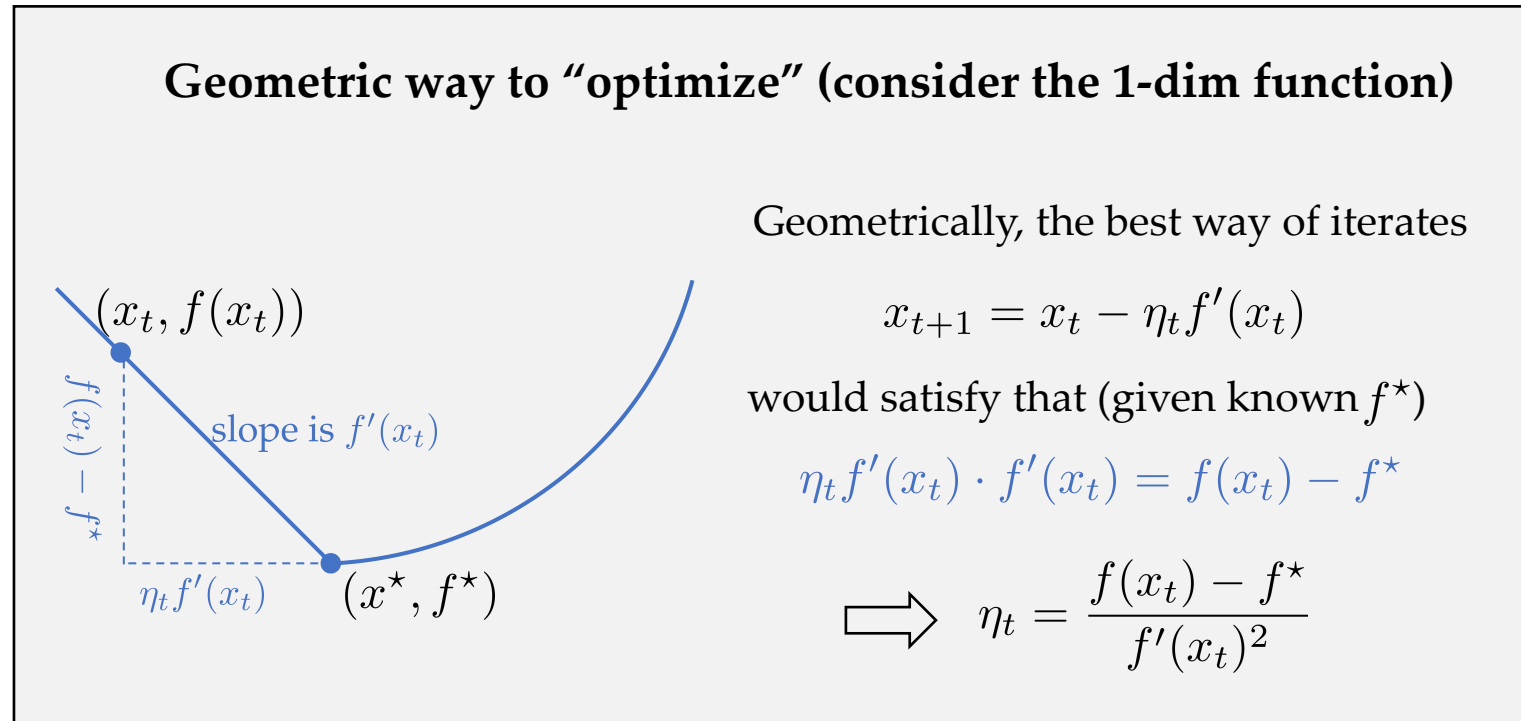
$$\text{Polyak step size: } \eta_t = \begin{cases} \frac{f(\mathbf{x}_t) - f^*}{\|\nabla f(\mathbf{x}_t)\|^2}, & \nabla f(\mathbf{x}_t) \neq \mathbf{0} \\ 1, & \nabla f(\mathbf{x}_t) = \mathbf{0} \end{cases}$$

Without loss of generality, we assume $\nabla f(\mathbf{x}_t) \neq \mathbf{0}$ from now on.

A Geometric View of Polyak Step Size



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



Gradient Descent with Polyak Step Size

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

Polyak Step Size

Polyak step size:

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

assume known f^ for a moment.*

 **Peter Richtarik** @peter_richtarik · 2月4日

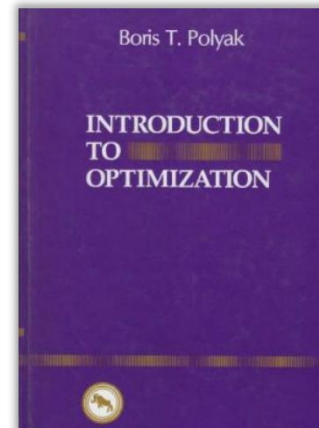
Boris Polyak (1935) passed away today. An immensely gentle person the way I knew him, and a giant of science in general and optimization in particular. His name, results and legacy will live forever.

 **Peter Richtarik** @peter_richtarik

With Boris Polyak @ Optimization & Statistical Learning, Les Houches, France, 2015 (Alfonso S. Bandeira, John Duchi, Alexander Rakhlin, Vladimir Spokoiny, Boris T. Polyak, Ekaterina Krymova, Yury Maximov, Peter Richtarik = me)



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^* be the optimal value. Then,*

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

$$(ii) \quad f(\mathbf{x}_t) \rightarrow f^* \text{ as } t \rightarrow \infty.$$

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

Convergence

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

(the first GD lemma)

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

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

(i) is proved.

Convergence

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

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

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

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

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

Infinite summation is bounded by constants \rightarrow **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^* be the optimal value. Define $\bar{\mathbf{x}}_T = \arg \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$, we have*

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

Proof:

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

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^*

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

assume known f^ for a moment*

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

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*

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

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

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^*) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\ \implies \eta_t (f(\mathbf{x}_t) - f^*) &\leq \frac{1}{2} (\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2) + \frac{1}{2} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \quad \square \end{aligned}$$

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 $\bar{\mathbf{x}}_T \triangleq \frac{\sum_{t=1}^T \eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$, we have

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

Convergence Result

Proof:

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

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

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

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

we have completed the proof of the desired result:

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

Convergence Result

Proof:

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

$$\begin{aligned} \sum_{t=1}^T \eta_t (f(\mathbf{x}_t) - f^*) &= \left(\sum_{t=1}^T \eta_t \right) \left(\sum_{t=1}^T \left[\frac{\eta_t}{\sum_{t=1}^T \eta_t} \right] f(\mathbf{x}_t) - f^* \right) \\ &\geq \left(\sum_{t=1}^T \eta_t \right) \left(f \left(\sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t} \right) - f^* \right) \end{aligned}$$

(Jensen's inequality)

Thus, we achieve the desired result:

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

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

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

then $f(\bar{\mathbf{x}}_T) \rightarrow f^*$ as $T \rightarrow \infty$.

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

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

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

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

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

then $f(\bar{\mathbf{x}}_T) \rightarrow f^$ as $T \rightarrow \infty$.*

Example:

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

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

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

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

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:

$$\begin{aligned} f(\bar{\mathbf{x}}_T) - f^* &\leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2 \sum_{t=1}^T \eta_t} && \text{(the second GD lemma)} \\ &\leq \frac{G \|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2 \sum_{t=1}^T \eta_t \|\nabla f(\mathbf{x}_t)\|} + \frac{G \sum_{t=1}^T \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2}{2 \sum_{t=1}^T \eta_t \|\nabla f(\mathbf{x}_t)\|} && (\|\nabla f(\cdot)\| \leq G) \\ &\leq \frac{G \|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2 \sum_{t=1}^T \frac{1}{\sqrt{t}}} + \frac{G \sum_{t=1}^T \frac{1}{t}}{2 \sum_{t=1}^T \frac{1}{\sqrt{t}}} && \left(\sum_{t=1}^T \frac{1}{t} \leq \log T + 1 \right) \\ &&& (\sqrt{T} \leq \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}) \end{aligned}$$

Thus,

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

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

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

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^* be the optimal value. Define $\bar{\mathbf{x}}_T = \arg \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t)$, we have

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

with Polyak's step size (known f^)*

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.

We show that this can be improved with an additional assumption of *bounded domain*.

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

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

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

$$\text{step size } \eta_t = \frac{D}{G\sqrt{T}} \implies f(\bar{\mathbf{x}}_T) - f^* \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$
$$\bar{\mathbf{x}}_T \triangleq \arg \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t) \text{ or } \bar{\mathbf{x}}_T \triangleq \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$$

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

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

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

Optimal Result with Known T

$$\text{step size } \eta_t = \frac{D}{G\sqrt{T}} \implies f(\bar{\mathbf{x}}_T) - f^* \leq \frac{DG}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$
$$\bar{\mathbf{x}}_T \triangleq \arg \min_{\{\mathbf{x}_t\}_{t=1}^T} f(\mathbf{x}_t) \text{ or } \bar{\mathbf{x}}_T \triangleq \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$$

- $\frac{DG}{\sqrt{T}}$ convergence rate is equivalent to $T = \frac{D^2G^2}{\varepsilon^2}$ complexity result to achieve $f(\bar{\mathbf{x}}_T) - f^* \leq \varepsilon$.
- $\frac{DG}{\sqrt{T}}$ is already minimax optimal for convex and Lipschitz 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

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

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 $\|\mathbf{x}_t - \mathbf{x}^*\|$ (not just $\|\mathbf{x}_1 - \mathbf{x}^*\|$) 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 $t = 1, \dots, T$ to $t = \lceil \frac{T}{2} \rceil, \dots, T$:

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

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

$$\left(\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}} \geq \frac{T}{2} \cdot \frac{1}{\sqrt{T}} = \frac{\sqrt{T}}{2} \right) \leq \frac{DG}{2} \frac{1}{\left[\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}} \right]} + \frac{DG \left[\frac{\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{t}}{\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{\sqrt{t}}} \right]}{2} \left(\sum_{t=\lceil \frac{T}{2} \rceil}^T \frac{1}{t} \leq \log(T+1) - \log(\lceil \frac{T}{2} \rceil) \leq \log(3) \right)$$

$\approx \frac{1}{\sqrt{T}}$

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

□

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)

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

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)

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

Strongly Convex and Lipschitz

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

Strongly convex case:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t \left(f(\mathbf{x}_t) - f^* + \frac{\sigma}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 \right) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\ &\quad \text{(strong convexity: } f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\sigma}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle\text{)} \\ &\leq (1 - \sigma\eta_t) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^*) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\ \implies f(\mathbf{x}_t) - f^* &\leq \frac{\eta_t^{-1} - \sigma}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\eta_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \frac{\eta_t G^2}{2} \quad \text{(rearranging)}\end{aligned}$$

Strongly Convex and Lipschitz

$$\begin{aligned} f(\mathbf{x}_t) - f^* &\leq \frac{\eta_t^{-1} - \sigma}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\eta_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \frac{\eta_t G^2}{2} \\ &= \frac{\sigma}{4} \left((t-1) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - (t+1) \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \right) + \frac{G^2}{\sigma(t+1)} \end{aligned}$$

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

telescope now

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

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

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 $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{2t}{T(T+1)} \mathbf{x}_t$.

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

$$\begin{aligned} \text{Case 2: } \sum_{t=1}^T t(f(\mathbf{x}_t) - f^*) &= \sum_{t=1}^T t f(\mathbf{x}_t) - \frac{T(T+1)}{2} f^* = \frac{T(T+1)}{2} \left(\sum_{t=1}^T \underbrace{\left[\frac{2t}{T(T+1)} \right]}_{\text{(distribution)}} f(\mathbf{x}_t) - f^* \right) \\ &\geq \frac{T(T+1)}{2} (f(\bar{\mathbf{x}}_T) - f^*) \\ &\quad \text{(Jensen's inequality)} \end{aligned}$$

(i) is proved. \square

Strongly Convex and Lipschitz

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

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

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

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

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

(ii) is proved. \square

Summary

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

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

Summary

