Gradient Descent

DSAI5104 (2025-26)

LASSO: Motivation

- LASSO: Least Absolute Shrinkage and Selection Operator is a major statistical methodology that has found many applications.
- Suppose we have N data points $\mathbf{x}_i \in \mathbb{R}^n$ with corresponding observations b_i, $i \in [N] := \{1, \ldots, N\}$. The linear regression is

$$b_i \approx \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i \in [N].$$

  The least squares model is

$$\min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) := \frac{1}{2} \sum_{i=1}^{N} \left(\mathbf{x}_i^\top \boldsymbol{\beta} - b_i \right)^2 = \frac{1}{2} \| X\boldsymbol{\beta} - \mathbf{b} \|^2.$$

- L_1-regularization leads to LASSO:

$$\min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

  where $\lambda > 0$ is a parameter and $\| \cdot \|_1$ is the ℓ_1-norm.

LASSO: Generalization

- LASSO has the hallmark of many machine learning problems:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \ell_i(\mathbf{x}_i, b_i, \boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta})$$

  where ℓ_i is the loss function at data point $(\mathbf{x}_i, b_i)$ and $\mathcal{R}$ is the regularization terms.

- There are many choices, e.g.,

$$\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2 \quad \text{(Euclidean norm)}$$

  We will see some of them in action later on.

- The problem is structural: data can be fed in batches.
- While $f(\boldsymbol{\beta})$ is differentiable, $\|\boldsymbol{\beta}\|_1$ is not (i.e., non-differentiable).

Problem Set-up

Consider the minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}).$$

where $f \in C^1(\mathbb{R}^n)$:
f is once continuously differentiable on its domain $\mathbb{R}^n$.
$\nabla f(\mathbf{x})$ denotes the gradient of f at $\mathbf{x}$.

Example:

$$\min f(x_1, x_2) = x_1^2 + x_2^2, \qquad (x_1, x_2) \in \mathbb{R}^2.$$

It is easy to see $(x_1 = 0, x_2 = 0)$ is the optimal solution, as it gives the lowest values of the objective function f.
Unfortunately, practical problems are not as simple. Fortunately, they are NOT as hard as impossibly to solve. We usually need iterative steps to find a solution.

GD: Gradient Descent

Suppose $\mathbf{x}^k$ is the current iterate with $\nabla f(\mathbf{x}^k) \neq 0$. Gradient Descent (GD) finds the next iterate by

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \Big(-\nabla f(\mathbf{x}^k) \Big),$$

where we take a step from the current iterate $\mathbf{x}^k$ along its negative gradient direction $(-\nabla f(\mathbf{x}^k))$ with a steplength $\alpha_k > 0$.

Let

$$\mathbf{d}^k = -\nabla f(\mathbf{x}^k).$$

GD Generates Better Points

We consider the functional values along the half line $\mathbf{x}^k + \alpha \mathbf{d}^k$, $\alpha \geq 0$. By the mean-value theorem, we have

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) = f(\mathbf{x}^k) + \underbrace{\langle \nabla f(\mathbf{x}^k + \tilde{\alpha}\mathbf{d}^k),\ \alpha \mathbf{d}^k \rangle}_{<0 \text{ when } \alpha \text{ is small enough}}$$

$$< f(\mathbf{x}^k), \qquad \text{(when } \alpha \text{ is sufficiently small)}$$

where $\tilde{\alpha} \in (0, \alpha]$. The inner product above is negative because when α is small, the gradient $\nabla f(\mathbf{x}^k + \tilde{\alpha}\mathbf{d}^k)$ is close to $\nabla f(\mathbf{x}^k)$ using $f \in C^1$. We can always ensure

$$f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$$

We may generate a sequence $\{\mathbf{x}^k\}$ with the functional values $\{f(\mathbf{x}^k)\}$ decreasing.

Two questions:

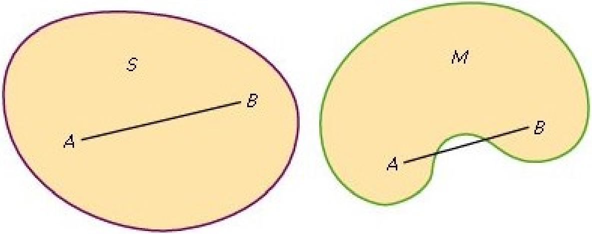
- Where the sequence $\{\mathbf{x}^k\}$ leads? (convergence analysis)
- How fast it leads? (convergence rate)

Convex Sets

Definition

A set $C \subset \mathbb{R}^b$ is convex if for any pair $\mathbf{x}, \mathbf{y} \in C$ we have

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C \qquad \forall \ \lambda \in [0, 1].$$



Convex Functions

Definition

A function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is said to be Convex if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}),\ \forall\, \mathbf{x}, \mathbf{y} \in \text{dom} f,\ \lambda \in [0,1],$$

where $\text{dom} f$ (the domain of f) is convex:

$$\text{dom} f = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) < +\infty\}.$$

Examples

$$\begin{aligned} f_1(\mathbf{x}) &= x_1 + x_2, & \text{dom} f_1 &= \mathbb{R}^2 \\ f_2(\mathbf{x}) &= x_1^2 + x_2^2, & \text{dom} f_2 &= \mathbb{R}^2 \\ f_3(x) &= -\sqrt{x}, & \text{dom} f_3 &= [0, \infty) \end{aligned} \qquad f_4(\mathbf{x}) = \delta_C(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ \infty & \text{otherwise,} \end{cases}$$

where $C \subset \mathbb{R}^n$ is a convex set. f_4 is called the indicator function of C and $\text{dom} f_4 = C$.

Two Consequences of Convex Functions

Theorem (Optimality of Convex Optimization)

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable and convex. We must have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \qquad \forall\ \mathbf{x}, \mathbf{y}.$$

Furthermore, if $\mathbf{x}^*$ is an optimal solution of

$$\min_{\mathbf{x}}\ f(\mathbf{x}),$$

we must have $\nabla f(\mathbf{x}^*) = 0$.

Proof. Since f is convex, we have

$$f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) = f(t\mathbf{y}+(1-t)\mathbf{x}) \leq tf(\mathbf{y})+(1-t)f(\mathbf{x}), \quad \forall\, t \in [0,1]$$

We then have

$$t(f(\mathbf{y}) - f(\mathbf{x})) \geq f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x}),$$

which implies

$$f(\mathbf{y})-f(\mathbf{x}) \geq \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \to \langle \nabla f(\mathbf{x}),\ \mathbf{y}-\mathbf{x}\rangle \text{ as } t \to 0^+$$

This proves the first part.

Now suppose $\mathbf{x}^*$ is optimal. We then have $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x}$. If $\nabla f(\mathbf{x}^*) \neq 0$, then there is a new point $\mathbf{x} = \mathbf{x}^* + \alpha(-\nabla f(\mathbf{x}^*))$ such that $f(\mathbf{x}) < f(\mathbf{x}^*)$ when $\alpha > 0$ is small enough. This contradicts the optimality of $\mathbf{x}^*$. Hence, we must have $\nabla f(\mathbf{x}^*) = 0$. □

Function Class: L-Lipschitz

Definition (L-Lipschitz)

A function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is L-Lipschitz if there exists $L > 0$ such that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\| \qquad \forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom} f.$$

Examples:

$$f_1(x) = |x| \quad \text{(absolute value function)}$$

$$f_2(\mathbf{x}) = \sqrt{\|\mathbf{x}\|^2 + \epsilon} \quad \text{with } \epsilon > 0.$$

Boundedness of the gradients: For f being L-Lipschitz, if it is also differentiable, then

$$\|\nabla f(\mathbf{x})\| \leq L \qquad \forall\ \mathbf{x} \in \mathrm{dom} f.$$

ϵ-Optimality

Definition (ϵ-optimal solution)

Consider the optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \; f(\mathbf{x}).$$

Suppose $\mathbf{x}^*$ is an optimal solution. For a given $\epsilon > 0$ (a small number), a given point $\mathbf{x}$ is said to be an ϵ-optimal solution if

$$f(\mathbf{x}) \leq f(\mathbf{x}^*) + \epsilon.$$

Example: Let $f(x) = x^2$, $x \in \mathbb{R}$. It is easy to see $x_* = 0$ is the (unique) optimal solution. Any point $x \in [-\sqrt{\epsilon}, \sqrt{\epsilon}]$ is an ϵ-optimal solution.

GD for Convex L-Lipschitz Optimization

Theorem (GD for Convex L-Lipschitz Optimization: $O(1/\epsilon^2)$)

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be (i) differentiable, (ii) L-Lipschitz, and (iii) convex. Let $\mathbf{x}^*$ be an optimal minimizer of f. Let $\mathbf{x}^1$ be the initial point and $\epsilon > 0$ be given. We generate:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \nabla f(\mathbf{x}^t) \quad t = 1, \ldots, T-1, \quad \alpha = \frac{\epsilon}{L^2}$$

where $T = \lceil \frac{D^2 L^2}{\epsilon^2} \rceil$ and $D = \|\mathbf{x}^1 - \mathbf{x}^*\|$. We must have

$$f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^t\right) \leq f(\mathbf{x}^*) + \epsilon.$$

Note: The notation $\lceil a \rceil$ denotes the smallest integer no less than $a \geq 0$.

Proof.

We have

$$
\begin{aligned}
& f(\mathbf{x}^t) - f(\mathbf{x}^*) \le \langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle \quad \text{(convexity)} \\
= \; & \frac{1}{\alpha} \langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^t - \mathbf{x}^* \rangle \quad \text{(by GD)} \\
= \; & \frac{1}{2\alpha} \left(\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \right) \quad \text{(cosine law)} \\
= \; & \frac{1}{2\alpha} \left(\|\mathbf{x}^t - \mathbf{x}^*\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \right) + \frac{\alpha}{2} \|\nabla f(\mathbf{x}^t)\|^2 \quad \text{(by GD)} \\
\le \; & \frac{1}{2\alpha} \left(\|\mathbf{x}^t - \mathbf{x}^*\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \right) + \frac{\alpha L^2}{2} \quad \text{(boundedness of } \nabla f(\mathbf{x})\text{)}
\end{aligned}
$$

Proof Continued

Then

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^t) - f(\mathbf{x}^*) &\leq \frac{1}{2\alpha T}\left(\|\mathbf{x}^1 - \mathbf{x}^*\|^2 - \|\mathbf{x}^{T+1} - \mathbf{x}^*\|^2\right) + \frac{\alpha L^2}{2} \\
&\leq \frac{1}{2\alpha T}\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{\alpha L^2}{2} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}
$$

By convexity again, we have

$$
f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^k\right) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \epsilon. \qquad \square
$$

Function Class: L-Smooth Functions

- We say f is L-smooth if $f \in C^1(\mathbb{R}^n)$ and there exists $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The Lipschitz constant of the gradient function ∇f is L.

Examples

$$f_1(\mathbf{x}) = \|\mathbf{x}\|^2,\ \nabla f(\mathbf{x}) = 2\mathbf{x},\ L = 2$$

$$f_2(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x},\ \nabla f_2(\mathbf{x}) = A\mathbf{x}$$

$$\|\nabla f_2(\mathbf{x}) - \nabla f_2(\mathbf{y})\| = \|A(\mathbf{x} - \mathbf{y})\| \leq \rho(A)\|\mathbf{x} - \mathbf{y}\|$$

where A is symmetric and $\rho(A)$ is the latest absolute eigenvalue of A (i.e., the spectral norm of A).

Descent Lemma for L-smooth Functions

Theorem (Descent Lemma)

Suppose f is L-smooth. Then for all $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$, it holds that

$$f(\mathbf{y}) \leq f(\mathbf{x}) + [\nabla f(\mathbf{x})]^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define $\psi(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. Then $\psi(0) = f(\mathbf{x})$ and $\psi'(s) = (\mathbf{y} - \mathbf{x})^\top \nabla f(\mathbf{x} + s(\mathbf{y} - \mathbf{x}))$. Hence

$$\begin{aligned}
\psi(1) = \psi(0) + \int_0^1 \psi'(s) ds &= \psi(0) + \psi'(0) + \int_0^1 (\psi'(s) - \psi'(0)) ds \\
&= \psi(0) + \psi'(0) + \int_0^1 (\mathbf{y} - \mathbf{x})^\top [\nabla f(\mathbf{x} + s(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})] ds \\
&\leq \psi(0) + \psi'(0) + L \int_0^1 s \|\mathbf{y} - \mathbf{x}\|^2 ds.
\end{aligned}$$

GD for Convex L-smooth Minimization

Theorem ($O(1/k)$ complexity)

Suppose f is L-smooth and convex and $\mathbf{x}^*$ is a minimizer of f. Let $\{\mathbf{x}^k\}$ be generated by the following procedure:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k), \quad k = 0, 1, \ldots.$$

Then for all $k > 1$, it holds that

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \le \frac{L}{2k}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Proof.

Note that $\mathbf{x}^{k+1}$ is the optimal solution of the convex problem:

$$\mathbf{x}^{k+1} = \arg\min \left\{ \Theta_k(\mathbf{x}) := f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k),\ \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 \right\}$$

and $\nabla \Theta_k(\mathbf{x}^{k+1}) = 0$ (by Theorem of Optimality for Convex Optimization). Moreover, $\Theta_k(\cdot)$ is quadratic and its second-order Taylor expansion is exact:

$$\begin{aligned}\Theta_k(\mathbf{x}) &= \Theta_k(\mathbf{x}^{k+1}) + \langle \nabla \Theta_k(\mathbf{x}^{k+1}),\ \mathbf{x} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{k+1}\|^2 \\ &= \Theta_k(\mathbf{x}^{k+1}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{k+1}\|^2.\end{aligned}$$

Since f is L-smooth, we have

$$\begin{aligned}f(\mathbf{x}^{k+1}) \le \Theta_k(\mathbf{x}^{k+1}) &= \Theta_k(\mathbf{x}) - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 \\ &= f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k),\ \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2.\end{aligned} \tag{1}$$

Setting $\mathbf{x} = \mathbf{x}^k$ in (1), we get

$$f(\mathbf{x}^{k+1}) \le f(\mathbf{x}^k) - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2,$$

showing that $\{f(\mathbf{x}^k)\}$ is nonincreasing.
Let $\mathbf{x} = \mathbf{x}^*$ in (1), we get

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k),\ \mathbf{x}^* - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2 \\ &\quad - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \\ &\le f(\mathbf{x}^*) + \frac{L}{2}\Big(\|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2\Big), \end{aligned}$$

where the last inequality is due to the convexity of f.

Hence

$$
\begin{aligned}
(k+1)[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)] &\leq \sum_{i=0}^{k} \Big(f(\mathbf{x}^{i+1} - f(\mathbf{x}^*)) \Big) \\
&\leq \frac{L}{2} \sum_{i=0}^{k} [\|\mathbf{x}^* - \mathbf{x}^i\|^2 - \|\mathbf{x}^{i+1} - \mathbf{x}^*\|^2] \\
&= \frac{L}{2} \Big(\|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \Big) \\
&\leq \frac{L}{2} \|\mathbf{x}^0 - \mathbf{x}^*\|^2.
\end{aligned}
$$

This completes the proof. $\square$

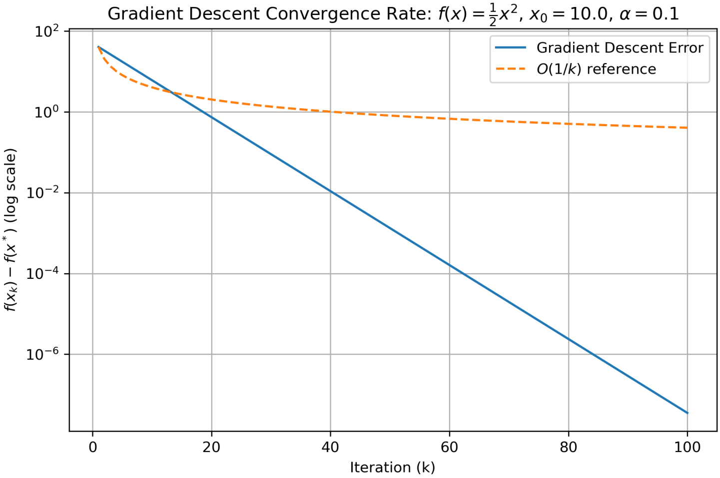


Figure: $O(1/k)$ Complexity of GD

Example: Piecewise-Linear-Quadratic Function (plq)

Consider one-dimensional optimization problem

$$\min \quad f(x),$$

where $f : \mathbb{R} \mapsto \mathbb{R}$ is given by

$$f(x) = \begin{cases} \frac{3(1-x)^2}{4} - 2(1-x) & \text{if } x > 1 \\ \frac{3(1+x)^2}{4} - 2(1+x) & \text{if } x < -1 \\ x^2 - 1 & \text{if } -1 \leq x \leq 1. \end{cases}$$

f is continuously differentiable everywhere (f is L-smooth and convex)

$$\nabla f(x) = \begin{cases} \frac{3x}{2} + \frac{1}{2} & \text{if } x > 1 \\ \frac{3x}{2} - \frac{1}{2} & \text{if } x < -1 \\ 2x & \text{if } -1 \leq x \leq 1. \end{cases}$$

Finite Termination of GD

We first note that the optimal solution is $x_* = 0$ and $L = 2$.

Let us use GD to solve the problem. We start with $x_0 = 2$.

Step 1: $x_0 = 2$, $\nabla f(x_0) = (3/2) \times 2 + 1/2 = 7/2$. Then

$$x_1 = x_0 - \frac{1}{L}\nabla f(x_0) = 2 - \frac{1}{2} \times \frac{7}{2} = \frac{1}{4}.$$

Step 2: $x_1 = 1/4$, $\nabla f(x_1) = 2 \times (1/4) = 1/2$. Then

$$x_2 = x_1 - \frac{1}{L}\nabla f(x_1) = \frac{1}{4} - \frac{1}{2} \times \frac{1}{2} = 0.$$

In 2 steps, we reached the optimal solution. This is known as finite termination. This only happens to some quadratic functions. In general, it would take infinitely many steps to observe the convergence.

Summary

- We considered the problem:

$$\min_{\mathbf{x}} \; f(\mathbf{x})$$

  where f is differentiable and often convex.

- GD:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k) = \mathbf{x}^k + \underbrace{\alpha}_{\text{steplength}} \times \underbrace{\Big(- \nabla f(\mathbf{x}^k) \Big)}_{\text{search direction}}$$

- Function class: L-Lipschitz and convex:

$$\alpha = \frac{\epsilon}{L^2}, \quad T = \frac{D^2 L}{\epsilon^2}$$

  and in $O(1/\epsilon^2)$ (i.e., T) iterations, GD can find ϵ-optimal solution:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{x}^t \quad \text{(uniform average of all iterates)}$$

Summary

- Function class: L-smooth and convex:

$$\alpha = \frac{1}{L}, \quad f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{L}{2k}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

  Let $T := LD^2/(2\epsilon)$. In other words, in T steps $(O(1/\epsilon))$, GD find an ϵ-optimal solution $\mathbf{x}^T$ such that

$$f(\mathbf{x}^T) \leq f(\mathbf{x}^*) + \epsilon.$$

- The complexity $O(1/\epsilon^2)$ vs $O(1/\epsilon)$: the latter is way faster than the former. For example, if we want a solution that is upto $\epsilon = 10^{-4}$ accuracy, the number of iterations take are:

$$10^8 \quad \text{vs} \quad 10^4.$$

- Note: the only difference between the two complexity results is the steplength choice (ϵ/L^2 vs $1/L$). This shows that the choice of α is extremely important (linesearch strategy).