Convergence of Gradient Descent: Smooth and Convex Settings

Avishek Ghosh
Assistant Professor, Dept. of CSE.,IIT Bombay

IIT Bombay

August 6, 2025

Outline

1. Gradient Descent and Decent lemma for Smooth functions
    - GD for Non-Convex Smooth Functions
    - GD for Convex and Smooth function
    - GD for Smooth and Strongly Convex functions

2. Projection

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent Update:

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent Update:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k > 0 \text{ (step size)}.$$

Problem:

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

.

Gradient Descent Update:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k > 0 \text{ (step size).}$$

- Choose α_k appropriately(for example, through line search methods)

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent Update:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k > 0 \text{ (step size)}.$$

- Choose α_k appropriately(for example, through line search methods)

Now from next slide ,we'll see about GD convergence results in the following cases:

- Smooth functions

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent Update:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k > 0 \text{ (step size)}.$$

- Choose α_k appropriately(for example, through line search methods)

Now from next slide ,we'll see about GD convergence results in the following cases:

- Smooth functions
- Smooth + Convex functions

Gradient Descent(GD) Formulation

Problem:

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

.

Gradient Descent Update:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k > 0 \text{ (step size)}.$$

- Choose α_k appropriately(for example, through line search methods)

Now from next slide ,we'll see about GD convergence results in the following cases:

- Smooth functions
- Smooth + Convex functions
- Smooth + Strongly Convex functions

Descent lemma for smooth function

We consider the unconstrained optimization problem

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

where:

Descent lemma for smooth function

We consider the unconstrained optimization problem

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

where:

- $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$.

Descent lemma for smooth function

We consider the unconstrained optimization problem

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

where:

- $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$.
- $f^* = \inf_x f(x) > -\infty$ is a global lower bound.

Descent lemma for smooth function

We consider the unconstrained optimization problem

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

where:

- $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$.
- $f^* = \inf_x f(x) > -\infty$ is a global lower bound.

Lemma (Descent Lemma)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be M-smooth. Using gradient descent

$$x^{k+1} = x^k - \frac{1}{M}\nabla f(x^k),$$

we have

Descent lemma for smooth function

We consider the unconstrained optimization problem

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

where:

- $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \le M\|x - y\|$.
- $f^* = \inf_x f(x) > -\infty$ is a global lower bound.

Lemma (Descent Lemma)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be M-smooth. Using gradient descent

$$x^{k+1} = x^k - \tfrac{1}{M}\nabla f(x^k),$$

we have

$$f(x^{k+1}) \le f(x^k) - \tfrac{1}{2M}\|\nabla f(x^k)\|^2 \le f(x^k).$$

Proof of Descent lemma

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|^2.$$

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

$$f(x^{k+1}) \leq f(x^k) - \langle \nabla f(x^k), \alpha^k \nabla f(x^k) \rangle + \frac{M}{2}\|\alpha^k \nabla f(x^k)\|^2$$

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \langle \nabla f(x^k), \alpha^k \nabla f(x^k) \rangle + \frac{M}{2} \|\alpha^k \nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{1}{M} \|\nabla f(x^k)\|^2 + \frac{M}{2} \frac{1}{M^2} \|\nabla f(x^k)\|^2 \end{aligned}$$

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \langle \nabla f(x^k), \alpha^k \nabla f(x^k) \rangle + \frac{M}{2}\|\alpha^k \nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{1}{M}\|\nabla f(x^k)\|^2 + \frac{M}{2}\frac{1}{M^2}\|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{1}{2M}\|\nabla f(x^k)\|^2 \end{aligned}$$

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \langle \nabla f(x^k), \alpha^k \nabla f(x^k) \rangle + \frac{M}{2}\|\alpha^k \nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{1}{M}\|\nabla f(x^k)\|^2 + \frac{M}{2}\frac{1}{M^2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{1}{2M}\|\nabla f(x^k)\|^2 \\
&\leq f(x^k).
\end{aligned} \tag{2}$$

Proof of Descent lemma

Since f is M-smooth, for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Setting $x = x^k$, $y = x^{k+1}$ and $\alpha^k = 1/M$ gives

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \langle \nabla f(x^k), \alpha^k \nabla f(x^k) \rangle + \frac{M}{2}\|\alpha^k \nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{1}{M}\|\nabla f(x^k)\|^2 + \frac{M}{2}\frac{1}{M^2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{1}{2M}\|\nabla f(x^k)\|^2 \\
&\leq f(x^k).
\end{aligned} \tag{2}$$

Which Complete the proof of Descent Lemma.

Case1: Non-Convex Smooth Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be* M*-smooth (but not necessarily convex).*

Case1: Non-Convex Smooth Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be* M*-smooth (but not necessarily convex). Running gradient descent with step size* $\alpha_k = \frac{1}{M}$

Case1: Non-Convex Smooth Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be* M*-smooth (but not necessarily convex). Running gradient descent with step size* $\alpha_k = \frac{1}{M}$

$$x^{k+1} = x^k - \frac{1}{M}\nabla f(x^k) \quad (k = 0, 1, \dots, T-1)$$

Case1: Non-Convex Smooth Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be M-smooth (but not necessarily convex). Running gradient descent with step size $\alpha_k = \frac{1}{M}$

$$x^{k+1} = x^k - \tfrac{1}{M}\nabla f(x^k) \quad (k = 0, 1, \ldots, T-1)$$

yields

$$\min_{0 \le k \le T-1} \|\nabla f(x^k)\| \;\le\; \sqrt{\frac{2M\big(f(x^0) - \bar{f}\big)}{T}},$$

where $\bar{f} = \inf_x f(x)$.

Case1: Non-Convex Smooth Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be* M*-smooth (but not necessarily convex). Running gradient descent with step size* $\alpha_k = \frac{1}{M}$

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

yields

$$\min_{0 \le k \le T-1} \|\nabla f(x^k)\| \le \sqrt{\frac{2M(f(x^0) - \bar{f})}{T}},$$

where $\bar{f} = \inf_x f(x)$*. Moreover,*

$$\lim_{T\to\infty} \|\nabla f(x^T)\| = 0.$$

Remarks

1. The above result is a first-order necessary condition (FONC) guarantee.

Remarks

1. The above result is a first-order necessary condition (FONC) guarantee.
2. In general for non-convex functions, this is the best one can establish for GD.

Remarks

1. The above result is a first-order necessary condition (FONC) guarantee.
2. In general for non-convex functions, this is the best one can establish for GD.
3. In the above theorem, we don't have any condition on the final iterate but

$$\lim_{T\to\infty} \|\nabla f(x^T)\| = 0$$

Proof

Use Decent lemma with $k = T - 1$,and so on..

Proof

Use Decent lemma with $k = T - 1$,and so on..

$$f(x^T) \le f(x^{T-1}) - \frac{1}{2M} \|\nabla f(x^{T-1})\|^2$$

Proof

Use Decent lemma with $k = T - 1$,and so on..

$$
\begin{aligned}
f(x^T) &\le f(x^{T-1}) - \frac{1}{2M}\left\|\nabla f(x^{T-1})\right\|^2 \\
&\le f(x^{T-2}) - \frac{1}{2M}\left\|\nabla f(x^{T-2})\right\|^2 - \frac{1}{2M}\left\|\nabla f(x^{T-1})\right\|^2
\end{aligned}
$$

Proof

Use Decent lemma with $k = T-1$,and so on..

$$
\begin{aligned}
f(x^T) &\le f(x^{T-1}) - \frac{1}{2M}\|\nabla f(x^{T-1})\|^2 \\
&\le f(x^{T-2}) - \frac{1}{2M}\|\nabla f(x^{T-2})\|^2 - \frac{1}{2M}\|\nabla f(x^{T-1})\|^2 \\
&\le \dots \\
&\le f(x^0) - \frac{1}{2M}\sum_{k=0}^{T-1}\|\nabla f(x^k)\|^2.
\end{aligned}
$$

Proof

Use Decent lemma with $k = T-1$,and so on..

$$
\begin{aligned}
f(x^T) &\le f(x^{T-1}) - \frac{1}{2M}\left\|\nabla f(x^{T-1})\right\|^2 \\
&\le f(x^{T-2}) - \frac{1}{2M}\left\|\nabla f(x^{T-2})\right\|^2 - \frac{1}{2M}\left\|\nabla f(x^{T-1})\right\|^2 \\
&\le \dots \\
&\le f(x^0) - \frac{1}{2M}\sum_{k=0}^{T-1}\left\|\nabla f(x^k)\right\|^2.
\end{aligned}
$$

Since $\bar{f} \le f(x^T)$, it follows

$$
\sum_{k=0}^{T-1}\left\|\nabla f(x^k)\right\|^2 \;\le\; 2M\big(f(x^0) - \bar{f}\big).
$$

(a) Asymptotic: Taking $T \to \infty$ gives

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0$.

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0$.

(b) Non-asymptotic: For finite T,

proof contd..

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0$.

(b) Non-asymptotic: For finite T,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T},$$

proof contd..

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0$.

(b) Non-asymptotic: For finite T,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T},$$

Since,minimum of non-negative sum term is upper-bounded by average sum, hence

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0$.

(b) Non-asymptotic: For finite T,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T},$$

Since,minimum of non-negative sum term is upper-bounded by average sum, hence

$$\min_{0\le k\le T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T}.$$

Remarks:

- Convergence rate: $\mathcal{O}(\frac{1}{\sqrt{T}})$.

(a) Asymptotic: Taking $T \to \infty$ gives $\lim_{T\to\infty} \|\nabla f(x^T)\|^2 = 0.$

(b) Non-asymptotic: For finite T,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T},$$

Since,minimum of non-negative sum term is upper-bounded by average sum, hence

$$\min_{0\le k\le T-1} \|\nabla f(x^k)\|^2 \;\le\; \frac{2M\big(f(x^0) - \bar{f}\big)}{T}.$$

Remarks:

- Convergence rate: $\mathcal{O}(\frac{1}{\sqrt{T}})$.
- In practice, choose a large T and apply early stopping.

Case 2 :GD for Convex & M-Smooth

Case 2 :GD for Convex & M-Smooth

Theorem (GD For Convex and M- Smooth)

f is M-smooth and convex. and x^* be a global minimizer of f. Running GD

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

yields

Case 2 :GD for Convex & M-Smooth

Theorem (GD For Convex and M- Smooth)

f is M-smooth and convex. and x^* be a global minimizer of f. Running GD

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

yields

$$f(x^T) - f(x^*) \ \le \ \frac{M}{2T}\,\|x^0 - x^*\|^2 \,.$$

Remarks:

- Convergence rate: $O(1/T)$.

Case 2 :GD for Convex & M-Smooth

Theorem (GD For Convex and M- Smooth)

f is M-smooth and convex. and x^* be a global minimizer of f. Running GD

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

yields

$$f(x^T) - f(x^*) \le \frac{M}{2T}\left\|x^0 - x^*\right\|^2 .$$

Remarks:

- Convergence rate: $O(1/T)$.
- Ensures global convergence to the minimizer.

Case 2 :GD for Convex & M-Smooth

Theorem (GD For Convex and M- Smooth)

f is M-smooth and convex. and x^* be a global minimizer of f. Running GD

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

yields

$$f(x^T) - f(x^*) \ \leq \ \frac{M}{2T}\,\|x^0 - x^*\|^2\,.$$

Remarks:

- Convergence rate: $O(1/T)$.
- Ensures global convergence to the minimizer.
- Guarantee on the final iterate, not on the gradient norm.

Case 2 :GD for Convex & M-Smooth

Theorem (GD For Convex and M- Smooth)

f is M-smooth and convex. and x^* be a global minimizer of f. Running GD

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

yields

$$f(x^T) - f(x^*) \ \le \ \frac{M}{2T}\,\|x^0 - x^*\|^2 \,.$$

Remarks:

- Convergence rate: $O(1/T)$.
- Ensures global convergence to the minimizer.
- Guarantee on the final iterate, not on the gradient norm.
- Requires differentiability; non-smooth cases call for subgradient methods.

Proof..

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \quad \text{(convexity)}$$

Proof..

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \quad \text{(convexity)}$$
$$f(x^k) \leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k),\, x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k),\, x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)}
\end{aligned}
$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)} \\
f(x^{k+1}) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2
\end{aligned}
$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)} \\
f(x^{k+1}) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \langle \nabla f(x^k), x^k - x^* \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2
\end{aligned}
$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)} \\
f(x^{k+1}) &\leq f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \langle \nabla f(x^k), x^k - x^* \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \tfrac{M}{2} \Big(\|x^k - x^*\|^2 - \|x^k - x^* - \tfrac{1}{M} \nabla f(x^k)\|^2 \Big)
\end{aligned}
$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k),\, x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k),\, x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)} \\
f(x^{k+1}) &\leq f(x^*) - \langle \nabla f(x^k),\, x^* - x^k \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \langle \nabla f(x^k),\, x^k - x^* \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \tfrac{M}{2} \left(\|x^k - x^*\|^2 - \|x^k - x^* - \tfrac{1}{M} \nabla f(x^k)\|^2 \right) \\
&= f(x^*) + \tfrac{M}{2} \left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right).
\end{aligned}
$$

Proof..

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k),\, x^* - x^k \rangle \quad \text{(convexity)} \\
f(x^k) &\leq f(x^*) - \langle \nabla f(x^k),\, x^* - x^k \rangle \\
f(x^{k+1}) &\leq f(x^k) - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \quad \text{(Descent Lemma)} \\
f(x^{k+1}) &\leq f(x^*) - \langle \nabla f(x^k),\, x^* - x^k \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \langle \nabla f(x^k),\, x^k - x^* \rangle - \tfrac{1}{2M} \|\nabla f(x^k)\|^2 \\
&= f(x^*) + \tfrac{M}{2} \left(\|x^k - x^*\|^2 - \|x^k - x^* - \tfrac{1}{M} \nabla f(x^k)\|^2 \right) \\
&= f(x^*) + \tfrac{M}{2} \left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right).
\end{aligned}
$$

Proof Contd..

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} \left(f(x^{k+1}) - f(x^*)\right) \ \leq$$

Proof Contd..

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} \left(f(x^{k+1}) - f(x^*)\right) \;\le\; \frac{M}{2}\left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2\right) \;\le$$

Proof Contd..

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} \left(f(x^{k+1}) - f(x^*) \right) \;\leq\; \frac{M}{2} \left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2 \right) \;\leq\; \frac{M}{2} \|x^0 - x^*\|^2.$$

Proof Contd..

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} \left(f(x^{k+1}) - f(x^*)\right) \;\leq\; \tfrac{M}{2}\left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2\right) \;\leq\; \tfrac{M}{2}\|x^0 - x^*\|^2.$$

Since $\{f(x^k)\}$ is non-increasing due to Decent lemma and hence any term will be bounded by avarage quantity,

Proof Contd..

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} (f(x^{k+1}) - f(x^*)) \;\leq\; \tfrac{M}{2}(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2) \;\leq\; \tfrac{M}{2}\|x^0 - x^*\|^2.$$

Since $\{f(x^k)\}$ is non-increasing due to Decent lemma and hence any term will be bounded by avarage quantity,

$$f(x^T) - f(x^*) \;\leq\; \frac{1}{T}\sum_{k=0}^{T-1}(f(x^k) - f(x^*)) \;\leq\; \frac{M}{2T}\|x^0 - x^*\|^2.$$

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Proof:

- since f is m-strongly convex

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Proof:

- since f is m-strongly convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2 .$$

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Proof:

- since f is m-strongly convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2 .$$

- Fix x and optimize over y then

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Proof:

- since f is m-strongly convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2 .$$

- Fix x and optimize over y then

$$\nabla f(x) + \frac{m}{2}(2(y - x)) = 0$$

Case 3:M-Smooth and m-Strongly Convex

Lemma

If f is m-strongly convex and let x^* is a minima of f then

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)) \tag{3}$$

Proof:

- since f is m-strongly convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2 .$$

- Fix x and optimize over y then

$$\nabla f(x) + \frac{m}{2}(2(y - x)) = 0$$
$$\Rightarrow y = x - \frac{1}{m} \nabla f(x).$$

Proof..

- Substitute the value of y in RHS of above equation and LHS in $f(x^*)$,

Proof..

- Substitute the value of y in RHS of above equation and LHS in $f(x^*)$,

$$f(x^*) \geq f(x) + \langle \nabla f(x), -\frac{1}{m} \nabla f(x) \rangle + \frac{m}{2} \left\| -\frac{1}{m} \nabla f(x) \right\|^2$$

Proof..

- Substitute the value of y in RHS of above equation and LHS in $f(x^*)$,

$$
\begin{aligned}
f(x^*) &\geq f(x) + \langle \nabla f(x), -\frac{1}{m}\nabla f(x) \rangle + \frac{m}{2}\left\| -\frac{1}{m}\nabla f(x) \right\|^2 \\
&= f(x) - \frac{1}{m}\langle \nabla f(x), \nabla f(x) \rangle + \frac{1}{2m}\left\| \nabla f(x) \right\|^2
\end{aligned}
$$

Proof..

- Substitute the value of y in RHS of above equation and LHS in $f(x^*)$,

$$\begin{aligned} f(x^*) &\geq f(x) + \langle \nabla f(x), -\frac{1}{m}\nabla f(x)\rangle + \frac{m}{2}\left\| -\frac{1}{m}\nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{m}\langle \nabla f(x), \nabla f(x)\rangle + \frac{1}{2m}\left\|\nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{2m}\left\|\nabla f(x)\right\|^2 \end{aligned}$$

Proof..

- Substitute the value of y in RHS of above equation and LHS in $f(x^*)$,

$$
\begin{aligned}
f(x^*) &\geq f(x) + \langle \nabla f(x), -\frac{1}{m}\nabla f(x)\rangle + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(x)\right\|^2 \\
&= f(x) - \frac{1}{m}\langle \nabla f(x), \nabla f(x)\rangle + \frac{1}{2m}\left\|\nabla f(x)\right\|^2 \\
&= f(x) - \frac{1}{2m}\left\|\nabla f(x)\right\|^2
\end{aligned}
$$

$$\Rightarrow \left\|\nabla f(x)\right\|^2 \geq 2m\big(f(x) - f(x^*)\big).$$

GD For M-Smooth and m-strongly convex

Theorem

Let f be M-smooth and m-strongly convex. We run GD with $\alpha_k = \frac{1}{M}$ then,

GD For M-Smooth and m-strongly convex

Theorem

Let f be M-smooth and m-strongly convex. We run GD with $\alpha_k = \frac{1}{M}$ then,

$$f(x^T) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^T (f(x^0) - f(x^*)). \tag{4}$$

Proof:

GD For M-Smooth and m-strongly convex

Theorem

Let f be M-smooth and m-strongly convex. We run GD with $\alpha_k = \frac{1}{M}$ then,

$$f(x^T) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^T (f(x^0) - f(x^*)). \tag{4}$$

Proof:

- Using Descent lemma:we get

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2M} \left\|\nabla f(x^k)\right\|^2$$

GD For M-Smooth and m-strongly convex

Theorem

Let f be M-smooth and m-strongly convex. We run GD with $\alpha_k = \frac{1}{M}$ then,

$$f(x^T) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^T (f(x^0) - f(x^*)). \tag{4}$$

Proof:

- Using Descent lemma:we get

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2M} \left\| \nabla f(x^k) \right\|^2$$

- Using previous lemma for m-strongly convex function

GD For M-Smooth and m-strongly convex

Theorem

Let f be M-smooth and m-strongly convex. We run GD with $\alpha_k = \frac{1}{M}$ then,

$$f(x^T) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^T (f(x^0) - f(x^*)). \quad (4)$$

Proof:

- Using Descent lemma:we get

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2M} \left\|\nabla f(x^k)\right\|^2$$

- Using previous lemma for m-strongly convex function ,we get

$$f(x^{k+1}) \leq f(x^k) - \frac{m}{M}(f(x^k) - f(x^*))$$

Proof cond..

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*)$$

Proof cond..

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*) \\
&= \left(1 - \frac{m}{M}\right)[f(x^k) - f(x^*)]
\end{aligned}
$$

Proof cond..

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) &\le f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*) \\
&= \left(1 - \frac{m}{M}\right)[f(x^k) - f(x^*)] \\
&\le \left(1 - \frac{m}{M}\right)^2 [f(x^{k-1}) - f(x^*)]
\end{aligned}
$$

Proof cond..

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) &\le f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*) \\
&= \left(1 - \frac{m}{M}\right)[f(x^k) - f(x^*)] \\
&\le \left(1 - \frac{m}{M}\right)^2 [f(x^{k-1}) - f(x^*)] \\
&\vdots
\end{aligned}
$$

Proof cond..

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &\le f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*) \\
&= \left(1 - \frac{m}{M}\right)[f(x^k) - f(x^*)] \\
&\le \left(1 - \frac{m}{M}\right)^2 [f(x^{k-1}) - f(x^*)] \\
&\vdots \\
&\le \left(1 - \frac{m}{M}\right)^k [f(x^0) - f(x^*)]
\end{aligned}$$

Proof cond..

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq f(x^k) - \frac{m}{M}(f(x^k) - f(x^*)) - f(x^*) \\
&= \left(1 - \frac{m}{M}\right)[f(x^k) - f(x^*)] \\
&\leq \left(1 - \frac{m}{M}\right)^2 [f(x^{k-1}) - f(x^*)] \\
&\quad \vdots \\
&\leq \left(1 - \frac{m}{M}\right)^k [f(x^0) - f(x^*)]
\end{aligned}
$$

1. In this case, we can give a guarantee on $\|x^T - x^*\|$. The choice of step-size in that case would be $\alpha_k = \frac{2}{M+m}$
2. Left as exercise!

Convergence Rates of Gradient Descent

Assumptions	Rate	Guarantee

Convergence Rates of Gradient Descent

Assumptions	Rate	Guarantee
f is M-smooth	$\min_{0\leq k<T} \|\nabla f(x_k)\|^2 \leq \frac{2M\left(f(x_0)-f^*\right)}{T}$	Best-iterate gradient norm

Convergence Rates of Gradient Descent

Assumptions	Rate	Guarantee
f is M-smooth	$\min_{0 \le k < T} \\|\nabla f(x_k)\\|^2 \le \frac{2M\left(f(x_0) - f^*\right)}{T}$	Best-iterate gradient norm
f is M-smooth and convex	$f(x_T) - f^* \le \frac{M}{2T} \\|x_0 - x^*\\|^2$	Final-iterate function gap

Convergence Rates of Gradient Descent

Assumptions	Rate	Guarantee
f is M-smooth	$\min_{0 \le k < T} \|\nabla f(x_k)\|^2 \le \frac{2M\left(f(x_0) - f^*\right)}{T}$	Best-iterate gradient norm
f is M-smooth and convex	$f(x_T) - f^* \le \frac{M}{2T} \|x_0 - x^*\|^2$	Final-iterate function gap
f is M-smooth, m-strongly convex	$f(x_T) - f^* \le \left(1 - \frac{m}{M}\right)^T \left(f(x_0) - f^*\right)$	Exponential (linear) convergence

Let Here, $\kappa = M/m$, $R = \|x_0 - x^*\|$, and $\Delta_0 = f(x_0) - f^*$.

Assumptions	**Criterion**	**Iterations**

Gradient Descent: Rates under Various Assumptions

Let Here, $\kappa = M/m$, $R = \|x_0 - x^*\|$, and $\Delta_0 = f(x_0) - f^*$.

Assumptions	**Criterion**	**Iterations**
f is m-strongly convex, M-smooth	$\|x_T - x^*\| \leq \varepsilon$	$O\big(\kappa \ln(R/\varepsilon)\big)$

Let Here, $\kappa = M/m$, $R = \|x_0 - x^*\|$, and $\Delta_0 = f(x_0) - f^*$.

Assumptions	**Criterion**	**Iterations**
f is m-strongly convex, M-smooth	$\|x_T - x^*\| \leq \varepsilon$	$O\Big(\kappa \ln(R/\varepsilon)\Big)$
f is convex, M-smooth	$f(x_T) - f^* \leq \varepsilon$	$O\Big(\frac{M\,R^2}{\varepsilon}\Big)$

Let Here, $\kappa = M/m$, $R = \|x_0 - x^*\|$, and $\Delta_0 = f(x_0) - f^*$.

Assumptions	Criterion	Iterations
f is m-strongly convex, M-smooth	$\|x_T - x^*\| \le \varepsilon$	$O\Big(\kappa \ln(R/\varepsilon)\Big)$
f is convex, M-smooth	$f(x_T) - f^* \le \varepsilon$	$O\Big(\frac{M\,R^2}{\varepsilon}\Big)$
f is nonconvex, M-smooth	$\min_{0 \le k < T} \|\nabla f(x_k)\| \le \varepsilon$	$O\Big(\frac{M\,\Delta_0}{\varepsilon^2}\Big)$

Constrained Optimization

Consider the following constrained optimization problem:

$$\min_{x \in \Omega} f(x) \tag{5}$$

Constrained Optimization

Consider the following constrained optimization problem:

$$\min_{x \in \Omega} f(x) \tag{5}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable and M-smooth function,

Constrained Optimization

Consider the following constrained optimization problem:

$$\min_{x \in \Omega} f(x) \tag{5}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable and M-smooth function, and $\Omega \subseteq \mathbb{R}^d$ is a closed convex set.

Constrained Optimization

Consider the following constrained optimization problem:

$$\min_{x \in \Omega} f(x) \tag{5}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable and M-smooth function, and $\Omega \subseteq \mathbb{R}^d$ is a closed convex set.

Euclidean projection

Euclidean projection

Let $\Omega \subseteq \mathbb{R}^d$ be a closed and convex set.

Euclidean projection

Let $\Omega \subseteq \mathbb{R}^d$ be a closed and convex set. The projection of a point $x \in \mathbb{R}^d$ onto Ω is defined as:

Euclidean projection

Let $\Omega \subseteq \mathbb{R}^d$ be a closed and convex set. The projection of a point $x \in \mathbb{R}^d$ onto Ω is defined as:

$$P_\Omega(x)$$

Euclidean projection

Let $\Omega \subseteq \mathbb{R}^d$ be a closed and convex set. The projection of a point $x \in \mathbb{R}^d$ onto Ω is defined as:

$$P_\Omega(x) := \arg\min_{z \in \Omega} \frac{1}{2}\|z - x\|^2 \tag{6}$$

Example 1:Non-Negative Orthant

Example 1:Non-Negative Orthant

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as the **non-negative orthant**:

Example 1:Non-Negative Orthant

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as the **non-negative orthant**:

$$\Omega := \left\{ x \in \mathbb{R}^d \,;\; x_i \geq 0, \quad \forall i = 1, \ldots, d \right\} \tag{7}$$

Example 1:Non-Negative Orthant

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as the **non-negative orthant**:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; x_i \geq 0, \quad \forall i = 1, \ldots, d \right\} \tag{7}$$

The projection of a vector $y \in \mathbb{R}^d$ onto Ω is given component-wise by:

Example 1:Non-Negative Orthant

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as the **non-negative orthant**:

$$\Omega := \left\{x \in \mathbb{R}^d \,;\; x_i \geq 0, \quad \forall i = 1, \ldots, d\right\} \tag{7}$$

The projection of a vector $y \in \mathbb{R}^d$ onto Ω is given component-wise by:

$$[P_\Omega(y)]_i = \max\{0, y_i\}, \quad \text{for each } i = 1, \ldots, d \tag{8}$$

Example 2:Box

Example 2:Box

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_\infty \leq 1 \right\} \tag{9}$$

Example 2:Box

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_\infty \leq 1 \right\} \tag{9}$$

Exercise:Given a point $y \in \mathbb{R}^d$, determine the projection $P_\Omega(y)$ onto Ω.

L_2 Norm Ball

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{x \in \mathbb{R}^d \,;\; \|x\|_2 \leq 1\right\} \tag{10}$$

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$,

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$, then $P_\Omega(y) = y$

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$, then $P_\Omega(y) = y$
- If $\|y\|_2 > 1$,

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$, then $P_\Omega(y) = y$
- If $\|y\|_2 > 1$,then $P_\Omega(y) = \dfrac{y}{\|y\|_2}$

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \; ; \; \|x\|_2 \leq 1 \right\} \tag{10}$$

The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$, then $P_\Omega(y) = y$
- If $\|y\|_2 > 1$,then $P_\Omega(y) = \dfrac{y}{\|y\|_2}$

Contraction/Non-expansive Property of Projection:

L_2 Norm Ball

Consider the set $\Omega \subseteq \mathbb{R}^d$ defined as:

$$\Omega := \left\{ x \in \mathbb{R}^d \ ; \ \|x\|_2 \leq 1 \right\} \tag{10}$$

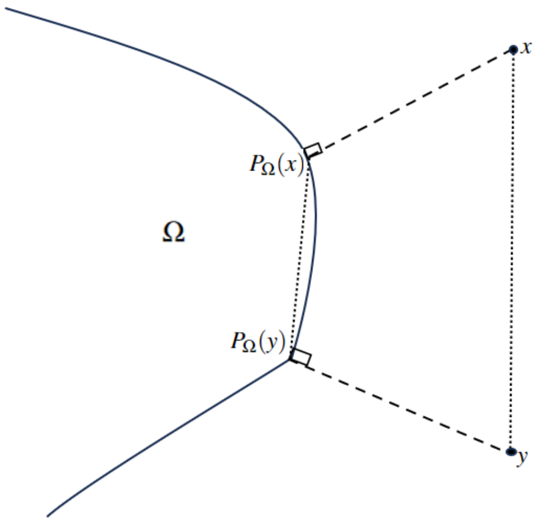
The projection $P_\Omega(y)$ onto Ω is given by:

- If $\|y\|_2 \leq 1$, then $P_\Omega(y) = y$
- If $\|y\|_2 > 1$,then $P_\Omega(y) = \dfrac{y}{\|y\|_2}$

Contraction/Non-expansive Property of Projection:

$$\boxed{\|P_\Omega(x) - P_\Omega(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n} \tag{11}$$

Figure: Illustration of Contraction Property



Convergence of GD for Constrained Optimization

Projected Gradient Descent:

$$x^{k+1} = P_{\Omega}[x^k - \alpha_k \nabla f(x_k)]$$

1. Thanks to the Contraction property of Projection, the distance gets smaller after projection
2. The GD algorithm remains the same with the projection step
3. The convergence rates remain the same

References I

[1] Benjamin Recht and Stephen J. Wright.
Optimization for Data Analysis. 2022.

[2] Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.

[3] Jorge Nocedal and Stephen J. Wright.
Numerical Optimization, 2nd Edition.
Springer, 2006.

Thank You!