The Gradient Method and Convergence Analysis

Presented on August 2, 2021

Outline

The Gradient Descent Method

Convergence under Convexity

Convergence under Smoothness

Convergence under Convexity and Smoothness

The Gradient Descent Method

The optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \; f^N(\theta),$$

$f : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

Gradient iteration:

$$\theta^{k+1} = \theta^k - \gamma \cdot \nabla f^N(\theta^k).$$

Questions:

- where: $\theta^k \to ?$
- when: rate of convergence $\theta^k \to \theta^\infty$
- why?

Two Important Classes of Functions: Convex Functions

- Definition:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

- First order condition:

$$f(x) \geq f(y) + \nabla f(y)^\top (x-y)$$

- Second order condition:

$$\nabla^2 f(x) \succeq 0.$$

Figure

[DIY] Prove equivalence of the statements.

Y. Nesterov, "Lectures on Convex Optimization." [Thm. 2.1.2]

Two Important Classes of Functions: Smooth Functions

- Definition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

- Descent Lemma:

$$|f(x) - f(y) + \nabla f(y)^\top (x - y)| \leq \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Figure

$$\boxed{|f(x) - f(y) + \nabla f(y)^\top (x-y)| \le \frac{L}{2}\|x-y\|^2, \quad \forall x, y \in \mathbb{R}^d.}$$

Proof. Using Taylor expansion:

$$f(x) = f(y) + \int_0^1 \nabla f\big(x + t(y-x)\big)^\top (x-y)\, dt.$$

Compare terms

$$\begin{aligned}
&|f(x) - f(y) + \nabla f(y)^\top (x-y)| \\
=&\left|\int_0^1 \nabla f\big(x + t(y-x)\big)^\top (x-y)\, dt - \nabla f(y)^\top (x-y)\right| \\
\le&\int_0^1 \left|\big(\nabla f\big(x + t(y-x)\big) - \nabla f(y)\big)^\top (x-y)\right| dt \quad \text{(why?)} \\
\le&\int_0^1 tL\|x-y\|^2 dt = \frac{L}{2}\|x-y\|^2.
\end{aligned}$$

□

First order expansion provides a good local approximation of f.

figure

Outline

The Gradient Descent Method

Convergence under Convexity

Convergence under Smoothness

Convergence under Convexity and Smoothness

Optimality Condition

The optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \; f^N(\theta),$$

$f : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

Global minimizer: $f(\theta^\star) \leq f(\theta)$ for all θ.

Local minimizer: $f(\theta^*) \leq f(\theta)$ for $\theta \in \mathcal{N}(\theta^*)$.

First order necessary condition

If θ^* is a local minimizer and f^N is continuously differentiable in an open neighborhood of θ^*, then $\nabla f^N(\theta^*) = 0$. [proof by contradiction]

Convexity: Local $\Leftrightarrow$ Global

Convergence Analysis

Optimality (stationarity) measures $M(\theta)$

- nonconvex: $\|\nabla f(\theta)\|$
- convex: $\|\theta - \theta^\star\|$, $f(\theta^k) - f^\star$.

Asymptotic convergence

Let $\{\theta^k\}_{k\in\mathbb{N}}$ be a sequence generated by an "algorithm", then $\lim_{k\to\infty} M(\theta^k) = 0$.

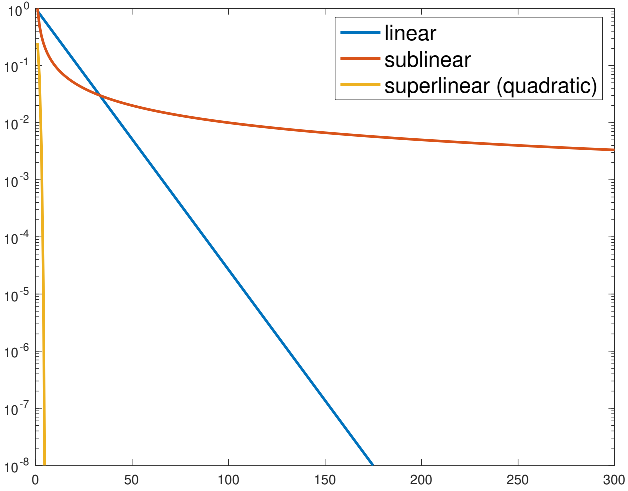
Rate of convergence: how fast?

Linear rate: $\lim\limits_{k\to\infty} \frac{M(\theta^{k+1})}{M(\theta^k)} = r < 1.$ $\quad \log M(\theta^{k+1}) \leq \log M(\theta^k) + \log r$

Sublinear rate: $\lim\limits_{k\to\infty} \frac{M(\theta^{k+1})}{M(\theta^k)} = 1.$

Superlinear rate: $\lim\limits_{k\to\infty} \frac{M(\theta^{k+1})}{M(\theta^k)} = 0.$

Order q convergence: $\lim\limits_{k\to\infty} \frac{M(\theta^{k+1})}{M(\theta^k)^q} \leq \text{const}.$


linear
sublinear
superlinear (quadratic)
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
0
50
100
150
200
250
300

Convex Functions

Implication of convexity:

$$\nabla f(\theta)^\top (\theta - \theta^\star) \geq f(\theta) - f^\star \geq 0$$

- $-\nabla f(\theta)$ is positively correlated to $\theta^\star - \theta$
- moving along $-\nabla f(\theta)$ direction gets closer to $\theta^\star$

Compute distance to $\theta^\star$:

$$\begin{aligned}
\|\theta^{k+1} - \theta^\star\|^2 &= \|\theta^k - \gamma \nabla f(\theta^k) - \theta^\star\|^2 \\
&= \|\theta^k - \theta^\star\|^2 - 2\gamma \nabla f(\theta^k)^\top (\theta^k - \theta^\star) + \gamma^2 \|\nabla f(\theta^k)\|^2 \\
&\leq \|\theta^k - \theta^\star\|^2 - 2\gamma (f(\theta^k) - f^\star) + \gamma^2 \|\nabla f(\theta^k)\|^2
\end{aligned}$$

Polyak's step size: $\gamma = \frac{f(\theta^k) - f^\star}{\|\nabla f(\theta^k)\|^2}$

$$\|\theta^{k+1} - \theta^\star\|^2 \leq \|\theta^k - \theta^\star\|^2 - \frac{(f(\theta^k) - f^\star)^2}{\|\nabla f(\theta^k)\|^2}.$$

Theorem

Let f be convex with bounded gradient, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma^k = \frac{f(\theta^k)-f^\star}{\|\nabla f(\theta^k)\|^2}$ satisfies

$$\min_{k\in[T]} f(\theta^k) - f^\star \leq \frac{B\|\theta^0 - \theta^\star\|}{\sqrt{T+1}}.$$

Proof. $\|\theta^{k+1} - \theta^\star\|^2 \leq \|\theta^k - \theta^\star\|^2 - \dfrac{(f(\theta^k) - f^\star)^2}{B^2}$.

Then

$$\sum_{k=0}^{T}(f(\theta^k) - f^\star)^2 \leq B^2\left(\|\theta^0 - \theta^\star\|^2 - \|\theta^{k+1} - \theta^\star\|^2\right)$$

□

Alternative Proof

Fixed step size γ:

$$\begin{aligned}\|\theta^{k+1}-\theta^\star\|^2 &\le \|\theta^k-\theta^\star\|^2 - 2\gamma(f(\theta^k)-f^\star) + \gamma^2\|\nabla f(\theta^k)\|^2 \\ &\le \|\theta^k-\theta^\star\|^2 - 2\gamma(f(\theta^k)-f^\star) + \gamma^2 B^2\end{aligned}$$

Regret interpretation: $f(\theta^k)-f^\star$ is large $\Rightarrow$ θ^{k+1} gets closer to $\theta^\star$

Rearranging terms

$$f(\theta^k)-f^\star \le \frac{1}{2\gamma}\left(\|\theta^k-\theta^\star\|^2 - \|\theta^{k+1}-\theta^\star\|^2 + \gamma^2 B^2\right)$$

Arrive at

$$\min_{k\in[T]} f(\theta^k)-f^\star \le \frac{1}{2(T+1)}\left(\frac{1}{\gamma}\|\theta^0-\theta^\star\|^2 + \gamma(T+1)B^2\right)$$

Optimal $\gamma^\star = \frac{\|\theta^0-\theta^\star\|}{\sqrt{T+1}B}$.

Outline

The Gradient Descent Method

Convergence under Convexity

Convergence under Smoothness

Convergence under Convexity and Smoothness

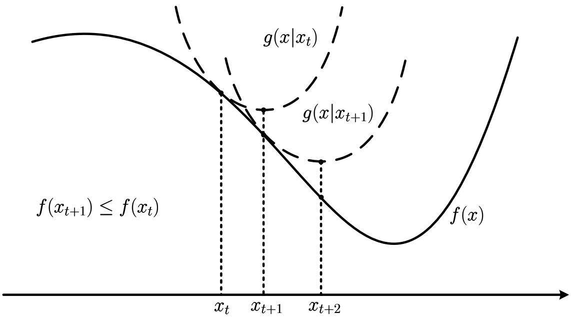
Smooth functions

- Definition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

- Descent Lemma:

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



Quadratic Upperbound - A Majorization Minimization Perspective

- GD:

$$\begin{aligned}\theta^{k+1} &= \theta^k - \gamma\nabla f(\theta^k)\\ &= \underset{\theta}{\operatorname{argmin}} \; \underbrace{\left\{ f(\theta^k)^\top(\theta-\theta^k) + \frac{1}{2\gamma}\|\theta-\theta^k\|^2 \right\}}_{L(\theta\,|\,\theta^k)} \text{ [verify it]}\end{aligned}$$

- Choose $\gamma \le 1/L$: $L(\theta\,|\,\theta^k) \ge f(\theta)$

Proof of Descent

Majorize: by descent lemma and $\gamma \le 1/L$

$$\begin{aligned} f(\theta) &\le f(\theta^k) + \nabla f(\theta^k)^\top(\theta - \theta^k) + \frac{L}{2}\|\theta - \theta^k\|^2 \\ &\le f(\theta^k) + \nabla f(\theta^k)^\top(\theta - \theta^k) + \frac{1}{2\gamma}\|\theta - \theta^k\|^2 \end{aligned}$$

Minimize: let $\theta = \theta^k - \gamma\nabla f(\theta^k)$

$$\begin{aligned} f(\theta^{k+1}) &\le f(\theta^k) - \gamma\|\nabla f(\theta^k)\|^2 + \frac{\gamma}{2}\|\nabla f(\theta^k)\|^2 \\ &= f(\theta^k) - \frac{\gamma}{2}\|\nabla f(\theta^k)\|^2. \end{aligned}$$

In fact, we can prove decay of $f(\theta^k)$ for $\gamma < 2/L$: [DIY]

$$f(\theta^{k+1}) \le f(\theta^k) - \gamma(1 - \frac{\gamma L}{2})\|\nabla f(\theta^k)\|^2$$

Outline

The Gradient Descent Method

Convergence under Convexity

Convergence under Smoothness

Convergence under Convexity and Smoothness

Upper and Lower Bounds

Quadratic upperbound by smoothness:

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

Linear lowerbounds by convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y).$$

Two implications:

- $f(\theta^{k+1}) \leq f(\theta^k) - \gamma(1 - \frac{\gamma L}{2})\|\nabla f(\theta^k)\|^2$
- $\nabla f(\theta^k)^\top (\theta^k - \theta^\star) \geq f(\theta^k) - f^\star$.

Convergence Analysis - Convex Smooth Functions

Distance to optimal point $\theta^\star$:

$$\begin{aligned}\|\theta^{k+1} - \theta^\star\|^2 &= \|\theta^k - \gamma \nabla f(\theta^k) - \theta^\star\|^2 \\ &= \|\theta^k - \theta^\star\|^2 \underbrace{-2\gamma \nabla f(\theta^k)^\top (\theta^k - \theta^\star)}_{\text{convexity}} + \underbrace{\gamma^2 \|\nabla f(\theta^k)\|^2}_{\text{smoothness}}\end{aligned}$$

Convexity:

$$-2\gamma \nabla f(\theta^k)^\top (\theta^k - \theta^\star) \leq -2\gamma \left(f(\theta^k) - f^\star \right).$$

Smoothness:

$$\gamma^2 \|\nabla f(\theta^k)\|^2 \leq \frac{\gamma}{(1 - \frac{\gamma L}{2})} \left(f(\theta^k) - f(\theta^{k+1}) \right) \overset{(\gamma L \leq 1)}{\leq} 2\gamma \left(f(\theta^k) - f(\theta^{k+1}) \right).$$

Combining

$$\|\theta^{k+1} - \theta^\star\|^2 \leq \|\theta^k - \theta^\star\|^2 - 2\gamma (f(\theta^{k+1}) - f^\star).$$

Convergence Analysis - Convex Smooth Functions

Theorem

Let f be convex and L-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma \le 1/L$ satisfies

$$f(\theta^T) - f^\star \le \frac{\|\theta^0 - \theta^\star\|}{2\gamma T}.$$

Proof.

$$\sum_{k=0}^{T-1} f(\theta^{k+1}) - f^\star \le \frac{1}{2\gamma}\left(\|\theta^0 - \theta^\star\|^2 - \|\theta^k - \theta^\star\|^2\right).$$

Plus monotonicity $f(\theta^{k+1}) \le f(\theta^k)$ completes the proof. □

Strong Convexity

- Definition:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$$

- First order condition:

$$f(x) \geq f(y) + \nabla f(y)^\top (x-y) + \frac{\mu}{2}\|x-y\|^2$$

- Second order condition:

$$\nabla^2 f(x) \succeq \mu I.$$

Figure

[DIY] Show equivalence

Upper and Lower bounds

Quadratic upperbound by smoothness:

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

Quadratic lowerbound by convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{\mu}{2}\|x - y\|^2.$$

Two implications:

- Same descent inequality: $f(\theta^{k+1}) \leq f(\theta^k) - \gamma(1 - \frac{\gamma L}{2})\|\nabla f(\theta^k)\|^2$
- Improved lower bound:

$$\nabla f(\theta^k)^\top (\theta^k - \theta^\star) \geq f(\theta^k) - f^\star + \frac{\mu}{2}\|\theta^k - \theta^\star\|^2$$

.

Convergence Analysis - Strongly Convex Smooth Functions

Distance to optimal point $\theta^\star$:

$$\begin{aligned}\|\theta^{k+1} - \theta^\star\|^2 &= \|\theta^k - \gamma\nabla f(\theta^k) - \theta^\star\|^2 \\ &= \|\theta^k - \theta^\star\|^2 \underbrace{-2\gamma\nabla f(\theta^k)^\top(\theta^k - \theta^\star)}_{\text{s-cvx}} + \underbrace{\gamma^2\|\nabla f(\theta^k)\|^2}_{\text{smoothness}}\end{aligned}$$

Strong convexity:

$$-2\gamma\nabla f(\theta^k)^\top(\theta^k - \theta^\star) \leq -2\gamma\left(f(\theta^k) - f^\star\right) - \mu\gamma\|\theta^k - \theta^\star\|^2.$$

Smoothness:

$$\gamma^2\|\nabla f(\theta^k)\|^2 \leq \frac{\gamma}{(1 - \frac{\gamma L}{2})}\left(f(\theta^k) - f(\theta^{k+1})\right) \overset{(\gamma L \leq 1)}{\leq} 2\gamma\left(f(\theta^k) - f(\theta^{k+1})\right).$$

Combining

$$\|\theta^{k+1} - \theta^\star\|^2 \leq (1 - \mu\gamma)\|\theta^k - \theta^\star\|^2 - 2\gamma(f(\theta^{k+1}) - f^\star).$$

Convergence Analysis - Strongly Convex Smooth Functions

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 1/L$ satisfies

$$\|\theta^{k+1} - \theta^\star\|^2 \leq \frac{1-\mu\gamma}{1+\mu\gamma}\|\theta^k - \theta^\star\|^2.$$

Proof. To complete the proof we lowerbound $f(\theta^{k+1}) - f^\star$ using strong convexity

$$f(\theta^{k+1}) \geq f(\theta^\star) + \nabla f(\theta^\star)^\top(\theta^{k+1} - \theta^\star) + \frac{\mu}{2}\|\theta^\star - \theta^\star\|^2.$$

Hence

$$\|\theta^{k+1} - \theta^\star\|^2 \leq (1-\mu\gamma)\|\theta^k - \theta^\star\|^2 - \mu\gamma\|\theta^{k+1} - \theta^\star\|^2$$

□

Alternative Proof From Descent Perspective

Gradient dominance: there exists constant $c > 0$ such that

$$\|\nabla f(\theta)\|^2 \geq c(f(\theta) - f^\star)$$

Strong convexity implies gradient dominance with $c = 2\mu$.

Proof:

$$\begin{aligned} f(\theta^\star) &\geq f(\theta) + \nabla f(\theta)^\top(\theta^\star - \theta) + \frac{\mu}{2}\|\theta^\star - \theta\|^2 \\ &= f(\theta) + \frac{\mu}{2}\left\|\theta^\star - \theta + \frac{1}{\mu}\nabla f(\theta)\right\|^2 - \frac{1}{2\mu}\|\nabla f(\theta)\|^2 \end{aligned}$$

In English: small gradient implies closeness to $\theta^\star$

NB: compare $f(\theta) = \theta^2$ and $f(\theta) = \theta^4$, θ^4 is super flat in the valley and θ can be far away from 0 even when $f'(\theta)$ is small.

Cont.

Recall that from descent lemma

$$\begin{aligned} f(\theta^{k+1}) &\leq f(\theta^k) + \nabla f(\theta^k)^\top (\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2 \\ &= f(\theta^k) - \gamma \cdot \|\nabla f(\theta^k)\|^2 + \frac{\gamma^2 L}{2}\|\nabla f(\theta^k)\|^2. \end{aligned}$$

Apply the gradient dominance property

$$\begin{aligned} f(\theta^{k+1}) &\leq f(\theta^k) - \gamma\left(1 - \frac{\gamma L}{2}\right)\|\nabla f(\theta^k)\|^2 \\ &\leq f(\theta^k) - \gamma\left(1 - \frac{\gamma L}{2}\right) 2\mu(f(\theta^k) - f^\star). \end{aligned}$$

Subtracting $f^\star$ from both sides completes the proof.

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 2/L$ satisfies

$$f(\theta^{k+1}) - f^\star \leq \left(1 - 2\mu\gamma\big(1 - \frac{\gamma L}{2}\big)\right)(f(\theta^k) - f^\star)$$

- Larger step size range $\gamma < 2/L$
- $\gamma = 1/L$ gives the fastest rate

HW: Can you prove sublinear rate for convex f in terms of the objective value?