Convex Optimization and Gradient Descent

Ruixin Guo

Department of Computer Science
Kent State University

May 19, 2023

Contents

1. Properties of Convex Function

2. Gradient Descent

Contents

1. Properties of Convex Function

2. Gradient Descent

Notations

Jacobian: Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^p$ be differentiable, and $x \in \mathbb{R}^d$. Then $D\mathcal{F}(x)$ is the Jacobian of $\mathcal{F}$ at x, which is the matrix defined by its first derivatives:

$$[D\mathcal{F}(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x), \qquad \text{for } i = 1, ..., p, \quad j = 1, ..., d.$$

where we write $\mathcal{F} = (f_1(x), f_2(x), ..., f_p(x))$. Consequently $D\mathcal{F}(x)$ is a $p \times d$ matrix.

Gradient: Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, then $Df(x) \in \mathbb{R}^{1\times d}$ is a row vector. The gradient of f is defined as the transpose of $Df(x)$:

$$\nabla f(x) = Df(x)^T = [\frac{\partial f}{\partial x_1} f(x), \frac{\partial f}{\partial x_2} f(x), ..., \frac{\partial f}{\partial x_d} f(x)]^T$$

which is a column vector.

Hessian: If $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, and $x \in \mathbb{R}^d$, then $\nabla^2 f(x)$ is the Hessian of f at x, which is the matrix defined by its second-order partial derivatives:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \qquad \text{for } i, j = 1, ..., d.$$

Consequently $\nabla^2 f(x)$ is a $d \times d$ matrix.

Notations

Norm: The norm $\| \cdot \|$ is defined as Euclidean norm, i.e., for $x = [x_1, x_2, ..., x_d]^T$, $\|x\| = \sqrt{x_1^2 + x_2^2 + ..., x_d^2}$

Inner Product: Let $x, y \in \mathbb{R}^d$, the inner product of x and y is defined as $\langle x, y \rangle = x^T y$.

Convex Problem

Convex Problem: A convex problem is an optimization problem of the form

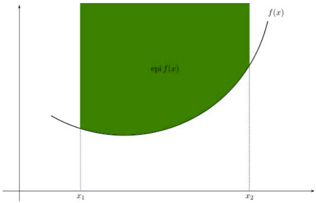
$$\min_{x \in C} f(x)$$

where f and C are convex.

Convex Set: If for any $x, y \in C$, $\lambda x + (1-\lambda)y \in C$ for any $\lambda \in [0,1]$, then C is a convex set.

Convex Function: Let $C \subseteq \mathbb{R}^d$ be a convex set, a function $f : C \to \mathbb{R}$ is convex if for all $x, y \in C$ and $0 \leq \lambda \leq 1$

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



Intuitively, define the epigraph of f:

$$\text{epi}(f) = \{(x,t) \in C \times \mathbb{R} \,|\, t \geq f(x)\}$$

Then f is a convex function means the epigraph epi(f) is a convex set.

Convex Function

Theorem 1: Let $C \subseteq \mathbb{R}^d$, and let the function $f : C \to \mathbb{R}$ be convex on C. Let x, y be any two elements in C. Then the following two conditions are equivalent:
(1) $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for any $\lambda \in [0,1]$
(2) $f(y) + \nabla f(y)^T(x-y) \leq f(x)$

Proof: $(1) \Rightarrow (2)$, (1) means

$$f(\lambda x + (1-\lambda)y) \leq f(y) + \lambda(f(x) - f(y))$$
$$f(x) - f(y) \geq \frac{f(y + \lambda(x-y)) - f(y)}{\lambda}$$

Since $\lim_{\lambda \to 0} \frac{f(y+\lambda(x-y))-f(y)}{\lambda(x-y)}(x-y) = \nabla f(y)^T(x-y)$, we get (2).
$(2) \Rightarrow (1)$, let $z = \lambda x + (1-\lambda)y$, by (2) we have $f(x) \geq f(z) + \nabla f(z)^T(x-z)$ and $f(y) \geq f(z) + \nabla f(z)^T(y-z)$. Multiplying the first inequality by λ and the second inequality by $1-\lambda$ we get

$$\lambda f(x) + (1-\lambda)f(y) \geq f(z) + \nabla f(z)^T(\lambda x + (1-\lambda)y - z) = f(z)$$

We get (1).

Lipschitz Smooth

Definition (Lipschitz Continuous): A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is Lipschitz continuous if for all $x, y \in \mathbb{R}^d$

$$\|f(x) - f(y)\| \leq K\|x - y\|$$

where K is referred to as Lipschitz constant.

If the gradient of f is Lipschitz continuous, we call f Lipschitz smooth.

Definition (Lipschitz Smooth): A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L-smooth if for all $x, y \in \mathbb{R}^d$

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

Convex and Smooth Function

Lemma 1: If $f : C \to \mathbb{R}$ is L-smooth and $C \subseteq \mathbb{R}^d$ is a convex set, then for all $x, y \in C$

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

Proof: Since C is convex, $x, y \in C$ implies the segment $[x, y] = \{z = \lambda x + (1 - \lambda) y \,|\, 0 \leq \lambda \leq 1\}$ is contained within C. Let $t \in [0, 1]$, then $x + t(y - x)$ is within C.

Define the function $\phi : [0, 1] \to \mathbb{R}$ as

$$\phi(t) = f(x + t(y - x))$$

Then $\phi(0) = f(x)$, $\phi(1) = f(y)$, $\phi'(t) = \nabla f(x + t(y - x))^T (y - x)$. Thus

$$\begin{aligned}
\phi'(t) - \phi'(0) &\leq |\phi'(t) - \phi'(0)| \\
&= |(\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x)| \\
&\leq \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| && \text{[Cauchy Inequality]} \\
&\leq L \|t(y - x)\| \|y - x\| && \text{[L-Smooth]} \\
&\leq tL \|y - x\|^2
\end{aligned}$$

Convex and Smooth Function

Thus,

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \phi(1) - \phi(0) - \phi'(0) \\ &= \int_0^1 (\phi'(t) - \phi'(0)) dt \\ &\leq \int_0^1 tL\|y - x\|^2 dt \\ &= \frac{L}{2}\|y - x\|^2 \end{aligned}$$

Rate and Order of Convergence

We use Rate of Convergence and Order of Convergence to theoretically evaluate how fast a convergence algorithm can be.

Definition: Let the sequence $\{x_k\}$ converge to x^*. Let $\|x_k - x^*\|$ be the error between x_k and x^*. If there exists a number $p > 1$ and a constant $r > 0$ (with $r < 1$ if $p = 1$) such that

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = r$$

We call p the Order of Convergence and r the Rate of Convergence.

- If $p = 1$ and $0 < r < 1$, the convergence is linear. Especially, when $r = 1$, we call the convergence sublinear (See [6]).
- If $p = 2$, the convergence is quadratic.
- If $p = 3$, the convergence is cubic.

Contents

1. Properties of Convex Function

2. Gradient Descent

Gradient Descent

Let $C \subseteq \mathbb{R}^d$ be a convex set, and let the $f : C \to \mathbb{R}$ be a convex function. The Gradient Descent algorithm is as follows:

- Choose $x_0 \in \mathbb{R}^d$ and step size $t > 0$.
- For $i = 0, 1, 2, ...$, define

$$x_{i+1} = x_i - t\nabla f(x_i)$$

Convergence Analysis of Gradient Descent

Theorem 2: Let $C \subseteq \mathbb{R}^d$ be a convex set, $f : C \to \mathbb{R}$ be a L-smooth convex function and $x^* = \arg\min_x f(x)$. Then the Gradient Descent Iteration

$$x_{i+1} = x_i - t\nabla f(x_i) \tag{1}$$

with step size $t \leq 1/L$ satisfies the following:

$$f(x_k) \leq f(x^*) + \frac{\|x_0 - x^*\|^2}{2tk}$$

Proof: Since f is convex, by Theorem 1,

$$f(x_i) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle \tag{2}$$

Since f is L-smooth, by Lemma 1

$$\begin{aligned} f(x_{i+1}) &\leq f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{L}{2}\|x_{i+1} - x_i\|^2 \\ &= f(x_i) - t\|\nabla f(x_i)\|^2 + \frac{Lt^2}{2}\|\nabla f(x_i)\|^2 && \text{[by (1)]} \\ &= f(x_i) - t(1 - \frac{Lt}{2})\|\nabla f(x_i)\|^2 \\ &\leq f(x_i) - \frac{t}{2}\|\nabla f(x_i)\|^2 && [Lt \leq 1] \qquad (3) \end{aligned}$$

Convergence Analysis of Gradient Descent

Remember that by (1) we have $\nabla f(x_i) = (1/t)(x_i - x_{i+1})$. Plugging (1) in (2) we get

$$\begin{aligned} f(x_{i+1}) &\leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle - \frac{t}{2}\|\nabla f(x_i)\|^2 \\ &= f(x^*) + \frac{1}{2t}\|x_i - x^*\|^2 - \frac{1}{2t}(\|x_i - x^*\|^2 - 2\langle t\nabla f(x_i), x_i - x^* \rangle + \|t\nabla f(x_i)\|^2) \\ &= f(x^*) + \frac{1}{2t}\|x_i - x^*\|^2 - \frac{1}{2t}\|x_i - x^* - t\nabla f(x_i)\|^2 \\ &= f(x^*) + \frac{1}{2t}(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2) \end{aligned}$$

Thus,

$$\sum_{i=0}^{k-1}(f(x_{i+1}) - f(x^*)) \leq \frac{1}{2t}(\|x_0 - x^*\|^2 - \|x_{i+1} - x^*\|^2) \leq \frac{\|x_0 - x^*\|^2}{2t}$$

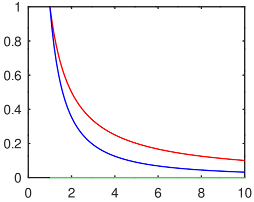
By (3), the sequence $f(x_0), f(x_1), f(x_2), \ldots$ is non-increasing, thus

$$k(f(x_k) - f(x^*)) \leq \sum_{i=0}^{k-1}(f(x_{i+1}) - f(x^*)) \leq \frac{\|x_0 - x^*\|^2}{2t}$$

$$f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2tk}$$

Convergence Speed of Gradient Descent in Theorem 2

Claim: In the worst case the convergence speed of the gradient descent algorithm in Theorem 2 is sublinear.



In the left figure, both blue and red line converge to green line. The red line is the upper bound of blue line, which guarantees the worst case convergence speed. The blue line converges faster than red line on average because it is always below red line.

Since the upper bound of sequence $\{f(x_k) - f(x^*)\}$ satisfies

$$\lim_{k\to\infty} \frac{\sup\{f(x_{k+1}) - f(x^*)\}}{\sup\{f(x_k) - f(x^*)\}} = \lim_{k\to\infty} \left(\frac{\|x_0 - x^*\|^2}{2t(k+1)} \Big/ \frac{\|x_0 - x^*\|^2}{2tk} \right) = 1$$

So if only the Lipschitz smooth condition is guaranteed for f, in the worst case the gradient decent will converge sublinearly, which is very slow.

Next time we will show that if both Lipschitz smooth and strong convexity conditions are guaranteed for f, the convergence of gradient decent will be improved to a linear one.

References

[1] Theory of Convex Functions. https://www.princeton.edu/~aaa/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf.

[2] Notes on convergence of gradient descent. https://raghumeka.github.io/CS289ML/gdnotes.pdf.

[3] Handbook of Convergence Theorems for (Stochastic) Gradient Methods. https://gowerrobert.github.io/pdf/M2_statistique_optimisation/grad_conv.pdf. (Important!)

[4] Gradient Descent: Convergence Analysis. https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf.

[5] Rate of Convergence. https://en.wikipedia.org/wiki/Rate_of_convergence.

[6] Linear Convergence https://www.math.drexel.edu/~tolya/linearconvergence.pdf