> 整理自同学的笔记

Gradient Descent

Part 1

$$\min f(\mathbf{x})$$

- $f(\mathbf{x})$ is continuously differentiable.
- $f(\mathbf{x})$ is convex.

Definition: A set C is convex if the line segment between any two points in C lies in C. that is $\forall x_1, x_2 \in C$, and $\forall \theta \in [0,1]$, we have $\theta x_1 + (1-\theta)x_2 \in C$.

Definition: A function $f:\mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom}\ f$ is convex and if $\mathbf{x},\mathbf{y} \in \mathrm{dom} f$ and $\theta \in [0,1]$, we have $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$.

Definition: A function is strict convex if strict inequality holds where $\mathbf{x} \neq \mathbf{y}$ and $\theta \in [0,1]$.

Definition: A function is strongly convex with parameter u if $f - \frac{u}{2}\|\mathbf{x}\|^2$ is convex.

Theorem 1: Suppose that f is continuously differentiable. Then f is convex if and only if $\mathrm{dom}\ f$ is convex and $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle, \forall \mathbf{x},\mathbf{y} \in \mathrm{dom} f$.

Proof:

$(\Rightarrow)$

$$\begin{aligned} f(\mathbf{x}+\theta(\mathbf{y}-\mathbf{x})) &\le f(\mathbf{x}) + \theta[f(\mathbf{y}) - f(\mathbf{x})] \\ f(\mathbf{y}) - f(\mathbf{x}) &\ge \lim_{\theta\to 0}\frac{f(\mathbf{x}+\theta(\mathbf{y}-\mathbf{x}))}{\theta} \text{（方向导数）} \\ &= \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \end{aligned}$$

$(\Leftarrow)$

$$\mathbf{z} = \theta\mathbf{x} + (1-\theta)\mathbf{y}$$

$$f(\mathbf{x}) \ge f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x}-\mathbf{z}\rangle \tag{1}$$

$$f(\mathbf{y}) \ge f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y}-\mathbf{z}\rangle \tag{2}$$

$\theta(1) + (1-\theta)(2)$可得。

Corollary: Suppose f is continuously differentiable. Then f is convex iff $\mathrm{dom} f$ is convex and $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \ge 0$.

Theorem 2: Suppose that f is continuously differentiable. Then f is convex iff $\mathrm{dom} f$ is convex and $\nabla^2 f(\mathbf{x}) \ge 0$.

Proof:

$(\Rightarrow)$

Let $\mathbf{x}_t = \mathbf{x} + t\mathbf{s}$, $t > 0$. Then

$$
\begin{aligned}
0 \le \frac{1}{t^2}\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}), \mathbf{x}_t - \mathbf{x}\rangle \\
&= \frac{1}{t}\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}), \mathbf{s}\rangle \\
&= \frac{1}{t}\int_0^t \langle \nabla^2 f(\mathbf{x}+\tau\mathbf{s})\mathbf{s}, \mathbf{s}\rangle d\tau \text{（微积分基本定理）} \\
&\xrightarrow{t\to 0} \langle \nabla^2 f(\mathbf{x})\mathbf{s}, \mathbf{s}\rangle \\
&= \mathbf{s}^T \nabla^2 f(\mathbf{x})\mathbf{s}
\end{aligned}
$$

($\Leftarrow$)

$$
\begin{aligned}
g(t) &= f(\mathbf{x}+t\mathbf{s}) \\
g'(0) &= \langle \nabla f(\mathbf{x}), \mathbf{s}\rangle \\
g''(0) &= \langle \nabla^2 f(\mathbf{x})\mathbf{s}, \mathbf{s}\rangle \\
g(1) &= g(0) + \int_0^1 g'(t)dt \\
&= g(0) + \int_0^1 [g'(0) + \int_0^t g''(\tau)d\tau]dt \\
&\ge g(0) + g'(0) \\
f(\mathbf{x}+\mathbf{s}) &\ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{s}\rangle
\end{aligned}
$$

Theorem 3: Suppose f is continuously differentiable. Then $\mathbf{x}^* \in \arg\min_{\mathbf{x}} f(\mathbf{x})$ iff $\nabla f(\mathbf{x}^*) = 0$, $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle = f(\mathbf{x})$.

Part 2

$$
\min f(\mathbf{x})
$$

- The objective function $f(\mathbf{x})$ is continuously differentiable.
- $f(\mathbf{x})$ is convex.
- $\exists \mathbf{x}^* \in \mathrm{dom} f$, s.t. $f(\mathbf{x}^*) = f^* = \min f(\mathbf{x})$.
- The gradient of f is Lipschitz continuous, that is, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|$, $L > 0$.

Algorithm: Gradient Descent

$$
\begin{aligned}
&\text{Input: An initial point } \mathbf{x_0}, \text{ a constant } \alpha \in (0, \frac{2}{L}),\ k=0 \\
&\text{while the termination condition does not hold, do} \\
&\qquad k = k+1 \\
&\qquad \mathbf{x_{k+1}} = \mathbf{x_k} - \alpha \nabla f(\mathbf{x_k}) \\
&\text{end while}
\end{aligned}
$$

Convergence Rate

Definition: Suppose that the sequence $\{a_k\}$ converges to a number L. Then, the sequence is said to converge linearly to L if there exists a number $\mu \in (0,1)$, s.t. $\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = \mu$.

Lemma 1: Suppose that a function $f \in C^1$. If ∇f is Lipschitz continuous with Lipschitz constant L, then

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

Proof:

$$
\begin{aligned}
f(\mathbf{y})-f(\mathbf{x}) &= \int_{\mathbf{x}}^{\mathbf{y}} \nabla f(\mathbf{z})\mathbf{dz} \\
&= \int_0^1 \langle \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})), \mathbf{y}-\mathbf{x}\rangle dt \\
&= \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \int_0^1 \langle \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle dt \\
&\le \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \int_0^1 \|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\| \|\mathbf{y}-\mathbf{x}\| dt \\
&\le \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + L\|\mathbf{y}-\mathbf{x}\|^2 \int_0^1 t dt \\
&= \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2
\end{aligned}
$$

（与凹凸性无关）

Lemma 2 （Descent Lemma） : Suppose that a function $f \in C^1$. If ∇f is Lipschitz continuous with Lipschitz constant $L > 0$, then $\forall\{\mathbf{x_k}\}$ generated by the Gradient Descent Algorithm satisfies

$$
f(\mathbf{x_{k+1}}) \le f(\mathbf{x_k}) - \alpha(1-\frac{L\alpha}{2})\|\nabla f(\mathbf{x_k})\|^2.
$$

（这也是为什么算法约定$\alpha \in (0, \dfrac{2}{L})$）

下面证明算法可以收敛到最小值，在前提条件下，可以考虑证明：

$$
\lim_{k\to\infty} \nabla f(\mathbf{x_k}) = \nabla f(\lim_{k\to\infty} \mathbf{x_k}) = 0.
$$

Proof:

由Lemma 2，

$$
\begin{aligned}
\|\nabla f(\mathbf{x_k})\|^2 &\le \frac{f(\mathbf{x_k}) - f(\mathbf{x_{k+1}})}{\alpha(1-\frac{L\alpha}{2})} \\
\sum_k \|\nabla f(\mathbf{x_k})\|^2 &\le \frac{f(\mathbf{x_0}) - f(\mathbf{x_{k+1}})}{\alpha(1-\frac{L\alpha}{2})} \\
&\le \frac{f(\mathbf{x_0}) - f^*}{\alpha(1-\frac{L\alpha}{2})}
\end{aligned}
$$

这个求和存在固有上界，故

$$
\lim_{k\to\infty} \nabla f(\mathbf{x_k}) = 0
$$

Efficiency and limitations

Theorem: Consider the Problem（$\min f(x)$）and the sequence generated by the Gradient Descent Algorithm. Then the sequence value $f(\mathbf{x_k})$ tends to the optimum function value in a rate of $O(\frac{1}{k})$.

1. If $\alpha \in (0, \dfrac{1}{L})$

$$
f(\mathbf{x_k}) - f^* \le \frac{1}{k}(\frac{1}{2\alpha}\|\mathbf{x_0} - \mathbf{x}^*\|^2)
$$

2. If $\alpha \in (\dfrac{1}{L}, \dfrac{2}{L})$

$$
f(\mathbf{x_k}) - f^* \le \frac{1}{k}(\frac{1}{2\alpha}\|\mathbf{x_0} - \mathbf{x}^*\|^2 + \frac{L\alpha - 1}{2 - L\alpha}(f(\mathbf{x_0}) - f(\mathbf{x}^*)))
$$

Proof:

As $\mathbf{x_{k+1}} = \mathbf{x_k} - \alpha\nabla f(\mathbf{x_k})$ and $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$,

$$f(\mathbf{x_{k+1}}) \leq f(\mathbf{x_k}) - (\frac{1}{\alpha} - \frac{L}{2})\|\mathbf{x_{k+1}} - \mathbf{x_k}\|^2$$

$$f(\mathbf{x_{k+1}}) - f^* \leq f(\mathbf{x_k}) - f^* - (\frac{1}{\alpha} - \frac{L}{2})\|\mathbf{x_{k+1}} - \mathbf{x_k}\|^2$$

Consider the convexity of f,

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

Summing up the inequalities,

$$\begin{aligned}
k(f(\mathbf{x_k}) - f(\mathbf{x^*})) &\leq \sum_{i=0}^{k-1}(f(\mathbf{x_{i+1}}) - f(\mathbf{x^*})) \\
&\leq \frac{1}{2\alpha}(\|\mathbf{x_0} - \mathbf{x^*}\|^2 - \|\mathbf{x_k} - \mathbf{x^*}\|^2) - (\frac{1}{2\alpha} - \frac{L}{2})\sum_{i=0}^{k-1}\|\mathbf{x_{i+1}} - \mathbf{x_i}\|^2
\end{aligned}$$

1. If $\alpha \in (0, \frac{1}{L})$, $\frac{1}{2\alpha} - \frac{L}{2} > 0$, then

$$k(f(\mathbf{x_k}) - f(\mathbf{x^*})) \leq \frac{1}{2\alpha}\|\mathbf{x_0} - \mathbf{x^*}\|^2.$$

2. If $\alpha \in (\frac{1}{L}, \frac{2}{L})$, $\frac{1}{2\alpha} - \frac{L}{2} > 0$, then

$$\begin{aligned}
k(f(\mathbf{x_k}) - f(\mathbf{x^*})) &\leq \frac{1}{2\alpha}(\|\mathbf{x_0} - \mathbf{x^*}\|^2 - \|\mathbf{x_k} - \mathbf{x^*}\|^2) + \frac{L\alpha - 1}{2\alpha}\sum_{i=0}^{k-1}\|\mathbf{x_{i+1}} - \mathbf{x_i}\|^2 \\
&\leq \frac{1}{2\alpha}\|\mathbf{x_0} - \mathbf{x^*}\|^2 + \frac{L\alpha - 1}{2\alpha}\sum_{i=0}^{\infty}\|\mathbf{x_{i+1}} - \mathbf{x_i}\|^2 \\
&\leq \frac{1}{2\alpha}\|\mathbf{x_0} - \mathbf{x^*}\|^2 + \frac{L\alpha - 1}{2\alpha}\frac{2\alpha}{2 - L\alpha}(f(\mathbf{x_0}) - f(\mathbf{x^*}))(Lemma\ 2)
\end{aligned}$$

Remark: $\|\mathbf{x_k} - \mathbf{x^*}\|$ doesn't always converge to 0.