Convergence of Gradient Descent

I. $f(x)$ has Lipschitz Continuous $\nabla f(x)$

Theorem Assume ∇f is L-continuous.

Assume $f(x) \geq f(x_*)$, $\forall x \in \mathbb{R}^n$

Then for $x_{k+1} = x_k - \eta \nabla f(x_k)$

where $\eta \in (0, \frac{2}{L})$ is a contant:

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

$\omega = \eta(1 - \frac{L}{2}\eta)$

② $\lim\limits_{x \to \infty} \|\nabla f(x_k)\| = 0$

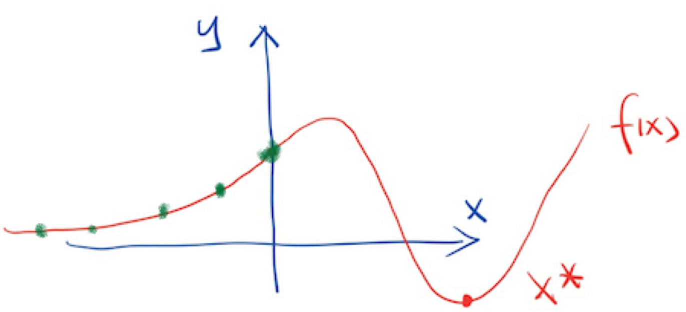
③ $\min\limits_{0 \leq k \leq N} \|\nabla f(x_k)\| \leq \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}[f(x_0) - f(x_*)]}$

Example 2.2. *We construct an example for which the gradient descent method produces almost* $\|\nabla f(\mathbf{x}_k)\| = \frac{1}{k}$. *Consider the following function*

$$f(x) = \begin{cases} e^x, & x \leq 0 \\ g(x), & x > 0 \end{cases},$$

where we pick a function $g(x)$ *such that*

1. $f(x)$ *is very smooth;*
2. $|f''(x)| \leq 1$ *for any* x, *which implies* $f'(x)$ *is L-continuous with* $L = 1$;
3. $f(x)$ *has a global minimizer* x_*.



So a stable step size can be chosen as any positive $\eta < 2$. We consider the following gradient descent iteration with $\eta = 1$:

$$\begin{cases} x_{k+1} = x_k - f'(x_k) \\ x_0 = 0 \end{cases}.$$

Notice that all iterates x_k stays non-positive, it can also be written as

$$x_{k+1} = x_k - e^{x_k}, \quad x_0 = 0.$$

One can easily implement this on MATLAB to verify that numerically we have $|f'(x_k)| \approx \frac{1}{k}$ for this iteration.

Remark: this example shows than $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$ does not imply convergence to critical point.

II. Convergence of GD for convex $f(x)$, L-cont. $\nabla f(x)$

<u>Theorem</u> Assume ∇f is L-continuous.
Assume $f(x) \geq f(x_*)$, $\forall x \in \mathbb{R}^n$
Assume $f(x)$ is convex.

Then for $x_{k+1} = x_k - \eta \nabla f(x_k)$

where $\eta \in (0, \frac{2}{L})$ is a constant:

$$f(x_k) - f(x_*) \leq \frac{1}{\frac{1}{f(x_0)-f(x_*)} + k\omega \frac{1}{\|x_0 - x_*\|^2}} < \frac{\|x_0 - x_*\|^2}{k\omega}$$

Remark 2.8. *We obtain convergence rate $\mathcal{O}(\frac{1}{k})$, assuming only convexity of the cost function and Lipschitz-continuity of its gradient. We cannot expect convergence of $\mathbf{x}_k$ to $\mathbf{x}_*$ because a convex function may have multiple global minimizers, e.g., $f(\mathbf{x}) \equiv 0$.*

III. Convergence of GD for $\begin{cases} \text{strongly convex } f(x) \\ L\text{-Cont } \nabla f(x) \end{cases}$

Theorem 2.12 (Global linear rate of gradient descent)**.** *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex with $\mu > 0$ and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L. Then $f(\mathbf{x})$ has a unique global minimizer: $f(\mathbf{x}) \geq f(\mathbf{x}_*), \forall \mathbf{x}$. The gradient descent method (2.1) with a constant step size $\eta \in (0, \frac{2}{L+\mu}]$ satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \leq \left(1 - \frac{2\eta\mu L}{L+\mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|^2.$$

In particular, if $\eta = \frac{2}{L+\mu}$, then we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|,$$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq \frac{L}{2}\left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Remark: $\mu I \leq \nabla^2 f(x) \leq L I, \ \forall x$

$$\Rightarrow \begin{cases} f(x) \text{ is strongly convex with } \mu > 0 \\ \nabla f(x) \text{ is } L\text{-Cont. with } L \end{cases}$$

Proofs of Theorems:

Lemma 2.1 (Descent Lemma)**.** *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then*

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

Convexity $= \langle \nabla f(x) - \nabla f(y), x - y \rangle \geqslant 0$

Theorem 2.8. *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L and $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex. Then for any $\mathbf{x}, \mathbf{y}$:*

1. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$
2. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.

Remark 2.7. *Without convexity, by the proof of Lemma 2.1, we only have*

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

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

With strong convexity, we can have

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

Proof: Define $\phi(x) = f(x) - \langle \nabla f(x_0), x \rangle$

Then $\nabla \phi(x) = \nabla f(x) - \nabla f(x_0)$

$\nabla \phi(y) = \nabla f(y) - \nabla f(x_0)$

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

$\Rightarrow$ $\phi(x)$ is L-continuous

Descent Lemma on $\phi(x)$:

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

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

$\langle a,b\rangle \leq \|a\|\cdot\|b\|$

$\gamma = \|x-y\|$

$$\min_{x\in\mathbb{R}^n}\phi(x) \leq \min_{x\in\mathbb{R}^n}\left[\phi(y)+\|\nabla\phi(y)\|\cdot\|y-x\|+\frac{L}{2}\|x-y\|^2\right]$$

$\parallel$

$\phi(x_0)$

$$= \min_{r\geq 0}\left[\phi(y) + r\|\nabla\phi(y)\| + \frac{L}{2}r^2\right]$$

$$= \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|^2$$

$r = -\frac{1}{L}\|\nabla\phi\|$

$\phi(x) = f(x) - \langle\nabla f(x_0), x\rangle \Rightarrow \phi(x)$ is convex

linear function satisfies

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

Sum of two convex functions is still convex.

$$\left.\begin{array}{l}\nabla\phi(X_0)=0\\ \phi \text{ is convex}\end{array}\right\} \Rightarrow X_0 \text{ minimizes } \phi(X)$$

So $\phi(X_0) \le \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|^2$

$$f(X_0) - \langle \nabla f(X_0), X_0\rangle \le f(y) - \langle \nabla f(X_0), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(X_0)\|^2$$

$$\Rightarrow f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le f(y) \quad (1)$$

$$f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \le f(x) \quad (2)$$

Add two inequalities $\Rightarrow$

$$\|\nabla f(x) - \nabla f(y)\|^2 \le L\,\langle \nabla f(x) - \nabla f(y), x - y\rangle$$

Theorem Assume ∇f is L-continuous.
Assume $f(x) \ge f(x_*)$, $\forall x \in \mathbb{R}^n$
Assume $f(x)$ is convex.

Then for $x_{k+1} = x_k - \eta \nabla f(x_k)$

where $\eta \in (0, \frac{2}{L})$ is a contant:

$$f(x_k) - f(x_*) \le \frac{1}{\frac{1}{f(x_0) - f(x_*)} + k\omega \frac{1}{\|x_0 - x_*\|^2}} < \frac{\|x_0 - x_*\|^2}{k\omega}$$

$$\omega = \eta\left(1 - \frac{L}{2}\eta\right)$$

Proof: Let $r_k = \|x_k - x_*\|$

$$r_{k+1}^2 = \|x_{k+1} - x_*\|^2$$

$$= \|x_k - \eta \nabla f(x_k) - x_*\|^2$$

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

$$= \|x_k - x_* - \eta \nabla f(x_k)\|^2$$

$$= r_k^2 - 2\eta\langle \nabla f(x_k), x_k - x_*\rangle + \eta^2 \|\nabla f(x_k)\|^2$$

$$= r_k^2 - 2\eta\langle \nabla f(x_k) \underline{- \nabla f(x_*)}, x_k - x_*\rangle + \eta^2 \|\nabla f(x_k)\|^2$$

$$\leq r_k^2 - 2\eta \frac{1}{L}\|\nabla f(x_k) - \nabla f(x_*)\|^2 + \eta^2 \|\nabla f(x_k)\|^2$$

$$\leq r_k^2 - \eta\left(\frac{2}{L} - \eta\right)\|\nabla f(x_k)\|^2$$

Let $R_k = f(x_k) - f(x_*)$

Convexity $\Rightarrow f(x) \geq f(x_k) + \langle \nabla f(x_k), x - x_k\rangle$

$$\Rightarrow f(x_*) \geq f(x_k) + \langle \nabla f(x_k), x_* - x_k\rangle$$

$$\Rightarrow f(x_k) - f(x_*) \le \langle \nabla f(x_k), x_k - x_* \rangle$$
$$\le \|\nabla f(x_k)\| \cdot \|x_k - x_*\|$$

$$\Rightarrow R_k \le \|\nabla f(x_k)\| \cdot r_k$$
$$\Rightarrow -\|\nabla f(x_k)\| \le -\frac{R_k}{r_k}$$

Recall we have $\qquad \omega = \eta(1 - \frac{L}{2}\eta)$

$$f(x_{k+1}) \le f(x_k) - \omega \|\nabla f(x_k)\|^2$$
$$\Rightarrow f(x_{k+1}) - f(x_*) \le f(x_k) - f(x_*) - \omega \|\nabla f(x_k)\|^2$$
$$0 < R_{k+1} \le R_k - \frac{\omega}{r_k^2} R_k^2$$

Multiply both sides by $\frac{1}{R_{k+1}} \frac{1}{R_k}$

$$\Rightarrow \frac{1}{R_k} \le \frac{1}{R_{k+1}} - \frac{\omega}{r_k^2} \frac{R_k}{R_{k+1}}$$
$$\Rightarrow \frac{1}{R_{k+1}} \ge \frac{1}{R_k} + \frac{\omega}{r_k^2} \frac{R_k}{R_{k+1}}$$
$$\ge \frac{1}{R_k} + \frac{\omega}{r_k^2}$$

$r_k \le r_0$

$\frac{1}{r_k} \ge \frac{1}{r_0}$

Summing up for $k = 0, 1, \cdots, N$

$$\Rightarrow \quad \frac{1}{R_{N+1}} \geq \frac{1}{R_0} + \frac{\omega}{r_0^2}(N+1)$$

$$\Rightarrow \quad R_{N+1} \leq \frac{1}{\frac{1}{R_0} + \frac{\omega}{r_0^2}(N+1)}$$