The method described in this section require a suitable starting point $x^{(0)}$. The starting point must lie in dom$f$, and in addition the sublevel set

$$S = \left\{ x \in \mathrm{dom} f : \; f(x) \le f(x^{(0)}) \right\}$$

must be closed. This condition is satisfied for all $x^{(0)} \in \mathrm{dom} f$ if the function f is closed. Continuous functions with $\mathrm{dom}(f) = \mathbb{R}^n$ are closed, so if $\mathrm{dom}(f) = \mathbb{R}^n$, the initial sublevel set condition is satisfied by any $x^{(0)}$.

Theorem 1. *Assume that f convex and differentiable, with* $\mathrm{dom}(f) = \mathbb{R}^n$ *and ∇f is Lipschitz continuous with constant $L > 0$, i.e.*

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2 \qquad \forall x, y$$

then the gradient descent with fixed step size $t \le 1/L$ satisfies

$$f(x^{(k)}) - f^\star \le \frac{\|x^{(0)} - x^\star\|}{2tk}$$

We say that the gradient descent has convergence rate $O(1/k)$.

Proof. **Part I:** With ∇f Lipschitz constant L, we have that

$$f(y) \le f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|x - y\|_2^2 \qquad \forall x, y \tag{1}$$

Suppose we are at x in the gradient descent and the next iteration go to

$$x^+ = x - t\nabla f(x)$$

We can use the above inequality with $y = x^+$ and

$$\begin{aligned} f(x^+) \;&\le f(x) + \nabla f(x)^T(-t\nabla f(x)) + \frac{L}{2}\| - t\nabla f(x)\|_2^2 \\ &= f(x) - t\|\nabla f(x)\|_2^2 + \frac{Lt^2}{2}\|\nabla f(x)\|_2^2 \\ &= f(x) - \left(1 - \frac{Lt}{2}\right) t\|\nabla f(x)\|_2^2 \end{aligned}$$

If $0 \le t \le 1/L$, we get $-t + \frac{Lt^2}{2} \le \frac{-t}{2}$ which gives us that

$$f(x^+) \;\le\; f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2. \tag{2}$$

This result also implies the descent property of the gradient descent algorithm

$$f(x^+) \;\le\; f(x).$$

Part II: Use convexity of f, we know that

$$f(x^\star) \;\ge\; f(x) + \nabla f(x)^T(x^\star - x)$$

$$f(x) \;\le\; f(x^\star) - \nabla f(x)^T(x^\star - x) \tag{3}$$

Plugin (3) into (2) and you get

$$f(x^+) \;\leq\; f(x^\star) + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2$$

$$\begin{aligned} f(x^+) - f(x^\star) \;&\leq\; \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \end{aligned}$$

The last equality is true because

$$\begin{aligned} \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - t\nabla f(x) - x^\star\|_2^2\right) &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star\|_2^2 + 2t\nabla f(x)^T(x - x^\star) - t^2\|\nabla f(x)\|_2^2\right) \\ &= \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2 \end{aligned}$$

Finally,

$$\begin{aligned} f(x^{(i)}) - f(x^\star) \;&\leq\; \frac{1}{2t}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|\right) \\ \sum_{i=1}^{k}\left(f(x^{(i)}) - f(x^\star)\right) \;&\leq\; \frac{1}{2t}\left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2\right) \;\leq\; \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2 \end{aligned}$$

because we've proved that $f(x^{(0)}) \geq f(x^{(1)}) \geq \ldots \geq f(x^{(k)})$. Thus

$$f(x^{(k)}) - f(x^\star) \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(x^{(i)}) - f(x^\star)\right) \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

□

Remark 1. We can show that in Theorem 1, the assumption that ∇f is Lipschitz continuous with constant $L > 0$ can be relaxed to that we only need Lipschitz gradient over the sublevel set

$$S = \left\{x \in \mathrm{dom} f : \; f(x) \leq f(x^{(0)})\right\}.$$

Theorem 2. *If the sublevel sets contained in S are bounded, so in particular, if S is bounded. Then ∇f is Lipschitz continuous with constant $L > 0$ over S.*

Proof. If S is bounded, then the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of x on S, is also bounded above on S. i.e., there exist a constant L such that

$$\nabla^2 f(x) \preceq LI \qquad \forall x \in S.$$

This upper bound on the Hessian implies for any $x, y \in S$

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2$$

Therefore we get a similar condition to the original Lipschitz continuous assumption (1) except that it is on the sublevel set S, which is sufficient to prove Theorem 1 since this condition can also lead to the descent property on the sublevel set

$$f(x^{(1)}) \;\leq\; f(x^{(0)}) - \frac{t}{2}\|\nabla f(x^{(0)})\|_2^2 \qquad \forall x \in S$$

□