1 Proof of gradient descent

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

$$ S = \{ x \in \text{dom} f : f(x) \leq f(x^{(0)}) \} $$

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

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

$$ \| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2 \quad \forall x, y $$

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

$$ f(x^{(k)}) - f^* \leq \frac{\| x^{(0)} - x^* \|}{2tk} $$
We say that the gradient descent has convergence rate $O(1/k)$.

**Proof.** **Part I:** With $\nabla f$ Lipschitz constant $L$, we have that

\[
f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|x - y\|_2^2 \quad \forall x, y \tag{1}
\]

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

\[
f(x^+) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Lt^2}{2} \|\nabla f(x)\|_2^2
\]

\[
= f(x) - \left(1 - \frac{Lt^2}{2}\right) t\|\nabla f(x)\|_2^2
\]

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

\[
f(x^+) \leq f(x) - \frac{t}{2} \|\nabla f(x)\|_2^2. \tag{2}
\]

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

\[
f(x^+) \leq f(x).
\]

**Part II:** Use convexity of $f$, we know that

\[
f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)
\]

\[2\]
\[ f(x) \leq f(x^*) - \nabla f(x)^T (x^* - x) \] (3)

Plugin (3) into (2) and you get

\[
\begin{align*}
  f(x^+) & \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{t}{2} \| \nabla f(x) \|^2_2 \\
  f(x^+) - f(x^*) & \leq \nabla f(x)^T (x - x^*) - \frac{t}{2} \| \nabla f(x) \|^2_2 \\
  & = \frac{1}{2t} (\| x - x^* \|^2_2 - \| x^+ - x^* \|^2_2)
\end{align*}
\]

The last equality is true because

\[
\frac{1}{2t} (\| x - x^* \|^2_2 - \| x - t \nabla f(x) - x^* \|^2_2) = \frac{1}{2t} (\| x - x^* \|^2_2 - \| x - x^* \|^2_2 + 2t \nabla f(x)^T (x - x^*) - t^2 \| \nabla f(x) \|^2_2)
\]

\[
= \nabla f(x)^T (x - x^*) - \frac{t}{2} \| \nabla f(x) \|^2_2
\]

Finally,

\[
f(x^{(i)}) - f(x^*) \leq \frac{1}{2t} (\| x^{(i-1)} - x^* \|^2_2 - \| x^{(i)} - x^* \|)
\]

\[
\sum_{i=1}^k (f(x^{(i)}) - f(x^*)) \leq \frac{1}{2t} (\| x^{(0)} - x^* \|^2_2 - \| x^{(k)} - x^* \|^2_2) \leq \frac{1}{2t} \| x^{(0)} - x^* \|^2_2
\]

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

\[
f(x^{(k)}) - f(x^*) \leq \frac{1}{k} \sum_{i=1}^k (f(x^{(i)}) - f(x^*)) \leq \frac{\| x^{(0)} - x^* \|^2_2}{2tk}
\]

\[ \square \]

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

\[ S = \{ x \in \text{dom} f : f(x) \leq f(x^{(0)}) \} . \]

**Theorem 2.** If the sublevel sets contained in \( S \) are bounded, so in particular, if \( S \) is bounded. Then \( \nabla f \) 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

\[ \nabla^2 f(x) \preceq LI \quad \forall x \in S. \]

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

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

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

\[ f(x^{(1)}) \leq f(x^{(0)}) - \frac{t}{2} ||\nabla f(x^{(0)})||^2 \quad \forall x \in S \]

\[ \square \]

**Remark.** For example, if \( f \) is strongly convex then \( S \) is bounded

\[ f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2} ||y - x||^2 \]

at \( x = 0 \)

\[ f(y) \geq f(0) + \nabla f(0)^T(y - 0) + \frac{M}{2} ||y||^2 \]
we can see that if $\|y\|_2 \to \infty$ then $f(y) \to \infty$, so $f(y)$ is bowl-shaped.

2 Convergence analysis for backtracking

The backtracking exit inequality:

$$f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$$

By Lipschitz continuous gradient, we can show that:

$$f(x + t\Delta x) \leq f(x) + \frac{1}{L} \nabla f(x)^T \Delta x_0$$

$$\leq f(x) + \frac{\alpha}{L} \nabla f(x)^T \Delta x$$
The backtracking exit inequality \((*)\) holds for \(t \geq 0\) in an interval \((0, \frac{1}{L})\).

The backtracking line search stops with a stepsize of length \(t\) that satisfies:

\[
t = 1 \quad \text{or} \quad t \in \left( \beta \frac{1}{L}, \frac{1}{L} \right)
\]

Case 1:
\( t = 1 \) already satisfies the (\ast) i.e. \( t = 1 \leq \frac{1}{L} \)

Case 2: Otherwise \( 1 > t \) then the stepsize \( t \in \left( \frac{\beta}{L}, \frac{1}{L} \right) \)

Therefore, the step length \( t \geq \min \{1, \frac{\beta}{L}\} \)

Iterative method, updates \( x^{(k)} \) by:

\[
\begin{align*}
x^{(1)} &= x^{(0)} + t \nabla f(x^{(0)}) \\
x^{(2)} &= x^{(1)} + t \nabla f[x^{(0)} + t \nabla f(x^{(0)})]
\end{align*}
\]

3 How to choose stepsize \( t \)

Gradient Descent with constant \( t = \frac{1}{L} \) converge rate = \( O\left(\frac{1}{k}\right) \) Gradient Descent with Backtracking \( t = \min \{1, \frac{\beta}{L}\} \) converge rate = \( O\left(\frac{1}{k}\right) \) Gradient Descent with constant \( t = \frac{1}{L} \) for strongly convex, converge rate = \( O(e^k) \).