Line search methods:

Recall that to obtain the definition of a descent direction, we Taylor expanded to get:

$$f(x+\alpha p) = f(x) + \alpha \nabla f(x)^T p + O(\|p\|_2^2).$$

Then, if p is a descent direction, $\nabla f(x)^T p < 0$, and:

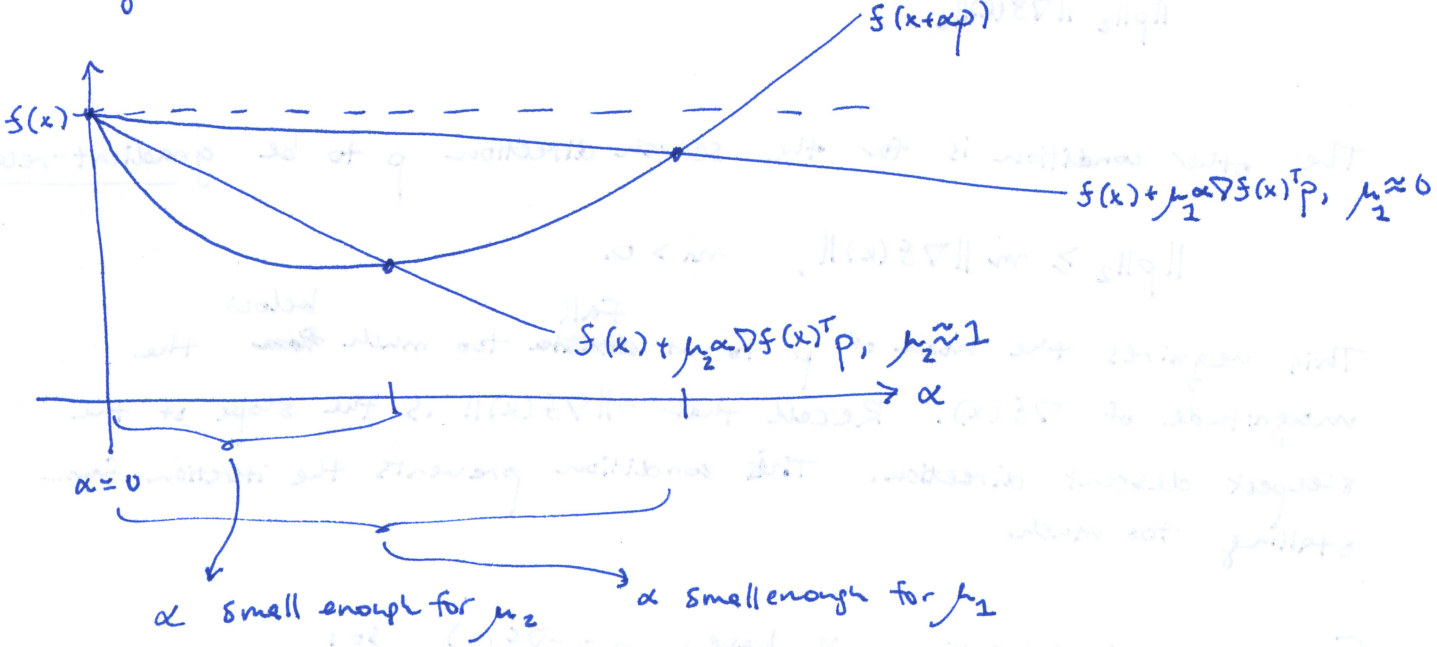
$$f(x+\alpha p) < f(x) + O(\|p\|_2^2).$$

To ensure descent with a line search, it is common to choose $\alpha > 0$ such that the Armijo condition holds:

$$f(x+\alpha p) \leq f(x) + \mu \alpha \nabla f(x)^T p,$$

where μ is a constant such that $0 < \mu < 1$. The Armijo condition sometimes is also refered to as the sufficient decrease condition.

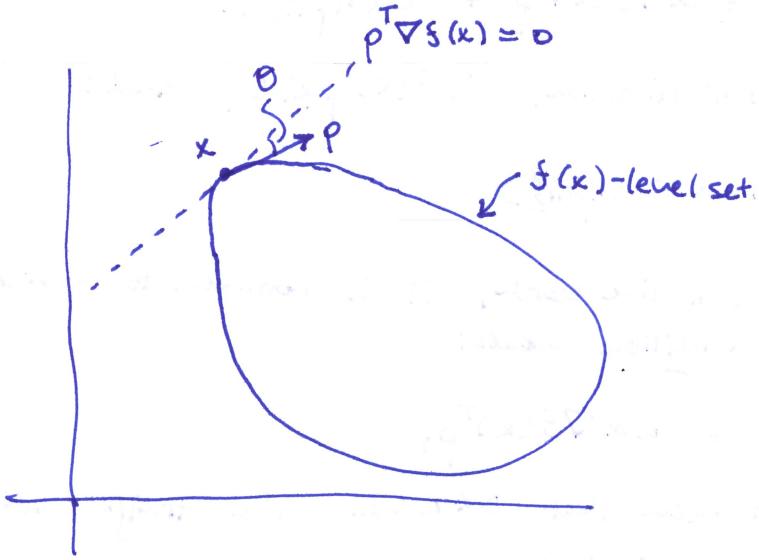
What is the idea behind the Armijo condition? Since p is a descent direction, and since $f(x) + \mu \alpha \nabla f(x)^T p$ is linear in μ, we can draw something like:



It should be clear why this condition is also known as the sufficient decrease condition. Since $\mu > 0$, it ensures that some progress is made at each iteration.

Along with the Armijo condition, there are two more conditions which can be assumed to ensure convergence.

First, is the angle condition. For a general step p, we may find that p and $\nabla f(x)$ become close to orthogonal to one another:



Clearly, if θ is small, the iteration might make little progress. To ensure progress, we can require:

$$-\frac{p^T \nabla f(x)}{\|p\|_2 \, \|\nabla f(x)\|_2} \geq \varepsilon > 0.$$

The other condition is for the search direction p to be gradient-related:

$$\|p\|_2 \geq m \|\nabla f(x)\|, \quad m > 0.$$

This requires the norm of p to not fall too much below the magnitude of $\nabla f(x)$. Recall that $-\|\nabla f(x)\|$ is the slope of the steepest descent direction. This condition prevents the iteration from stalling too much.

For gradient descent, we have: $p = -\nabla f(x)$, so:

$$\frac{-p^T \nabla f(x)}{\|p\|_2 \, \|\nabla f(x)\|_2} = \frac{\|\nabla f(x)\|_2^2}{\|\nabla f(x)\|_2^2} = 1 > 0;$$

so the angle condition is always satisfied. We also have:

$$\|p\|_2 = \|\nabla f(x)\|_2 \geq 1 \cdot \|\nabla f(x)\|_2,$$

so the search direction is clearly also gradient-related.

For Newton's method, $p = -\nabla^2 f(x)^{-1} \nabla f(x)$. If $\nabla^2 f(x)$ is positive definite, then $\nabla^2 f(x)^{-1}$ is positive definite (as we saw previously), so:

$$\frac{-p^T \nabla f(x)}{\|p\|_2 \|\nabla f(x)\|_2} = \frac{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)}{\|\nabla^2 f(x)^{-1} \nabla f(x)\|_2 \|\nabla f(x)\|_2} > 0.$$

To get this to be greater than some constant $\varepsilon > 0$ for each iteration, we need $\nabla^2 f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \geq \varepsilon C > 0$. A sufficient condition for this is to simply require $\lambda_{min}(\nabla^2 f(x)^{-1})$ to be bounded below as x varies.

We also have:

$$\|\nabla f(x)\| = \|\nabla^2 f(x) \nabla^2 f(x)^{-1} \nabla f(x)\| \leq \|\nabla^2 f(x)\| \underbrace{\|\nabla^2 f(x)^{-1} \nabla f(x)\|}_{-p}$$

Hence, if we take $m = \|\nabla^2 f(x)\|^{-1}$, we get:

$$m \|\nabla f(x)\| \leq \|\nabla^2 f(x)^{-1} \nabla f(x)\|.$$

Provided that $m > 0$ independent of x, Newton's method is also gradient-related.

If we take all of these conditions together, we get a theorem which indicates that a method with search steps satisfying these conditions based on backtracking line search converges to a local minimum.

To prove this theorem we need the following result:

Prop: Let $f: \mathbb{R}^n \to \mathbb{R}$, and assume that ∇f is Lipschitz on a convex set $S \subseteq \mathbb{R}^n$, with Lipschitz constant L. Then:

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

The main theorem is:

Theorem: (Grima 11.7) Let $f: \mathbb{R}^n \to \mathbb{R}$ such that ∇f is Lipschitz with Lipschitz constant $L > 0$; i.e:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x-y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Let $x_0 \in \mathbb{R}^n$ and consider the iteration:

$$x_{k+1} = x_k + \alpha_k p_k, \quad k \geq 0,$$

where the following conditions hold:

1) The set S (not necessarily convex!) defined to be:

$$S = \{ x \in \mathbb{R}^n : f(x) \leq f(x_0) \}$$

is bounded.

2) The vectors p_k $(k \geq 0)$ satisfy the angle condition with parameter $\varepsilon > 0$.

3) The search directions p_k $(k \geq 0)$ are gradient-related w/ parameter $m > 0$.

4) $\|p_k\| \leq M \quad (k \geq 0)$.

5) $\alpha_k \in (0, 1]$ is the first scalar among $1, \frac{1}{2}, \frac{1}{4}, \ldots$ for which the Armijo condition holds. w/ $\mu \in (0, 1)$.

Then: $\lim_{k \to 0} \|\nabla f(x_k)\|_2 = 0$. (The iteration converges to a local minimum)

Proof: we have to prove several pieces.

i) f bounded below on S:

- f is continuous
- S is bounded (by assumption)
- S is closed (by definition)
- a continuous function attains a minimum on a closed and bounded set — hence, a value bounding f from below on S exists ∎

ii) $\lim_{k \to \infty} f(x_k)$ exists:

- sufficient decrease condition $\Rightarrow f(x_{k+1}) < f(x_k) \quad \forall k \geq 0$
- hence, $x_k \in S \quad \forall k \geq 0$
- hence, $\{ f(x_k) \}_{k=0}^{\infty}$ is monotonic decreasing and bounded below $\Rightarrow$ limit exists ∎

iii) $\lim_{k \to \infty} \alpha_k \| \nabla f(x_k) \|_2^2 = 0$:

let $\bar{f} = \lim_{k \to \infty} f(x_k)$. Then:

$$
\begin{aligned}
f(x_0) - \bar{f} &= [f(x_0) - f(x_1)] + [f(x_1) - f(x_2)] + \cdots \\
&= \sum_{k=0}^{\infty} \left(f(x_k) - f(x_{k+1}) \right) \\
&\geq -\sum_{k=0}^{\infty} \mu \alpha_k p_k^T \nabla f(x_k) \qquad \left[\text{sufficient decrease, i.e. (5) (Armijo)} \right] \\
&\geq + \sum_{k=0}^{\infty} \mu \alpha_k \varepsilon \| p_k \|_2 \| \nabla f(x_k) \|_2 \qquad \left[\text{angle condition, (2)} \right] \\
&\geq \sum_{k=0}^{\infty} \mu \alpha_k \varepsilon m \| \nabla f(x_k) \|_2 . \qquad \left[\text{gradient-related, (3)} \right]
\end{aligned}
$$

Hence, this summation converges, implying that the individual terms go to zero:

$$\lim_{k\to\infty} \mu\alpha_k \varepsilon m \|\nabla f(x_k)\|^2 = 0.$$

Divide by $\mu\varepsilon m$ to get result ☑

iv) If $\alpha_k < 1$, then $\alpha_k \geq \gamma \|\nabla f(x_k)\|_2^2$, $\gamma = \frac{(1-\mu)\varepsilon m}{M^2 L} > 0$.

Two cases: $\alpha_k = 1$ or $\alpha_k < 1$. In latter case, we must have backtracked (reduced α_k by factor of $\frac{1}{2}$), so $2\alpha_k$ violates Armijo:

$$f(x_k + 2\alpha_k p_k) > f(x_k) + 2\mu\alpha_k p_k^T \nabla f(x_k). \qquad (*)$$

By our ~~theorem~~ proposition, since ∇f is L-Lipschitz:

$$f(x_k + 2\alpha_k p_k) - f(x_k) - 2\alpha_k p_k^T \nabla f(x_k) \leq \frac{1}{2} L \|2\alpha_k p_k\|_2^2.$$

(>0 by $(*)$)

Combining these inequalities and simplifying gives:

$$\alpha_k L \|p_k\|_2^2 \geq -(1-\mu) p_k^T \nabla f(x_k).$$
$$\geq (1-\mu)\varepsilon \|p_k\|_2 \|\nabla f(x_k)\|_2 \quad \text{[angle condition]}$$
$$\geq (1-\mu)\varepsilon m \|\nabla f(x_k)\|_2^2 \quad \text{[gradient-related]}$$

Since $\|p_k\|_2 \leq M$, conclude result ☑

☑ Finally: from (iv), $\alpha_k \geq \min\left(1, \gamma\|\nabla f(x_k)\|_2^2\right)$. Hence

$$\alpha_k \|\nabla f(x_k)\|_2^2 \geq \min\left(1, \gamma\|\nabla f(x_k)\|_2^2\right) \|\nabla f(x_k)\|_2^2 \geq 0.$$

Since $\alpha_k \|\nabla f(x_k)\|_2^2 \to 0$ by (iii), $\|\nabla f(x_k)\|_2^2 \to 0$. since $\gamma > 0$. ☑