# Newton's Method

- Gradient method uses only gradient information (first derivative).
- If we also use the second derivative (Hessian), we should be able to do better (but it may be more computationally demanding).
- Newton's method uses Hessian.
- For quadratics, converges in 1 step (order of convergence $\infty$).
- In general, it has order of convergence at least 2.

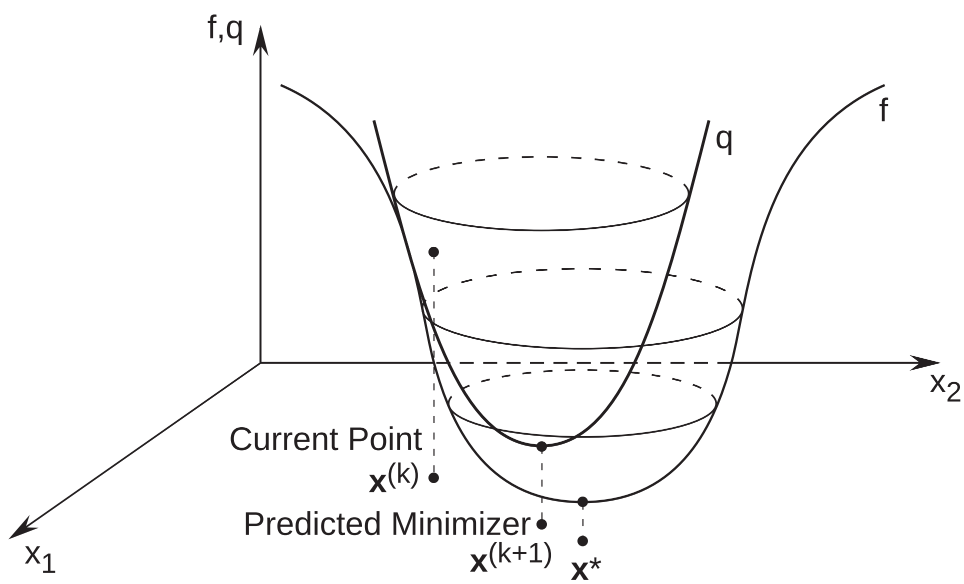
## Underlying idea (§9.1)

- Given: $f : \mathbb{R}^n \to \mathbb{R}$, and current iterate $\boldsymbol{x}^{(k)}$. Write $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$.
- To compute $\boldsymbol{x}^{(k+1)}$, approximate $f$ by a quadratic:

$$\begin{aligned} q(\boldsymbol{x}) &= f(\boldsymbol{x}^{(k)}) + (\boldsymbol{x} - \boldsymbol{x}^{(k)})^T \boldsymbol{g}^{(k)} \\ &\quad + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^T \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}). \end{aligned}$$

- Use minimizer of $q$ as next iterate $\boldsymbol{x}^{(k+1)}$.
- By FONC, we have $\nabla q(\boldsymbol{x}^{(k+1)}) = 0$, where

$$\nabla q(\boldsymbol{x}^{(k+1)}) = \boldsymbol{g}^{(k)} + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}).$$



- Newton's algorithm:
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}.$$

- Note: no step size (or, step size = 1).

- See Example 9.1.

- Can break down into two steps:

  1. Solve $\boldsymbol{F}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$ for $\boldsymbol{d}^{(k)}$
  2. Set $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)}$

- No need to explicitly compute $\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}$.

## Analysis of Newton's method (§9.2)

- Does the method work? When does it work? How well does it work?

- If $f$ is a quadratic (with invertible Hessian $\boldsymbol{Q}$), then Newton's method always converges to $\boldsymbol{x}^*$ in 1 step.

- For general $f$,

  - Hessian may not be invertible;
  - algorithm may not converge if we don't start close enough to $\boldsymbol{x}^*$;
  - it may not have descent property;
  - if/when it works, it is fast.

### Convergence of Newton's method

- What is the order of convergence of Newton's algorithm?

- Easy to show that it is $> 1$ ("superlinear") if the inverse Hessian is bounded.

- By Taylor's formula:
$$\begin{aligned}\boldsymbol{0} &= \nabla f(\boldsymbol{x}^*) \\ &= \nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^* - \boldsymbol{x}^{(k)}) + o(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|).\end{aligned}$$

- Rearranging, we obtain
$$\begin{aligned}&\boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\nabla f(\boldsymbol{x}^{(k)}) - \boldsymbol{x}^* \\ &\quad = \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1} o(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|) = o(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|)\end{aligned}$$

  by boundedness of $\boldsymbol{F}(\cdot)^{-1}$.

- Hence, $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* = o(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|)$.
- Thus,
$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|} = \lim_{k\to\infty} \frac{o(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|)}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|} = 0.$$
- The order of convergence is *superlinear* (if the order of convergence exists, it must be $> 1$).

Theorem (9.1): Suppose

1. $f \in \mathcal{C}^3$,
2. $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$,
3. $\boldsymbol{F}(\boldsymbol{x}^*)$ invertible.

Then, for all $\boldsymbol{x}^{(0)}$ sufficiently close to $\boldsymbol{x}^*$, Newton's method converges to $\boldsymbol{x}^*$ with order of convergence at least 2.

- Idea of proof: show
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2).$$

Sketch of proof:

- We have:
$$\begin{aligned}
&\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* \\
&\quad = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\nabla f(\boldsymbol{x}^{(k)}) - \boldsymbol{x}^* \\
&\quad = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\left(\nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^* - \boldsymbol{x}^{(k)})\right).
\end{aligned}$$
- By Taylor's formula and assumption 2,
$$\begin{aligned}
\boldsymbol{0} = \nabla f(\boldsymbol{x}^*) &= \nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^* - \boldsymbol{x}^{(k)}) \\
&\quad + O(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2).
\end{aligned}$$
- Thus
$$-\left(\nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^* - \boldsymbol{x}^{(k)})\right) = O(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2).$$
- Substituting Taylor's formula into first equation, we get
$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* = \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1} \cdot O(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2).$$
- Hence, taking norms,
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \leq \|\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\| \cdot O(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2)$$
- By assumptions 1 and 3, $\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}$ exists if $\boldsymbol{x}^{(k)}$ is sufficiently near $\boldsymbol{x}^*$, and is bounded.
- To make the argument rigorous, use induction and some technical lemmas (read proof in book).

**Newton's method and descent property**

- Newton's method may not have descent property.
- It is possible that for some $k$,
$$f(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k)}).$$
- Fortunately, the vector
$$\boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$$
points in a direction of decreasing $f$.
- Theorem (9.2): Suppose $\boldsymbol{F}(\boldsymbol{x}^{(k)}) > 0$ and $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$. Then, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$,
$$f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) < f(\boldsymbol{x}^{(k)}).$$
- Consequence: if we include a step size in Newton's algorithm,
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$$
and we choose $\alpha_k$ appropriately, e.g.,
$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}),$$
then the modified Newton's algorithm has a descent property.

Proof of theorem:

- As usual, write
$$\phi(\alpha) = f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}).$$
- By chain rule,
$$\phi'(0) = \nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)T}\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}.$$
- Because $\boldsymbol{F}(\boldsymbol{x}^{(k)}) > 0$ and $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$, we deduce that $\phi'(0) < 0$.
- Hence, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$, we have $\phi(\alpha) < \phi(0)$, or
$$f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) < f(\boldsymbol{x}^{(k)}).$$

**Summary**

- Newton's method performs well if we start close enough.
- We can incorporate a step size to ensure descent.
- For a quadratic, converges in one step.
- Is there some way of using only gradients, but still only converge in one or a finite number of steps for quadratics?
- Yes ... conjugate direction method.

**General algorithms**

- We have already seen two examples of algorithms of the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)},$$

  where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}).$$

- In steepest descent algorithm, $\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$.

- In (modified) Newton's algorithm,
  $\boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$.

- There are some general statements we can make about algorithms of the above form.

- Prop.: Suppose $\alpha_k > 0$. Then, the following equation holds:

$$\boldsymbol{d}^{(k)T}\boldsymbol{g}^{(k+1)} = 0.$$

- Proof: Consider $\phi(\alpha) = f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})$.

  By FONC, we have $\phi'(\alpha_k) = 0$.

  By chain rule,

$$\phi'(\alpha_k) = \boldsymbol{d}^{(k)T}\nabla f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}) = \boldsymbol{d}^{(k)T}\boldsymbol{g}^{(k+1)}.$$

- Prop.: If $\boldsymbol{d}^{(k)T}\boldsymbol{g}^{(k)} < 0$, then

  1. $\alpha_k > 0$,
  2. $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$.

- Proof: Exercise.

**Remark:**

- Steepest descent and Newton's algorithms satisfy $\boldsymbol{d}^{(k)T}\boldsymbol{g}^{(k)} < 0$ (if $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$).

- Prop.: Suppose

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b},$$

  with $\boldsymbol{Q} = \boldsymbol{Q}^T > 0$. Then,

$$\alpha_k = -\frac{\boldsymbol{d}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{d}^{(k)T}\boldsymbol{Q}\boldsymbol{d}^{(k)}}.$$

- Proof: Exercise. *Hint*: Consider $\phi$.