Newton's Methods

Consider $$\min_{x \in \mathbb{R}^n} f(x)$$

- $n = 1$ $\quad x_{k+1} = x_k - \left(f''(x_k)\right)^{-1} f'(x_k)$
- $n > 1$ $\quad x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$

Another form ① solve $\quad \nabla^2 f(x_k)\, d_k = -\nabla f(x_k)$

② update $\quad x_{k+1} = x_k + d_k$

Remark. Do not require $\nabla^2 f(x_k) > 0$ only needs non-singular (invertible).
Namely, Newton's method also works for non-convex optimization problems.
but may not find local min.

Pro. 1. Convergences super-fast (quadratic rate)

2. Affine invariant.

Con. 1. Local convergence. Require $\|x_0 - x^*\|$ is small enough.

2. Computational cost.

Form Hessian matrix $O(n^2)$. Compute $(\nabla^2 f)^{-1}$: $O(n^3)$.

Derivation. Given current approximate x_k, approximates f by its quadratic Taylor series

$$f(x) \approx f_q(x; x_k) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle$$

$$\min_{x\in\mathbb{R}^n} f(x) \rightsquigarrow \min_{x\in\mathbb{R}^n} f_q(x;x_k) \rightsquigarrow \nabla f_q(x_{k+1};x_k) = 0.$$

$\Downarrow$ Newton's method

$$\nabla f_q(x;x_k) = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k)$$

162 NEWTON'S METHOD

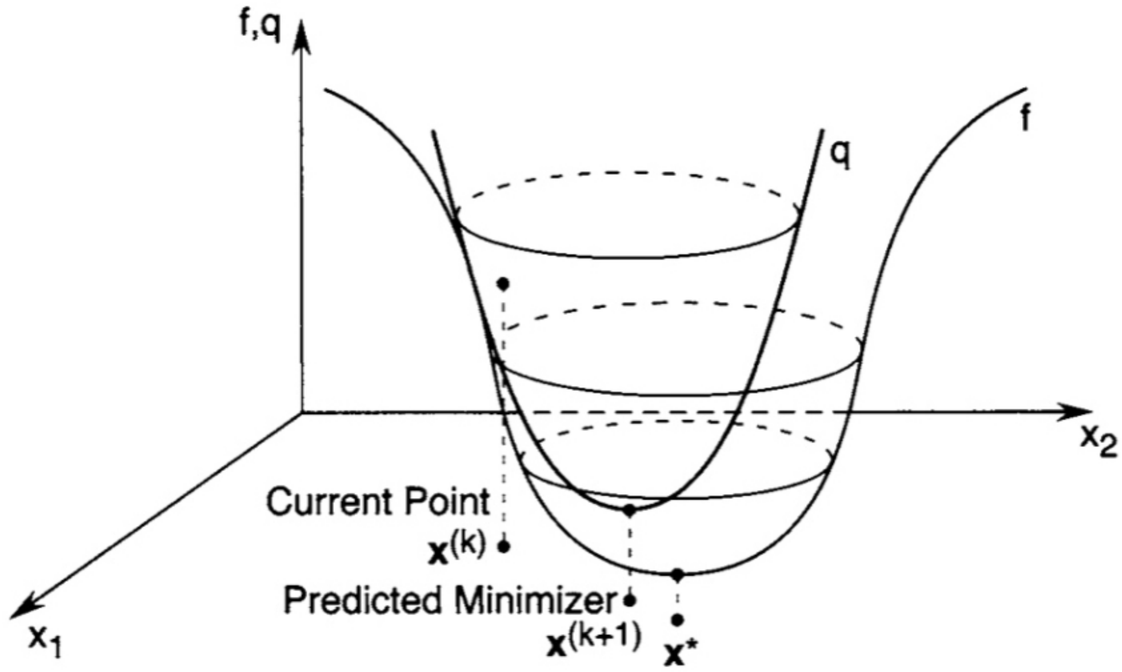


Figure 9.1 Quadratic approximation to the objective function using first and second derivatives.

Convergence Analysis.

Theorem. Suppose $f \in C^3$. x^* is a critical pt, i.e. $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*)$ is invertible. Then for all x_0 sufficiently close to x^*, Newton's method is well defined for all k, and $\|x_{k+1} - x^*\| \leqslant C\|x_k - x^*\|^2 \quad \forall k = 0, 1, 2, \cdots$

Proof. Denote by $F(x) = \nabla^2 f(x)$. Then $\det F(x) \in C^1$. As $\det F(x^*) \neq 0$, for sufficiently small ε, $\det F(x) \neq 0$, $\forall \|x - x^*\| < \varepsilon$. So $F(x)$ is invertible.

Furthermore $\|F^{-1}(x)\| \leq C$, $\forall \|x-x^*\| < \varepsilon$.

Assume x_k satisfies $\|x_k - x^*\| < \varepsilon$, then $F^{-1}(x_k)$ exists and $\|F^{-1}(x_k)\| \leq C$.

Then $x_{k+1} - x^* = x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

$$= (\nabla^2 f(x_k))^{-1} [\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k)].$$

We apply first order Taylor expansion to $\nabla f(x^*)$ at x_k to get

$$\nabla f(x^*) = \nabla f(x_k) + \nabla^2 f(x_k)(x^* - x_k) + O(\|x_k - x^*\|^2)$$

Note that $\nabla f(x^*) = 0$ and the sign change, we have

$$\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) = O(\|x_k - x^*\|^2).$$

Therefore $\|x_{k+1} - x^*\| \leq C\|(\nabla^2 f(x_k))^{-1}\| \, \|x_k - x^*\|^2$

$$\leq C_1 \|x_k - x^*\|^2$$

Again by choosing ε sufficiently small s.t. $C_1 \varepsilon^2 < \varepsilon$, we conclude

$\|x_{k+1} - x^*\| < \varepsilon$ and $F(x_{k+1})^{-1}$ exists and $\|F(x_{k+1})^{-1}\| \leq C$.

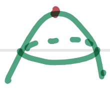
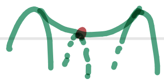
So if ε is small enough and $\|x_0 - x^*\| < \varepsilon$, all $\|x_k - x^*\| < \varepsilon$ and

$$\|x_{k+1} - x^*\| \leq C_1 \|x_k - x^*\|^2 \quad \forall k = 0, 1, 2, \cdots$$

which implies the local quadratic convergence. #.

Example. $C_1 = 10$. $\|x_0 - x^*\| < 10^{-1}$, $\|x_1 - x^*\| \leq 10^{-1}$ and $\|x_k - x^*\| \leq 10^{-1}$.

Not convergent. $\|x_0 - x^*\| \leq 10^{-2}$, $\|x_1 - x^*\| \leq 10^{-3}$, $\|x_2 - x^*\| \leq 10^{-5}$ super-fast.

Remark. The convergence is proved for $\|x_k - x^*\|$, where x^* is only a critical point, i.e. $\nabla f(x^*) = 0$. x^* may not be a local mininum. It could be local max  or a saddle pt . To be a local minimum, we need to further verify $\nabla^2 f(x^*) > 0$.

Modification of Newton's method.

Newton's method may not be a descent method, i.e. $f(x_{k+1}) > f(x_k)$ is possible (e.g. x^* is a local maximum). Have to restrict to strickly convex functions.

Lemma. Assume $\nabla^2 f(x) > 0$, $\forall x$. If $\nabla f(x_k) \neq 0$, then Newton's direction $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ is a descent direction in the sense that $f(x_k + \alpha d_k) < f(x_k)$ for sufficiently small α.

Proof. Let $\phi(\alpha) = f(x_k + \alpha d_k)$. Then $\phi'(\alpha) = (\nabla f(x_k + \alpha d_k), d_k)$ and

$$\phi'(0) = -(\nabla f(x_k), (\nabla^2 f(x_k))^{-1} \nabla f(x_k)) = -(g_k, g_k)_a < 0 \quad \text{where}$$

$g_k = \nabla f(x_k)$, $a = (\nabla^2 f(x_k))^{-1} > 0$.

Then for sufficiently small α, $f(x_k + \alpha d_k) = \phi(\alpha) < \phi(0) = f(x_k)$. #.

For convex functions, we can use the following modification

1. Compute d_k by solving $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$
2. Find $\alpha_k = \operatorname{argmin} f(x_k + \alpha d_k)$ by line search.
3. Update $x_{k+1} = x_k + \alpha_k d_k$.

What if $\nabla^2 f$ is not SPD? Note that for non-convex functions, the gradient method $x_{k+1} = x_k - \alpha_k I \nabla f(x_k)$ is always a descent method. This motivates the Levenberg-Marquardt modification

$$\boxed{x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k) + \mu_k I \right)^{-1} \nabla f(x_k),}$$

where $\mu_k > 0$ is chosen s.t. $\nabla^2 f(x_k) + \mu_k I \succ 0$ and $\alpha_k > 0$ is a step size.

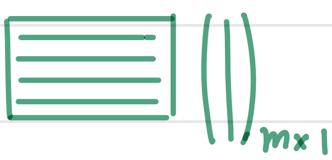
It is a mixture of Newton and gradient methods:

- $\mu_k = 0$. Newton's method.
- $\mu_k \to +\infty$. Gradient method.

Non-linear Least Squares

$f(x) = \frac{1}{2}\|r(x)\|^2$ where $r = (r_1, r_2, \cdots, r_m) \in \mathbb{R}^m$ and $x = (x_1, x_2, \cdots x_n) \in \mathbb{R}^n$.

- $\nabla f(x) = \left(\frac{\partial r}{\partial x}, r\right) = J(x)^T r(x)$, where $J(x) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(x) & \cdots & \frac{\partial r_1}{\partial x_n}(x) \\ \vdots & & \\ \frac{\partial r_m}{\partial x_1}(x) & \cdots & \frac{\partial r_m}{\partial x_n}(x) \end{bmatrix}_{m\times n}$



- $\nabla^2 f(x) = (\nabla^2 r, r) + \left(\frac{\partial r}{\partial x}, \frac{\partial r}{\partial x}\right) = S(x) + J^T(x) J(x)$.

- Newton's method $x_{k+1} = x_k - \left(J^T(x_k) J(x_k) + S(x_k)\right)^{-1} J^T(x_k)\, r(x_k)$
- Gauss-Newton method $x_{k+1} = x_k - \left(J^T(x_k) J(x_k)\right)^{-1} J^T(x_k)\, r(x_k)$
- Levenberg-Marquardt method $x_{k+1} = x_k - \left(J^T(x_k) J(x_k) + \mu_k I\right)^{-1} J^T(x_k)\, r(x_k)$.