MATH 4211/6211 – Optimization

Newton's Method

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

Newton's method

- Improve gradient method by using second-order (Hessian) information.
- Approximate f at $\boldsymbol{x}^{(k)}$ locally by a quadratic function, and use the minimizer of the quadratic function as $\boldsymbol{x}^{(k+1)}$.
- The Newton's method resolves to iterating

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{H}^{(k)})^{-1}\boldsymbol{g}^{(k)}$$

where $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ and $\boldsymbol{H}^{(k)} = \nabla^2 f(\boldsymbol{x}^{(k)})$.

The Newton's (or Newton-Raphson) method executes the two steps below in each iteration:

- Step 1: Solve $\boldsymbol{d}^{(k)}$ from $\boldsymbol{H}^{(k)}\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$;
- Step 2: Update $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)}$.

Therefore the key is solving a linear system in Step 1 in every iteration.

- **Pros**:
  - very fast convergence near solution x^* (more later).
- **Cons**:
  - not a descent method;
  - Hessian may not be invertible;
  - may diverge if initial guess is bad.

We will see how fast Newton's method is, and how to remedy the issues.

Let us first see what happens when applying Newton's method to minimize the quadratic functions with $\boldsymbol{Q} \succ 0$:

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

We know that

$$\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b} \qquad \text{and} \qquad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{Q}$$

In addition, the unique minimizer is $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$.

Therefore, given any initial $\boldsymbol{x}^{(0)}$, we have

$$\begin{aligned}
\boldsymbol{x}^{(1)} &= \boldsymbol{x}^{(0)} - (\boldsymbol{H}^{(0)})^{-1}\boldsymbol{g}^{(0)} \\
&= \boldsymbol{x}^{(0)} - (\boldsymbol{Q})^{-1}(\boldsymbol{Q}\boldsymbol{x}^{(0)} - \boldsymbol{b}) \\
&= \boldsymbol{Q}^{-1}\boldsymbol{b} \\
&= \boldsymbol{x}^*
\end{aligned}$$

which means the Newton's method converges in 1 iteration.

Convergence of the Newton's method for general case.

Theorem. Suppose $f \in \mathcal{C}^3(\mathbb{R}^n;\mathbb{R})$, and $\exists\, \boldsymbol{x}^* \in \mathbb{R}^n$ such that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*)$ is invertible. Then for all $\boldsymbol{x}^{(0)}$ sufficiently close to $\boldsymbol{x}^*$, the Newton's method is well-defined for all k, and $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ with order at least 2.

Proof. Since $f \in \mathcal{C}^3$ and $\nabla^2 f(\boldsymbol{x}^*)$ is invertible, we know $\exists\, r, c_1, c_2 > 0$, such that $\forall\, \boldsymbol{x} \in B(\boldsymbol{x}^*; r)$, there are

- $\|\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{x}^* - \boldsymbol{x})\| \le c_1 \|\boldsymbol{x}^* - \boldsymbol{x}\|^2$;
- $\nabla^2 f(\boldsymbol{x})$ is invertible;
- $\|(\nabla^2 f(\boldsymbol{x}))^{-1}\| \le c_2$.

Proof (cont). Let $\varepsilon = \min(r, \frac{1}{c_1 c_2}, 1^-)$ (here 1^- means any number slightly smaller than 1).

If $\boldsymbol{x}^{(k)} \in B(\boldsymbol{x}^*; \varepsilon)$, then

$$\begin{aligned}
\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| &= \|\boldsymbol{x}^{(k)} - (\boldsymbol{H}^{(k)})^{-1}\boldsymbol{g}^{(k)} - \boldsymbol{x}^*\| \\
&= \|(\boldsymbol{H}^{(k)})^{-1}(\boldsymbol{H}^{(k)}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) - \boldsymbol{g}^{(k)})\| \\
&\le \|(\boldsymbol{H}^{(k)})^{-1}\| \|\boldsymbol{H}^{(k)}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) - \boldsymbol{g}^{(k)}\| \\
&\le \|(\boldsymbol{H}^{(k)})^{-1}\| \|\boldsymbol{0} - \boldsymbol{g}^{(k)} - \boldsymbol{H}^{(k)}(\boldsymbol{x}^* - \boldsymbol{x}^{(k)})\| \\
&\le c_1 c_2 \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2 \\
&\le \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| \\
&\le \varepsilon
\end{aligned}$$

which implies

$$\boldsymbol{x}^{(k+1)} \in B(\boldsymbol{x}^*; \varepsilon) \quad \text{and} \quad \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \le c_1 c_2 \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2$$

for all k by induction. This implies the convergence is of order at least 2.

Now we consider modifications to overcome the issues of Newton's method.

Issue #1: $\boldsymbol{d}^{(k)} = -(\boldsymbol{H}^{(k)})^{-1}\boldsymbol{g}^{(k)}$ may not be a descent direction.

Theorem. If $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$ and $\boldsymbol{H}^{(k)} \succ \boldsymbol{0}$, then $\boldsymbol{d}^{(k)}$ is a descent direction.

Proof. Let $\boldsymbol{d}^{(k)} = -(\boldsymbol{H}^{(k)})^{-1}\boldsymbol{g}^{(k)}$, and denote $\phi(\alpha) = f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})$.

Then $\phi(0) = f(\boldsymbol{x}^{(k)})$, and

$$\phi'(0) = \nabla f(\boldsymbol{x}^{(k)})^\top \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}(\boldsymbol{H}^{(k)})^{-1}\boldsymbol{g}^{(k)} < 0$$

Therefore, $\exists\, \bar{\alpha} > 0$ such that $\phi(\alpha) < \phi(0)$, i.e.,

$$f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) < f(\boldsymbol{x}^{(k)})$$

for all $\alpha \in (0, \bar{\alpha})$. Therefore $\boldsymbol{d}^{(k)}$ is a descent direction.

Issue #2: $\boldsymbol{H}^{(k)}$ may not be positive definite (or invertible).

Observation. Suppose $\boldsymbol{H}$ is symmetric, then it has eigenvalue decomposition $\boldsymbol{H} = \boldsymbol{U}^\top \boldsymbol{\Lambda} \boldsymbol{U}$ for some orthogonal $\boldsymbol{U}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1 \geq \cdots \geq \lambda_n$.

Let $\mu > \max(0, -\lambda_n)$, then $\lambda_i + \mu > 0$ for all i.

Then $\boldsymbol{H} + \mu \boldsymbol{I} = \boldsymbol{U}^\top (\boldsymbol{\Lambda} + \mu \boldsymbol{I}) \boldsymbol{U} \succ 0$ since all eigenvalues $\lambda_i + \mu > 0$.

Levenberg-Marquardt's modification of Newton's method.

Replace $\boldsymbol{H}^{(k)}$ by $\boldsymbol{H}^{(k)} + \mu_k \boldsymbol{I}$ for sufficiently large $\mu_k > 0$, and

- $\boldsymbol{d}^{(k)} = -(\boldsymbol{H}^{(k)} + \mu_k \boldsymbol{I})^{-1} \boldsymbol{g}^{(k)}$ is a descent direction;

- choose α_k properly such that

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k (\boldsymbol{H}^{(k)} + \mu_k \boldsymbol{I})^{-1} \boldsymbol{g}^{(k)}$$

  is a descent method.

Newton's method for nonlinear least-squares.

Suppose we want to solve

$$\text{minimize } f(\boldsymbol{x}) \quad \text{where} \quad f(\boldsymbol{x}) = \sum_{i=1}^{m} (r_i(\boldsymbol{x}))^2$$

and $r_i : \mathbb{R}^n \to \mathbb{R}$ may not be affine.

Now denote $\boldsymbol{r}(\boldsymbol{x}) = [r_1(\boldsymbol{x}), \dots, r_m(\boldsymbol{x})]^\top \in \mathbb{R}^m$. Then the Jacobian of $\boldsymbol{r} : \mathbb{R}^n \to \mathbb{R}^m$ is

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Note that $f(\boldsymbol{x}) = \|\boldsymbol{r}(\boldsymbol{x})\|^2$, therefore,

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x})$$
$$\nabla^2 f(\boldsymbol{x}) = 2(\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x}))$$

where $\boldsymbol{S}(\boldsymbol{x}) = \sum_{i=1}^{m} r_i(\boldsymbol{x}) \nabla^2 r_i(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$.

In this case, Newton's method yields

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{J}^{(k)\top} \boldsymbol{J}^{(k)} + \boldsymbol{S}^{(k)})^{-1} \boldsymbol{J}^{(k)\top} \boldsymbol{r}^{(k)}$$

where $\boldsymbol{J}^{(k)} = \boldsymbol{J}(\boldsymbol{x}^{(k)}), \boldsymbol{S}^{(k)} = \boldsymbol{S}(\boldsymbol{x}^{(k)}), \boldsymbol{r}^{(k)} = \boldsymbol{r}(\boldsymbol{x}^{(k)})$.

- If $\boldsymbol{S}^{(k)} \approx \mathbf{0}$, then we have

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{J}^{(k)\top}\boldsymbol{J}^{(k)})^{-1}\boldsymbol{J}^{(k)\top}\boldsymbol{r}^{(k)}$$

This is known as the **Gauss-Newton's method**.

- If $\boldsymbol{J}^{(k)\top}\boldsymbol{J}^{(k)}$ is not positive definite, then we modify it:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{J}^{(k)\top}\boldsymbol{J}^{(k)} + \mu_k\boldsymbol{I})^{-1}\boldsymbol{J}^{(k)\top}\boldsymbol{r}^{(k)}$$

This is known as the **Levenberg-Marquardt's method**.