Statistical Optimization: Lecture 10

Newton's method

Zijian Guo

Zhejiang University
Center for Data Science

April 13, 2026

1-dimensional Newton method

Newton's method was originally introduced for solving

$$f(\theta) = 0.$$

Starting from $\theta_0 \in \mathbb{R}$, it computes

$$\theta_{t+1} = \theta_t - \frac{f(\theta_t)}{f'(\theta_t)}, \qquad t \geq 0.$$

Geometric view: θ_{t+1} is the intersection of the tangent line at $(\theta_t, f(\theta_t))$ with the horizontal axis, i.e. it solves

$$f(\theta_t) + f'(\theta_t)(\theta - \theta_t) = 0.$$

Remark. We can apply this idea to find zero gradient.

1-dimensional Newton method

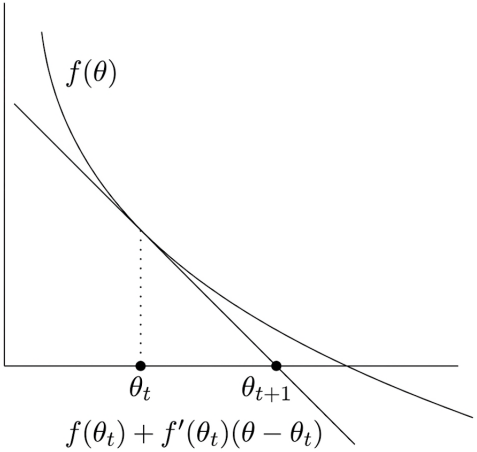

Figure: One step of Newton's method

Newton's method for optimization

Suppose we want to minimize a twice differentiable function

$$f : \mathbb{R} \to \mathbb{R}.$$

In one dimension, this can be reduced to finding a zero of f'. Applying Newton's method to f', we obtain

$$\theta_{t+1} := \theta_t - \frac{f'(\theta_t)}{f''(\theta_t)} = \theta_t - \big(f''(\theta_t)\big)^{-1} f'(\theta_t), \qquad t \geq 0. \tag{1}$$

There is no reason to restrict to $d = 1$. For

$$f : \mathbb{R}^d \to \mathbb{R},$$

Newton's method iterates

$$\theta_{t+1} := \theta_t - \big(\nabla^2 f(\theta_t)\big)^{-1} \nabla f(\theta_t), \qquad t \geq 0. \tag{2}$$

Second-order intuition and gradient descent

To minimize a smooth function f, we approximate it locally near θ_t by a second-order Taylor expansion:

$$f(\theta_t + p_t) \approx f(\theta_t) + \nabla f(\theta_t)^\top p_t + \frac{1}{2} p_t^\top \nabla^2 f(\theta_t) p_t.$$

Minimizing this quadratic model (when $\nabla^2 f(\theta_t)$ is positive definite) gives

$$p_t = -(\nabla^2 f(\theta_t))^{-1} \nabla f(\theta_t).$$

If we replace the Hessian by a scaled identity matrix,

$$\nabla^2 f(\theta_t) \approx \frac{1}{\gamma_t} I,$$

then minimizing this simplified model yields

$$p_t = -\gamma_t \nabla f(\theta_t).$$

which is exactly the gradient descent update.

Newton step minimizes the quadratic model

Theorem

Let f be convex and twice differentiable at θ_t, and assume

$$\nabla^2 f(\theta_t) \succ 0$$

is invertible. Then the Newton iterate θ_{t+1} satisfies

$$\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \Big[f(\theta_t) + \nabla f(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2} (\theta - \theta_t)^\top \nabla^2 f(\theta_t)(\theta - \theta_t) \Big].$$

Proof of the theorem

Proof. Consider the quadratic function

$$m_t(\theta) := f(\theta_t) + \nabla f(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla^2 f(\theta_t)(\theta - \theta_t).$$

Since $\nabla^2 f(\theta_t) \succ 0$, the function m_t is strictly convex and has a unique minimizer. Setting its gradient to zero gives

$$\nabla m_t(\theta) = \nabla f(\theta_t) + \nabla^2 f(\theta_t)(\theta - \theta_t) = 0.$$

Hence

$$\theta = \theta_t - \big(\nabla^2 f(\theta_t)\big)^{-1} \nabla f(\theta_t) = \theta_{t+1}.$$

Therefore θ_{t+1} is the unique minimizer of m_t. □

Gradient step minimizes the simplified quadratic model

Theorem

Let f be differentiable at θ_t, then the gradient descent iterate θ_{t+1} satisfies

$$\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \left[f(\theta_t) + \nabla f(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2\gamma_t} (\theta - \theta_t)^\top (\theta - \theta_t) \right].$$

Theorem (Quadratic Convergence)

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be twice differentiable with a critical point $\theta^\star$. Suppose there is a ball $\Theta \subseteq \mathrm{dom}(f)$ centered at $\theta^\star$ such that:

(i) Bounded inverse Hessians: *$\exists\, \mu > 0$ with*

$$\|(\nabla^2 f(\theta))^{-1}\| \leq \frac{1}{\mu}, \qquad \forall \theta \in \Theta.$$

(ii) Lipschitz continuous Hessians: *$\exists\, B \geq 0$ with*

$$\|\nabla^2 f(\theta) - \nabla^2 f(\eta)\| \leq B\|\theta - \eta\|, \qquad \forall \theta, \eta \in \Theta.$$

(Here $\|\cdot\|$ is the spectral norm.) Then for $\theta_t \in \Theta$ and θ_{t+1} given by Newton's method,

$$\|\theta_{t+1} - \theta^\star\| \leq \frac{B}{2\mu}\|\theta_t - \theta^\star\|^2.$$

Proof. By Newton's method,

$$\theta_{t+1} - \theta^\star = \theta_t - \theta^\star - \big(\nabla^2 f(\theta_t)\big)^{-1} \nabla f(\theta_t).$$

Since $\nabla f(\theta^\star) = 0$, we expand $\nabla f(\theta^\star)$ around θ_t:

$$\nabla f(\theta^\star) = \nabla f(\theta_t) + \int_0^1 \nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big)(\theta^\star - \theta_t)\, ds.$$

Therefore,

$$0 = \nabla f(\theta_t) + \nabla^2 f(\theta_t)(\theta^\star - \theta_t) + \int_0^1 \Big(\nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big) - \nabla^2 f(\theta_t)\Big)(\theta^\star - \theta_t)\, ds.$$

Rearranging gives

$$\nabla f(\theta_t) = \nabla^2 f(\theta_t)(\theta_t - \theta^\star) - \int_0^1 \Big(\nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big) - \nabla^2 f(\theta_t)\Big)(\theta^\star - \theta_t)\, ds.$$

Substituting this identity into the Newton step, we obtain

$$\theta_{t+1} - \theta^\star = \big(\nabla^2 f(\theta_t)\big)^{-1} \int_0^1 \Big(\nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big) - \nabla^2 f(\theta_t)\Big)(\theta^\star - \theta_t)\, ds.$$

Taking norms, and using the spectral norm, gives

$$\|\theta_{t+1} - \theta^\star\| \le \big\|(\nabla^2 f(\theta_t))^{-1}\big\| \int_0^1 \Big\|\Big(\nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big) - \nabla^2 f(\theta_t)\Big)(\theta^\star - \theta_t)\Big\|\, ds.$$

Hence

$$\|\theta_{t+1} - \theta^\star\| \le \big\|(\nabla^2 f(\theta_t))^{-1}\big\| \|\theta^\star - \theta_t\| \int_0^1 \big\|\nabla^2 f\big(\theta_t + s(\theta^\star - \theta_t)\big) - \nabla^2 f(\theta_t)\big\|\, ds.$$

By assumptions (i) and (ii),

$$\|\theta_{t+1} - \theta^\star\| \le \frac{1}{\mu}\|\theta^\star - \theta_t\| \int_0^1 B\, \|s(\theta^\star - \theta_t)\|\, ds \le \frac{B}{2\mu}\|\theta_t - \theta^\star\|^2.$$

Therefore,

$$\|\theta_{t+1} - \theta^\star\| \le \frac{B}{\mu}\|\theta^\star - \theta_t\|^2 \int_0^1 s\, ds = \frac{B}{2\mu}\|\theta_t - \theta^\star\|^2.$$

Convergence of Newton's method

Corollary (Convergence of Newton's method)

With the assumptions of Theorem, if $\theta_0 \in \Theta$ satisfies

$$\|\theta_0 - \theta^\star\| \leq \frac{\mu}{B},$$

then Newton's method yields

$$\|\theta_T - \theta^\star\| \leq \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T-1}, \qquad T \geq 0.$$

Proof. Let $e_t := \|\theta_t - \theta^\star\|$. Theorem gives $e_{t+1} \leq \frac{B}{2\mu} e_t^2$. Assuming $e_t \leq \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^t-1}$,

$$e_{t+1} \leq \frac{B}{2\mu}\left(\frac{\mu}{B}\right)^2\left(\frac{1}{2}\right)^{2(2^t-1)} = \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^{t+1}-1}.$$

□