Newton's Method

Last modified: 2026-10-31 15:22:31 (PDT)

Newton's method

> **Source**
>
> This section is adapted from the Newton-Raphson and gradient descent parts of Zhou [1], used under the MIT License (see the license text in Linear Algebra). The source motivates the Newton step by minimizing a quadratic approximation; here that step is proved to be the exact minimizer of the quadratic model. In the source's zigzag example the starting value of the objective is $\frac{1}{2}(b^2 + b)$, not $\frac{1}{2}(1 + b^2)$ as the source writes; the ratio of successive values is unaffected. In the same example the source's formula for $\underset{\sim}{x}^{(1)}$ has $\underset{\sim}{x}^{(1)}$ on its right side where $\underset{\sim}{x}^{(0)}$ is meant. These parts of the source are not part of this section:
>
> - the quadratic convergence theorem for Newton's method in general, and its remedies for instability (a positive definite substitute for the Hessian, and shortening the Newton step by a search along its direction)
> - its remark that Hessians are expensive to compute
> - the convergence-rate theorem for strongly convex functions
> - gradient descent with momentum, and its optimal parameters
> - Nesterov acceleration, and stochastic gradient descent
> - its code and animations

Definition 0.1 (Iterative method)**.** An **iterative method** [2] starts from a point $\underset{\sim}{x}^{(0)}$ and computes $\underset{\sim}{x}^{(t+1)}$ from $\underset{\sim}{x}^{(t)}$ by a fixed rule, for $t = 0, 1, 2, \ldots$.

Definition 0.2 (Iterate)**.** In an iterative method (Definition 0.1), the points $\underset{\sim}{x}^{(t)}$ are its **iterates**.

Definition 0.3 (Error of an iterate)**.** For a point $\underset{\sim}{x}^*$ that an iterative method (Definition 0.1) is meant to reach, $\|\underset{\sim}{x}^{(t)} - \underset{\sim}{x}^*\|$ is the **error** of the iterate $\underset{\sim}{x}^{(t)}$ (Definition 0.2).

Definition 0.4 (Convergence of a sequence of vectors)**.** A sequence $\underset{\sim}{x}^{(0)}, \underset{\sim}{x}^{(1)}, \ldots$ in $\mathbb{R}^p$ **converges** to $x^* \in \mathbb{R}^p$ if the sequence of distances $\|\underset{\sim}{x}^{(t)} - \underset{\sim}{x}^*\|$ converges to 0. For $p = 1$, $\|x^{(t)} - x^*\| = |x^{(t)} - x^*|$, and this is the convergence of a sequence of numbers to x^*.

Example 0.1 (Halving a vector)**.** The iterative method $\underset{\sim}{x}^{(t+1)} = \frac{1}{2}\,\underset{\sim}{x}^{(t)}$ from $\underset{\sim}{x}^{(0)} = (4,2)^\top$ has iterates $(4,2)^\top, (2,1)^\top, (1,0.5)^\top, \ldots$, that is, $\underset{\sim}{x}^{(t)} = 2^{-t}\,(4,2)^\top$. With $\underset{\sim}{x}^* = \underset{\sim}{0}$, the error of $\underset{\sim}{x}^{(t)}$ is

$$\begin{aligned} \left\|2^{-t}\,(4,2)^\top - \underset{\sim}{0}\right\| &= 2^{-t}\left\|(4,2)^\top\right\| && \text{(norm of a multiple)} \\ &= 2^{-t}\sqrt{20}, && (4^2 + 2^2 = 20) \end{aligned}$$

which converges to 0, so the iterates converge to $\underset{\sim}{0}$.

Definition 0.5 (Newton's method for minimization)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ have first and second partial derivatives, and let $\underset{\sim}{x}^{(0)}$ be a starting point. **Newton's method** [3] computes, for $t = 0, 1, 2, \ldots$, as long as $\mathbf{H}_f(\underset{\sim}{x}^{(t)})$ is invertible,

$$\underset{\sim}{x}^{(t+1)} \overset{\text{def}}{=} \underset{\sim}{x}^{(t)} - \left[\mathbf{H}_f(\underset{\sim}{x}^{(t)})\right]^{-1} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t)}).$$

Definition 0.6 (Newton step)**.** In Newton's method (Definition 0.5), the vector $-\left[\mathbf{H}_f(\underset{\sim}{x}^{(t)})\right]^{-1} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t)})$ is the **Newton step** at $\underset{\sim}{x}^{(t)}$.

Example 0.2 (Newton's method for a cubic)**.** Let $f(x) = \frac{1}{3}x^3 - 4x$, so $f'(x) = x^2 - 4$ and $f''(x) = 2x$. For $x^{(t)} \neq 0$, one Newton step gives

$$\begin{aligned} x^{(t+1)} &= x^{(t)} - \frac{(x^{(t)})^2 - 4}{2x^{(t)}} && \text{(Newton's method with } p = 1) \\ &= \frac{2(x^{(t)})^2 - (x^{(t)})^2 + 4}{2x^{(t)}} && \text{(common denominator)} \\ &= \frac{1}{2}\left(x^{(t)} + \frac{4}{x^{(t)}}\right). && \text{(simplify)} \end{aligned}$$

From $x^{(0)} = 2.5$:

- $$x^{(1)} = \tfrac{1}{2}(2.5 + 1.6) \\ = 2.05;$$
- $$x^{(2)} = \tfrac{1}{2}(2.05 + 1.951219\ldots) \\ = 2.000610$$

  (to six decimals);
- $x^{(3)} = 2.0000001$ (to seven decimals).

The iterates (Definition 0.2) approach 2, the local minimizer of f (a stationary point, since $f'(2) = 0$, with $f''(2) = 4 > 0$; Theorem 7 in Critical Points and Optimality Conditions).

Definition 0.7 (Quadratic model, second-order Taylor polynomial)**.** Let f be as in Definition 0.5 and let $\underset{\sim}{x}$ be a point, with $g \overset{\text{def}}{=} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x})$ and $\mathbf{H} \overset{\text{def}}{=} \mathbf{H}_f(\underset{\sim}{x})$. The **quadratic model** of f at $\underset{\sim}{x}$ is the function of a step $\underset{\sim}{\delta} \in \mathbb{R}^{\tilde{p}}$

$$q(\underset{\sim}{\delta}) \overset{\text{def}}{=} f(\underset{\sim}{x}) + \underset{\sim}{g} \cdot \underset{\sim}{\delta} + \tfrac{1}{2}\, \underset{\sim}{\delta}^{\top} \mathbf{H} \underset{\sim}{\delta},$$

also called the **second-order Taylor polynomial** of f at $\underset{\sim}{x}$.

Example 0.3 (The model at $x = 2.5$)**.** For $f(x) = \frac{1}{3}x^3 - 4x$ at $x = 2.5$ (Example 0.2),

$$g = 2.5^2 - 4 \\ = 2.25$$

and

$$H = [2 \cdot 2.5] \\ = [5],$$

so $q(\delta) = f(2.5) + 2.25\,\delta + 2.5\,\delta^2$. For $f(\underset{\sim}{x}) = \frac{1}{2}(x_1^2 + 3x_2^2)$, whose Hessian is $\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$ at every point, the model at any $\underset{\sim}{x}$ equals $f(\underset{\sim}{x} + \underset{\sim}{\delta})$ exactly (Example 11 in Critical Points and Optimality Conditions with $\mathbf{S} = \begin{bmatrix} 1/2 & 0 \\ 0 & 3/2 \end{bmatrix}$).

Remark (The quadratic model and Taylor's theorem)

Remark 0.1 (The quadratic model and Taylor's theorem). When the second partial derivatives of f are continuous, Theorem 6 in Critical Points and Optimality Conditions says that $f(\underset{\sim}{x}+\underset{\sim}{\delta})$ equals the expression for $q(\underset{\sim}{\delta})$ in Definition 0.7 with the Hessian evaluated at a point between $\underset{\sim}{x}$ and $\underset{\sim}{x}+\underset{\sim}{\delta}$.

Theorem 0.1 (The Newton step minimizes the quadratic model)**.** *Let q be the quadratic model of f at $\underset{\sim}{x}$ (Definition 0.7), with $\mathbf{H}=\mathbf{H}_f(\underset{\sim}{x})$ positive definite. Its unique minimizer is the Newton step $\underset{\sim}{\delta}^*=-\mathbf{H}^{-1}\underset{\sim}{g}$.*

Proof

Proof. $\mathbf{H}$ is symmetric, by the definition of positive definite, and invertible (positive definite inverse), so $\underset{\sim}{\delta}^*$ exists and $\mathbf{H}\underset{\sim}{\delta}^*=-\underset{\sim}{g}$. For any $\underset{\sim}{d}\in\mathbb{R}^p$, first

$$\begin{aligned}\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{\delta}^* &= -\underset{\sim}{d}\cdot\underset{\sim}{g} \quad (\mathbf{H}\underset{\sim}{\delta}^*=-\underset{\sim}{g})\\ &= -\underset{\sim}{g}\cdot\underset{\sim}{d}, \quad (\text{a } 1\times 1 \text{ matrix equals its transpose})\end{aligned}$$

and $\underset{\sim}{\delta}^{*\top}\mathbf{H}\underset{\sim}{d}=\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{\delta}^*$ in the same way, using $\mathbf{H}^{\top}=\mathbf{H}$. Then

$$\begin{aligned}q(\underset{\sim}{\delta}^*+\underset{\sim}{d}) &= f(\underset{\sim}{x})+\underset{\sim}{g}\cdot(\underset{\sim}{\delta}^*+\underset{\sim}{d})+\tfrac{1}{2}(\underset{\sim}{\delta}^*+\underset{\sim}{d})^{\top}\mathbf{H}(\underset{\sim}{\delta}^*+\underset{\sim}{d}) && \text{(definition of } q)\\ &= f(\underset{\sim}{x})+\underset{\sim}{g}\cdot\underset{\sim}{\delta}^*+\underset{\sim}{g}\cdot\underset{\sim}{d}+\tfrac{1}{2}\left(\underset{\sim}{\delta}^{*\top}\mathbf{H}\underset{\sim}{\delta}^*+\underset{\sim}{\delta}^{*\top}\mathbf{H}\underset{\sim}{d}+\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{\delta}^*+\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{d}\right) && \text{(distribute)}\\ &= f(\underset{\sim}{x})+\underset{\sim}{g}\cdot\underset{\sim}{\delta}^*+\underset{\sim}{g}\cdot\underset{\sim}{d}+\tfrac{1}{2}\,\underset{\sim}{\delta}^{*\top}\mathbf{H}\underset{\sim}{\delta}^*-\underset{\sim}{g}\cdot\underset{\sim}{d}+\tfrac{1}{2}\,\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{d} && \text{(both cross terms equal } -\underset{\sim}{g}\cdot\underset{\sim}{d})\\ &= f(\underset{\sim}{x})+\underset{\sim}{g}\cdot\underset{\sim}{\delta}^*+\tfrac{1}{2}\,\underset{\sim}{\delta}^{*\top}\mathbf{H}\underset{\sim}{\delta}^*+\tfrac{1}{2}\,\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{d} && \text{(cancel } \underset{\sim}{g}\cdot\underset{\sim}{d})\\ &= q(\underset{\sim}{\delta}^*)+\tfrac{1}{2}\,\underset{\sim}{d}^{\top}\mathbf{H}\underset{\sim}{d}, && \text{(definition of } q)\end{aligned}$$

which is greater than $q(\underset{\sim}{\delta}^*)$ for every $\underset{\sim}{d}\neq\underset{\sim}{0}_{p\times 1}$, since $\mathbf{H}$ is positive definite. □

Example 0.4 (The Newton step at $x=2.5$)**.** For the model $q(\delta)=f(2.5)+2.25\,\delta+2.5\,\delta^2$ of Example 0.3, $H=[5]$ is positive definite, so by Theorem 0.1 its minimizer is

$$\begin{aligned}\delta^* &= -H^{-1}g\\ &= -2.25/5\\ &= -0.45.\end{aligned}$$

Setting the derivative $2.25+5\delta$ to 0 gives the same δ^*, and

$$\begin{aligned} 2.5 - 0.45 &= 2.05 \\ &= x^{(1)} \end{aligned}$$

(Example 0.2).

Definition 0.8 (Quadratic convergence)**.** A sequence $\underset{\sim}{x}^{(0)}, \underset{\sim}{x}^{(1)}, \ldots$ in $\mathbb{R}^p$ that converges (Definition 0.4) to $\underset{\sim}{x}^*$ **converges quadratically** [4] if there is a constant C with

$$\left\|\underset{\sim}{x}^{(t+1)} - \underset{\sim}{x}^*\right\| \le C \left\|\underset{\sim}{x}^{(t)} - \underset{\sim}{x}^*\right\|^2 \qquad \text{for every } t.$$

Example 0.5 (Halving the error is not quadratic convergence)**.** $x^{(t)} = 2^{-t}$ converges to 0, but it does not converge quadratically: the bound would need $2^{-(t+1)} \le C\,(2^{-t})^2 = C\,2^{-2t}$, that is, $2^{t-1} \le C$, which fails once $t > 1 + \log_2 C$. Each step only halves the error.

Example 0.6 (The error squares at each step)**.** For the iteration $x^{(t+1)} = \frac{1}{2}(x^{(t)} + 4/x^{(t)})$ of Example 0.2,

$$\begin{aligned} x^{(t+1)} - 2 &= \frac{1}{2}\left(x^{(t)} + \frac{4}{x^{(t)}}\right) - 2 && \text{(the iteration)} \\ &= \frac{(x^{(t)})^2 + 4 - 4x^{(t)}}{2x^{(t)}} && \text{(common denominator)} \\ &= \frac{(x^{(t)} - 2)^2}{2x^{(t)}}. && \text{(the numerator is a perfect square)} \end{aligned}$$

So near 2 the error is roughly squared and divided by 4 at each step: the errors are 0.05, then 0.00061, then 0.000000093, and $0.05^2/(2 \cdot 2.05) \approx 0.00061$. The iterates converge quadratically (Definition 0.8): $C = \frac{1}{2}$ works for every $x^{(t)} \ge 1$, since then $\frac{1}{2x^{(t)}} \le \frac{1}{2}$, and every iterate from 2.5 is at least 2, since the identity makes $x^{(t+1)} - 2 \ge 0$ whenever $x^{(t)} > 0$. Zhou [1] states such a bound for Newton's method in general; it holds from starting points close enough to a minimizer where the Hessian is positive definite, when near that minimizer the Hessian changes by at most a constant times the distance moved, which these notes do not prove.

Example 0.7 (Newton's method can find a maximum)**.** From $x^{(0)} = -2.5$ the same iteration gives -2.05, -2.000610, -2.0000001, approaching -2. But -2 is not a minimizer of f: it is a stationary point of $-f$, whose second derivative there is $-f''(-2) = 4 > 0$, so it is a strict local minimizer of $-f$ (Theorem 7 in Critical Points and Optimality Conditions), that is, f is strictly larger at -2 than at all nearby points. Newton's method looks only for a stationary point, and here the Hessian $f''(x) = 2x$ is negative along the way, so the

quadratic model has no minimizer (with $H < 0$, $q(\delta) = f + g\,\delta + \frac{1}{2}H\delta^2$ decreases without bound) and Theorem 0.1 does not apply: the positive definite hypothesis there is needed.

Theorem 0.2 (Newton's method minimizes a quadratic in one step)**.** *Let* $\mathbf{S}$ *be a* $p \times p$ *positive definite matrix,* $\underset{\sim}{c} \in \mathbb{R}^p$*, and* $f(\underset{\sim}{x}) = \frac{1}{2}\,\underset{\sim}{x}^\top \mathbf{S}\underset{\sim}{x} - \underset{\sim}{c} \cdot \underset{\sim}{x}$*. From any* $\underset{\sim}{x}^{(0)}$*, Newton's method (Definition 0.5) gives* $\underset{\sim}{x}^{(1)} = \mathbf{S}^{-1}\underset{\sim}{c}$*, the unique minimizer of* f*.*

Proof

Proof. Partial derivatives of a sum or of a constant multiple are the sum or multiple of the partial derivatives, so by derivative of a quadratic form for $\frac{1}{2}\,\underset{\sim}{x}^\top \mathbf{S}\underset{\sim}{x}$ and derivative of a linear combination for $\underset{\sim}{c} \cdot \underset{\sim}{x}$,

$$\begin{aligned}\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}) &= \tfrac{1}{2}(2\mathbf{S}\underset{\sim}{x}) - \underset{\sim}{c} \\ &= \mathbf{S}\underset{\sim}{x} - \underset{\sim}{c};\end{aligned}$$

by Hessian of a quadratic form with $\frac{1}{2}\mathbf{S}$ in place of $\mathbf{S}$, and because the gradient $-\underset{\sim}{c}$ of the linear term is constant (constant with respect to a vector), $\mathbf{H}_f(\underset{\sim}{x}) = \mathbf{S}$, which is invertible (positive definite inverse). So

$$\begin{aligned}\underset{\sim}{x}^{(1)} &= \underset{\sim}{x}^{(0)} - \mathbf{S}^{-1}(\mathbf{S}\underset{\sim}{x}^{(0)} - \underset{\sim}{c}) && \text{(definition of Newton's method)} \\ &= \underset{\sim}{x}^{(0)} - \mathbf{S}^{-1}\mathbf{S}\underset{\sim}{x}^{(0)} + \mathbf{S}^{-1}\underset{\sim}{c} && \text{(distribute)} \\ &= \underset{\sim}{x}^{(0)} - \underset{\sim}{x}^{(0)} + \mathbf{S}^{-1}\underset{\sim}{c} && (\mathbf{S}^{-1}\mathbf{S} = \mathbf{I}_p) \\ &= \mathbf{S}^{-1}\underset{\sim}{c}. && \text{(cancel)}\end{aligned}$$

The gradient at $\mathbf{S}^{-1}\underset{\sim}{c}$ is $\mathbf{S}\mathbf{S}^{-1}\underset{\sim}{c} - \underset{\sim}{c} = \underset{\sim}{0}_{p\times 1}$, and f is strictly convex (Theorem 5 in Convexity in Several Variables, part 2), so $\mathbf{S}^{-1}\underset{\sim}{c}$ is a global minimizer (Corollary 1 in Convexity in Several Variables). It is the only one: a strictly convex function has at most one global minimizer, since the midpoint of two distinct global minimizers, which have equal values, would have a strictly smaller value (Definition 3 in Convexity in Several Variables with $t = \frac{1}{2}$). □

Example 0.8 (One step to the minimizer)**.** Let $f(\underset{\sim}{x}) = \frac{1}{2}(x_1^2 + 3x_2^2)$, so $\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$ and $\underset{\sim}{c} = \underset{\sim}{0}_{2\times 1}$. From $\underset{\sim}{x}^{(0)} = (3,1)^\top$ the gradient is $\mathbf{S}\underset{\sim}{x}^{(0)} = (3,3)^\top$, and $\mathbf{S}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}$, so the Newton step is $-\mathbf{S}^{-1}(3,3)^\top = -(3,1)^\top$ and

$$\begin{aligned} \underset{\sim}{x}^{(1)} &= (3,1)^\top - (3,1)^\top \\ &= \underset{\sim}{0}_{2\times 1}, \end{aligned}$$

the minimizer.

[1] H. Zhou, "Optimization and multivariate calculus." Lecture notes for Biostat 216, Mathematical Methods for Biostatistics, University of California, Los Angeles, 2024. Accessed: Oct. 05, 2026. [Online]. Available: https://ucla-biostat-216.github.io/2024fall/slides/13-optim/13-optim.html

[2] Wikipedia contributors, "Iterative method — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Iterative_method

[3] Wikipedia contributors, "Newton's method in optimization — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization

[4] Wikipedia contributors, "Rate of convergence — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Rate_of_convergence