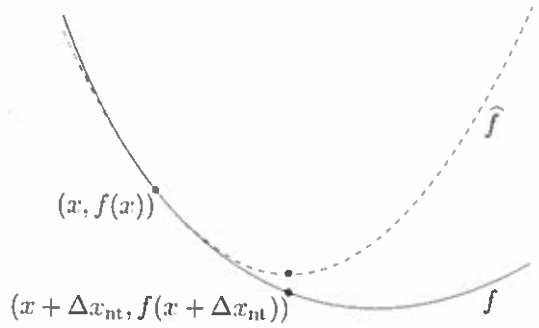

Figure 9.16 The function f (shown solid) and its second-order approximation $\widehat{f}$ at x (dashed). The Newton step $\Delta x_{\rm nt}$ is what must be added to x to give the minimizer of $\widehat{f}$.

9.5 Newton's method

9.5.1 The Newton step

For $x \in \mathbf{dom}\, f$, the vector

$$\Delta x_{\rm nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

is called the *Newton step* (for f, at x). Positive definiteness of $\nabla^2 f(x)$ implies that

$$\nabla f(x)^T \Delta x_{\rm nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$

unless $\nabla f(x) = 0$, so the Newton step is a descent direction (unless x is optimal). The Newton step can be interpreted and motivated in several ways.

Minimizer of second-order approximation

The second-order Taylor approximation (or model) $\widehat{f}$ of f at x is

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v, \tag{9.28}$$

which is a convex quadratic function of v, and is minimized when $v = \Delta x_{\rm nt}$. Thus, the Newton step $\Delta x_{\rm nt}$ is what should be added to the point x to minimize the second-order approximation of f at x. This is illustrated in figure 9.16.

This interpretation gives us some insight into the Newton step. If the function f is quadratic, then $x + \Delta x_{\rm nt}$ is the exact minimizer of f. If the function f is nearly quadratic, intuition suggests that $x + \Delta x_{\rm nt}$ should be a very good estimate of the minimizer of f, *i.e.*, $x^\star$. Since f is twice differentiable, the quadratic model of f will be very accurate when x is near $x^\star$. It follows that when x is near $x^\star$, the point $x + \Delta x_{\rm nt}$ should be a very good estimate of $x^\star$. We will see that this intuition is correct.

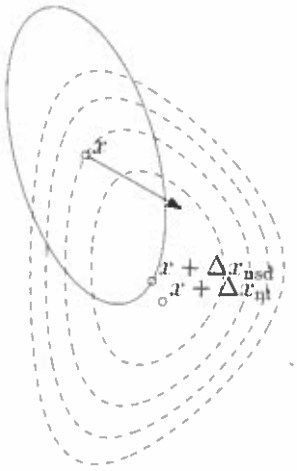

Figure 9.17 The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x + v \mid v^T \nabla^2 f(x) v \leq 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step Δx_{nt} is the steepest descent direction in the norm $\|\cdot\|_{\nabla^2 f(x)}$. The figure also shows Δx_{nsd}, the normalized steepest descent direction for the same norm.

Steepest descent direction in Hessian norm

The Newton step is also the steepest descent direction at x, for the quadratic norm defined by the Hessian $\nabla^2 f(x)$, *i.e.*,

$$\|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}.$$

This gives another insight into why the Newton step should be a good search direction, and a very good search direction when x is near $x^\star$.

Recall from our discussion above that steepest descent, with quadratic norm $\|\cdot\|_P$, converges very rapidly when the Hessian, after the associated change of coordinates, has small condition number. In particular, near $x^\star$, a very good choice is $P = \nabla^2 f(x^\star)$. When x is near $x^\star$, we have $\nabla^2 f(x) \approx \nabla^2 f(x^\star)$, which explains why the Newton step is a very good choice of search direction. This is illustrated in figure 9.17.

Solution of linearized optimality condition

If we linearize the optimality condition $\nabla f(x^\star) = 0$ near x we obtain

$$\nabla f(x + v) \approx \nabla f(x) + \nabla^2 f(x) v = 0,$$

which is a linear equation in v, with solution $v = \Delta x_{\text{nt}}$. So the Newton step Δx_{nt} is what must be added to x so that the linearized optimality condition holds. Again, this suggests that when x is near $x^\star$ (so the optimality conditions almost hold), the update $x + \Delta x_{\text{nt}}$ should be a very good approximation of $x^\star$.

When $n = 1$, *i.e.*, $f : \mathbf{R} \to \mathbf{R}$, this interpretation is particularly simple. The solution $x^\star$ of the minimization problem is characterized by $f'(x^\star) = 0$, *i.e.*, it is

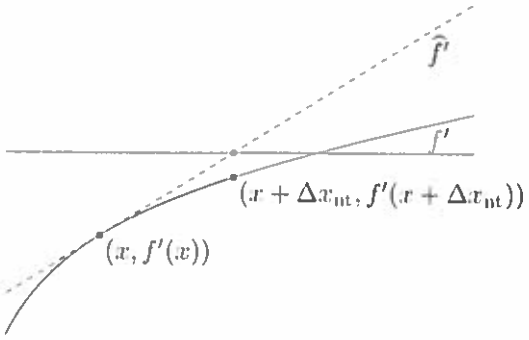

Figure 9.18 The solid curve is the derivative f' of the function f shown in figure 9.16. $\widehat{f}'$ is the linear approximation of f' at x. The Newton step $\Delta x_{\rm nt}$ is the difference between the root of $\widehat{f}'$ and the point x.

the zero-crossing of the derivative f', which is monotonically increasing since f is convex. Given our current approximation x of the solution, we form a first-order Taylor approximation of f' at x. The zero-crossing of this affine approximation is then $x + \Delta x_{\rm nt}$. This interpretation is illustrated in figure 9.18.

Affine invariance of the Newton step

An important feature of the Newton step is that it is independent of linear (or affine) changes of coordinates. Suppose $T \in \mathbf{R}^{n\times n}$ is nonsingular, and define $\bar{f}(y) = f(Ty)$. Then we have

$$\nabla \bar{f}(y) = T^T \nabla f(x), \qquad \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T,$$

where $x = Ty$. The Newton step for $\bar{f}$ at y is therefore

$$\begin{aligned} \Delta y_{\rm nt} &= -\left(T^T \nabla^2 f(x) T\right)^{-1} \left(T^T \nabla f(x)\right) \\ &= -T^{-1} \nabla^2 f(x)^{-1} \nabla f(x) \\ &= T^{-1} \Delta x_{\rm nt}, \end{aligned}$$

where $\Delta x_{\rm nt}$ is the Newton step for f at x. Hence the Newton steps of f and $\bar{f}$ are related by the same linear transformation, and

$$x + \Delta x_{\rm nt} = T(y + \Delta y_{\rm nt}).$$

The Newton decrement

The quantity

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$$

is called the *Newton decrement* at x. We will see that the Newton decrement plays an important role in the analysis of Newton's method, and is also useful

as a stopping criterion. We can relate the Newton decrement to the quantity $f(x) - \inf_y \widehat{f}(y)$, where $\widehat{f}$ is the second-order approximation of f at x:

$$f(x) - \inf_y \widehat{f}(y) = f(x) - \widehat{f}(x + \Delta x_{\rm nt}) = \frac{1}{2}\lambda(x)^2.$$

Thus, $\lambda^2/2$ is an estimate of $f(x) - p^*$, based on the quadratic approximation of f at x.

We can also express the Newton decrement as

$$\lambda(x) = \left(\Delta x_{\rm nt}^T \nabla^2 f(x) \Delta x_{\rm nt}\right)^{1/2}. \tag{9.29}$$

This shows that λ is the norm of the Newton step, in the quadratic norm defined by the Hessian, *i.e.*, the norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}.$$

The Newton decrement comes up in backtracking line search as well, since we have

$$\nabla f(x)^T \Delta x_{\rm nt} = -\lambda(x)^2. \tag{9.30}$$

This is the constant used in a backtracking line search, and can be interpreted as the directional derivative of f at x in the direction of the Newton step:

$$-\lambda(x)^2 = \nabla f(x)^T \Delta x_{\rm nt} = \frac{d}{dt} f(x + \Delta x_{\rm nt} t)\Big|_{t=0}.$$

Finally, we note that the Newton decrement is, like the Newton step, affine invariant. In other words, the Newton decrement of $\bar{f}(y) = f(Ty)$ at y, where T is nonsingular, is the same as the Newton decrement of f at $x = Ty$.

9.5.2 Newton's method

Newton's method, as outlined below, is sometimes called the *damped* Newton method or *guarded* Newton method, to distinguish it from the *pure* Newton method, which uses a fixed step size $t = 1$.

Algorithm 9.5 *Newton's method.*

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.

repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\rm nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$; $\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\rm nt}$.

This is essentially the general descent method described in §9.2, using the Newton step as search direction. The only difference (which is very minor) is that the stopping criterion is checked after computing the search direction, rather than after the update.