A few remarks about optimization

Optimization means maximizing or minimizing a given function over a given domain. In statistics, one main use is in maximum likelihood estimation. There is also the related notion of M-estimation, which is as follows. Let $M(x, \theta)$ be a function of a variable x and a parameter θ. Let $X_1, ..., X_n$ be observations. Then an M-estimate of θ given the sample is a value that minimizes $\sum_{j=1}^{n} M(X_j, \theta)$. For example if $f(x, \theta)$ is a probability density function or mass function of x given θ, and $M(x, \theta) = -\log f(x, \theta)$, then an M-estimate is a maximum likelihood estimate. If $M(x, \theta) = |x - \theta| - |x|$, then in dimension $d = 1$, an M-estimate is a sample median of $(X_1, ..., X_n)$, which is unique if n is odd but not usually if n is even. In dimension $d \geq 2$, the M-estimate is called the "spatial median" and is unique unless the sample is concentrated in a line and there has a non-unique median.

Let's begin with a seemingly different problem: solving an equation $F(x) = u$ for x given u, numerically, where F can be computed and has a derivative F' that can also be computed. First suppose that F is a probability distribution function with $F'(x) = f(x) > 0$ at all x where $0 < F(x) < 1$ on the real line and $0 < u < 1$. Suppose we don't have an explicit form of the inverse function F^{-1}, which we might want use, for example, in generating $X_1, X_2, ...$ i.i.d. (F) via $X_j = F^{-1}(U_j)$ for U_j i.i.d. $U[0, 1]$. We also may want a way to compute quantiles of F, even if the inversion method is not the most efficient way to generate X_j i.i.d. (F).

The very well known Newton–Raphson method proceeds as follows. Start with some x_0. Given x_n, find $F(x_n)$. If this is not equal to u to the desired accuracy, then find x_{n+1} as follows. Use the first-order Taylor approximation as $x \to x_n$,

$$F(x) = F(x_n) + F'(x_n)(x - x_n) + o(|x - x_n|).$$

Neglecting the remainder term and setting $F(x) = u$ gives

$$x = x_{n+1} := x_n + (u - F(x_n))/F'(x_n).$$

If x_∞ is the solution with $F(x_\infty) = u$ and if we are at an x_n in a neighborhood of x_∞ where the first-order Taylor approximation works well enough and F has a bounded second derivative, then x_m converge very fast to x_∞, with $|x_{m+1} - x_\infty| = O(|x_m - x_\infty|^2)$. If, however, we started too far away from x_∞, then the method may overshoot, and x_{n+1} may actually be farther away from x_∞ than x_n is. So the method needs to be supplemented by another one. One can first find a_0 and b_0 such that $F(a_0) < u < F(b_0)$. Let x_0 be one of a_0 or b_0. Generate a candidate x_1 by Newton–Raphson and if it's in the interval $[a_0, b_0]$, accept it. Otherwise reject it and set $x_1 = (a_0 + b_0)/2$ (bisect). If $F(x_1) < u$ let $a_1 = x_1$ and $b_1 = b_0$, or if $F(x_1) > u$ let $a_1 = a_0$ and $b_1 = x_1$. If $F(x_1) = u$ to the desired accuracy, we're done. Iterate this process. At each step, we either get a Newton–Raphson improvement in the error (of the order of squaring it) or "at worst" we cut the error in half. This is a satisfactory rate of convergence.

Now, suppose we have an actual optimization problem. Let's say for example we want to find a maximum of a function f on some interval $[a, b]$. The usual method is to look for a point ξ with $a < \xi < b$ where the derivative $f'(\xi) = 0$, and check the values of f at all such ξ and at the endpoints a, b to see which is largest. Suppose we can compute f' and f'' but it's hard to solve $f'(x) = 0$ explicitly. First let f be unimodal on $[a, b]$. Then f is nondecreasing, so $f' \geq 0$, on $[a, \xi]$ and nonincreasing, so $f' \leq 0$, on $[\xi, b]$. To have a relative maximum of f at a ξ for which $f'(\xi) = 0$, it's necessary that $f''(\xi) \leq 0$ and sufficient that $f''(\xi) < 0$. So suppose that $f''(x) < 0$ for $c < x < d$ such that $a \leq c < d \leq b$ where $f'(c) > 0$ and $f'(d) < 0$. To solve for ξ by the Newton–Raphson method, restricting to the interval $[c, d]$, f' plays the role of F (decreasing, rather than increasing) and f'', of F'. Then the method works well. But if $a < x < c$ or $d < x < b$, then $f''(x)$ may be positive and the method starting at such an x may fail, so we may again need to do bisection.

There may in general be multiple critical points where $f'(x) = 0$. In particular, a density f may have multiple modes, or a likelihood function may have multiple relative maxima or at least extrema. So, even in one dimension, we can't restrict ourselves to Newton–Raphson and bisection methods. Some method is needed for escaping the neighborhood of a local maximum to search for other maxima, and in general no better way is known than to make random, Metropolis–Hastings type moves, as in simulated annealing, on which there is a handout.

In more than one dimension, a Newton–Raphson type method exists, where we're seeking "critical points" where the gradient of f is 0, and the analogue of the second derivative is the Hessian matrix of second partial derivatives. In general the method doesn't work as well as in one dimension, so instead, there is a so-called "gradient descent" method, if the problem is a minimization rather than maximization, moving in a direction where one decreases the function as fast as possible. Again there may be multiple critical points and one needs to supplement the derivative-based methods with random search.

A critical point can be unique without giving a maximum or minimum, as with $f(x) = x^3$ in one dimension. A possibly more surprising example is, in two dimensions, the function $f(x, y) = x^2(1 + y)^3 + y^2$, a polynomial of degree 5. It has a unique critical point at $(0, 0)$ which is a strict relative minimum, but it is not an absolute minimum since $f(1, y) \to -\infty$ as $y \to -\infty$. So, in statistics, one should beware of definitions in terms of critical points, although I don't know of such an example relating to statistics. (The example was found in the paper by Durfee et al., which by the way resulted from an REU [Research Experience for Undergraduates] project.)

REFERENCES

Durfee, A., Kronenfeld, N., Munson, H., Roy, J., and Westby, I. (1993). Counting critical points of real polynomials in two variables. *Amer. Math. Monthly* **100**, 255-271.