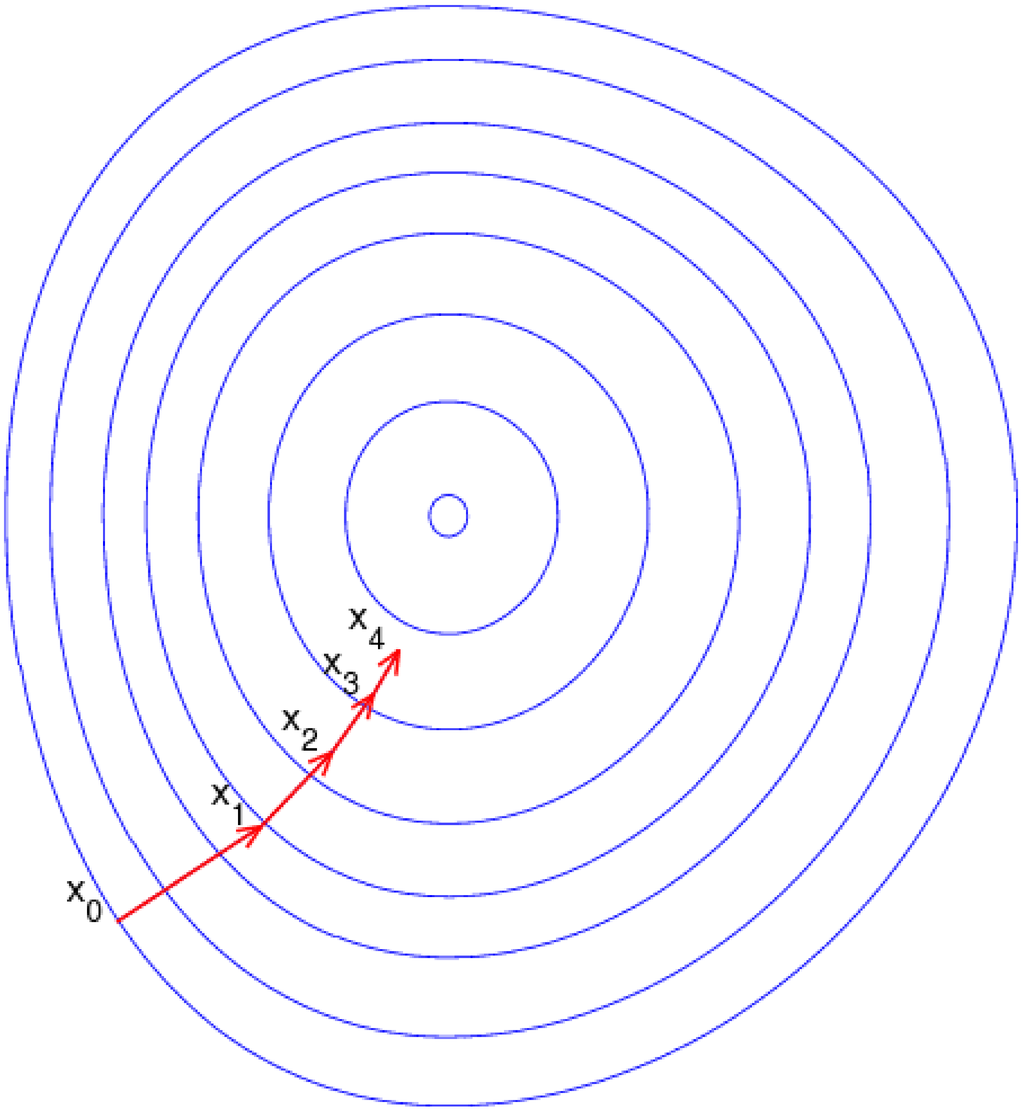
Jang, Sun, and Mizutani
Neuro-Fuzzy and Soft Computing
Chapter 6
Derivative-Based Optimization

Outline

1. Gradient Descent
2. The Newton-Raphson Method
3. The Levenberg–Marquardt Algorithm
4. Trust Region Methods

Contour plot



Gradient descent: head downhill

http://en.wikipedia.org/wiki/Gradient_descent

Fuzzy controller optimization: Find the MF parameters that minimize tracking error

min $E(\theta)$ with respect to θ

θ = n-element vector of MF parameters

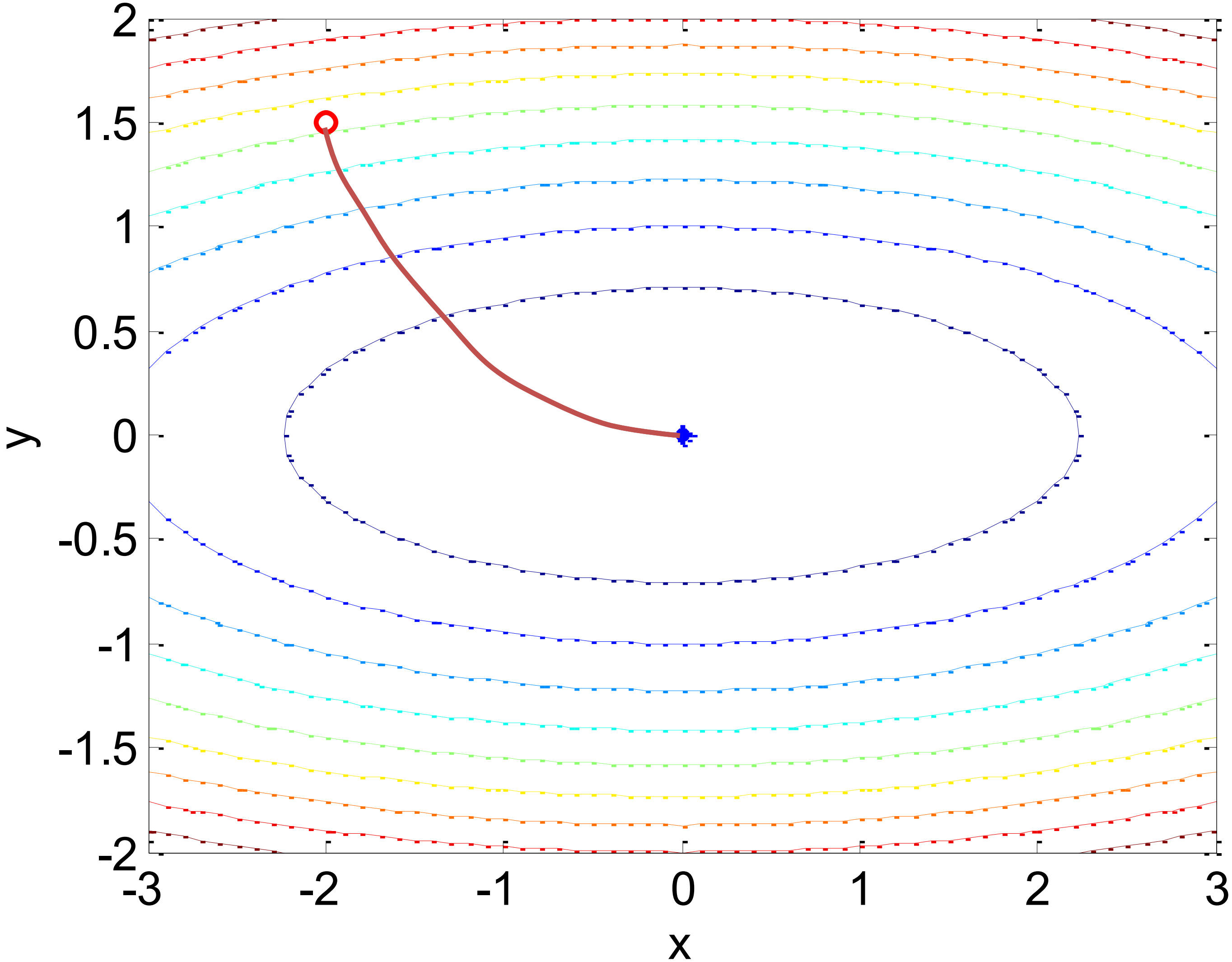
$E(\theta)$ = controller tracking error

$$\theta_{k+1} = \theta_k - \eta \frac{\partial E}{\partial \theta_k}$$

$$\frac{\partial E}{\partial \theta_k} = \left[\frac{\partial E}{\partial \theta_{1k}} \quad \cdots \quad \frac{\partial E}{\partial \theta_{nk}} \right]^T = \left. \frac{\partial E}{\partial \theta} \right|_{\theta=\theta_k}$$

$$\eta = \text{step size}$$

$$k = \text{step number}$$

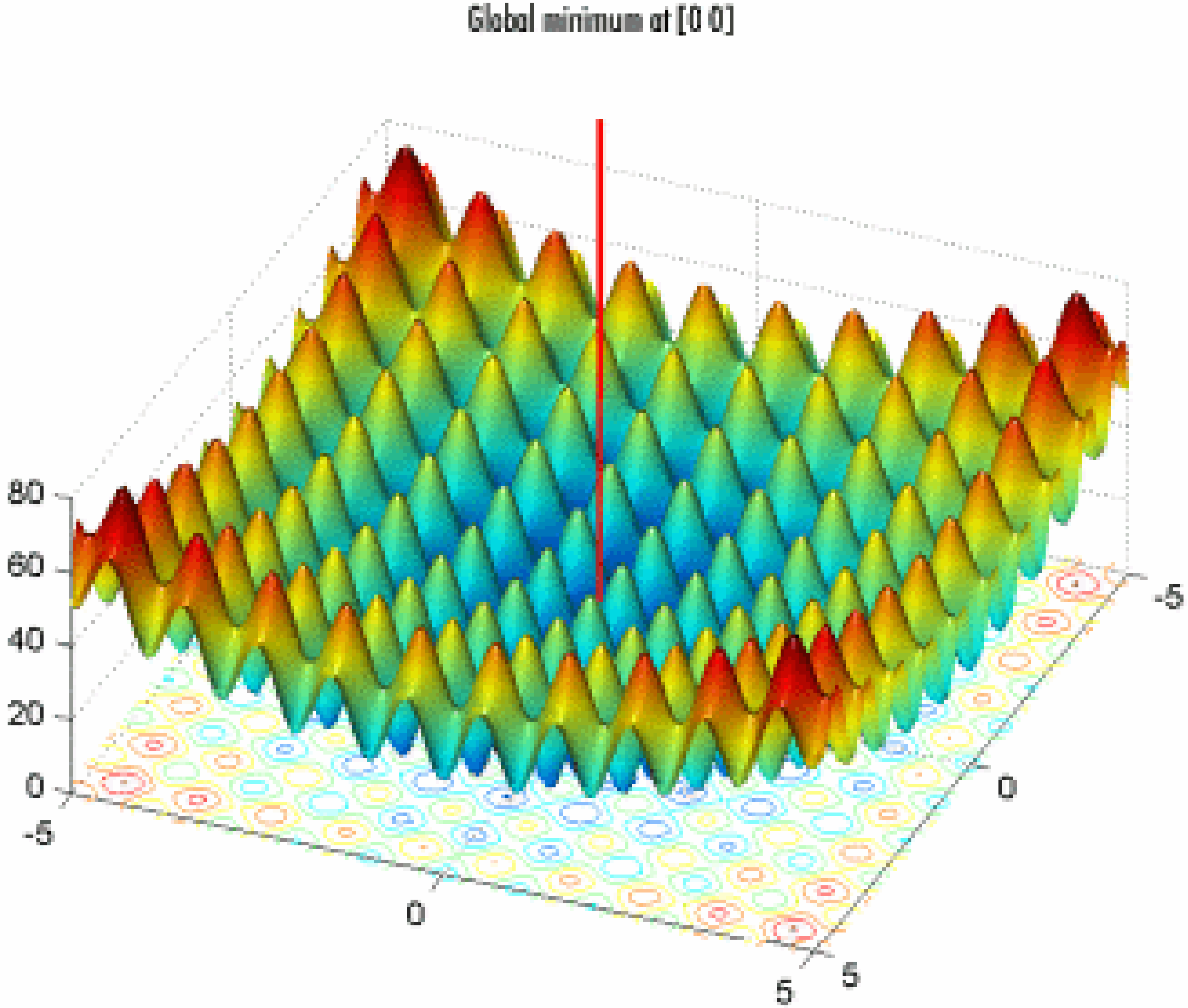


Contour plot of x^2+10y^2

η too small: convergence takes long time

η too large: overshoot minimum

```
x=-3: 0.1: 3; y=-2: 0.1: 2;
for i=1:length(x), for j=1:length(y), z(i,j)=x(i)^2+10*y(j)^2; end, end
contour(x,y,z)
```



Gradient descent is a *local* optimization method (Rastrigin function)

Step Size Selection

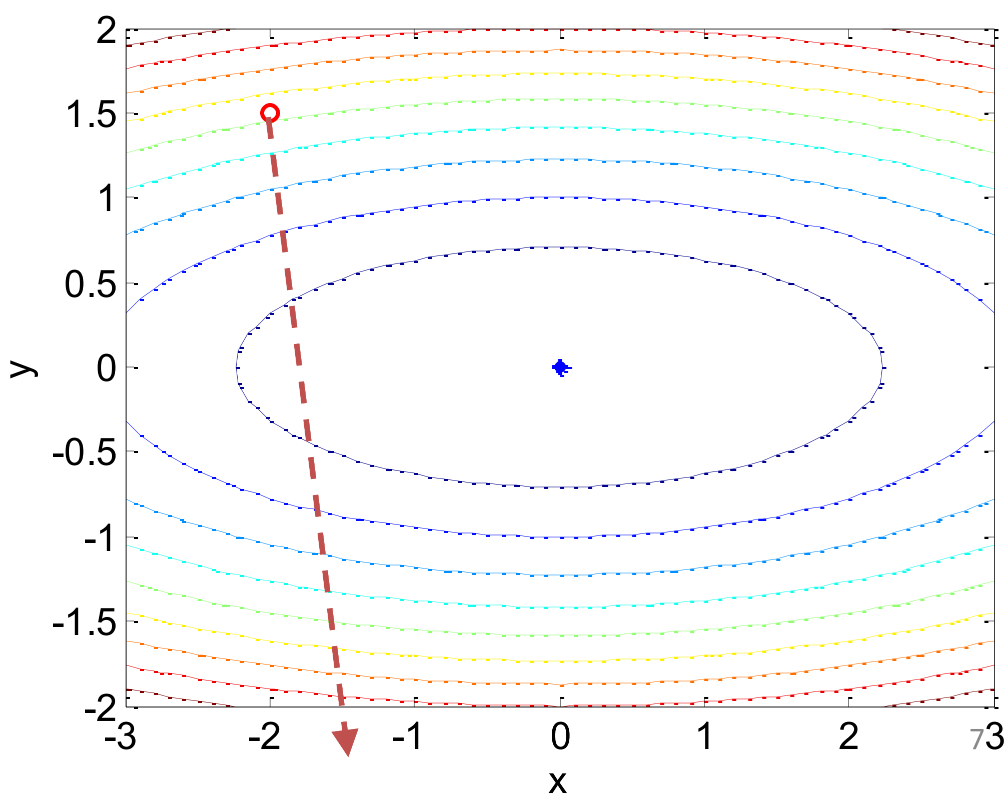
$$\theta_{k+1} = \theta_k - \eta_k \frac{\partial E}{\partial \theta_k}$$

How should we select the step size?

- η_k too small: convergence takes long time
- η_k too large: overshoot minimum

Line minimization:

$$\eta_k = \arg\min \theta_{k+1}$$



Step Size Selection

Recall the general Taylor series expansion:

$f(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots$ Therefore,

$$\frac{\partial E(\theta_{k+1})}{\partial \theta_k} \approx \frac{\partial E(\theta_k)}{\partial \theta_k} + \frac{\partial^2 E(\theta_k)}{\partial \theta_k^2}(\theta_{k+1} - \theta_k)$$

The minimum of $E(\theta_{k+1})$ occurs when its derivative is 0, which means that:

$$\frac{\partial E(\theta_k)}{\partial \theta_k} + \frac{\partial^2 E(\theta_k)}{\partial \theta_k^2}(\theta_{k+1} - \theta_k) = 0$$

$$(\theta_{k+1} - \theta_k) = -\left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2}\right]^{-1} \frac{\partial E(\theta_k)}{\partial \theta_k}$$

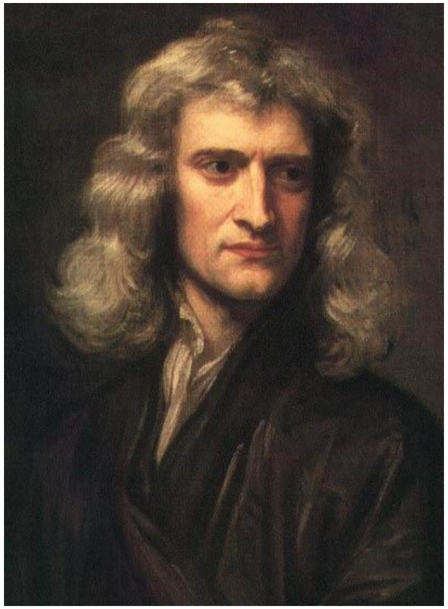
Step Size Selection

Compare the two equations:

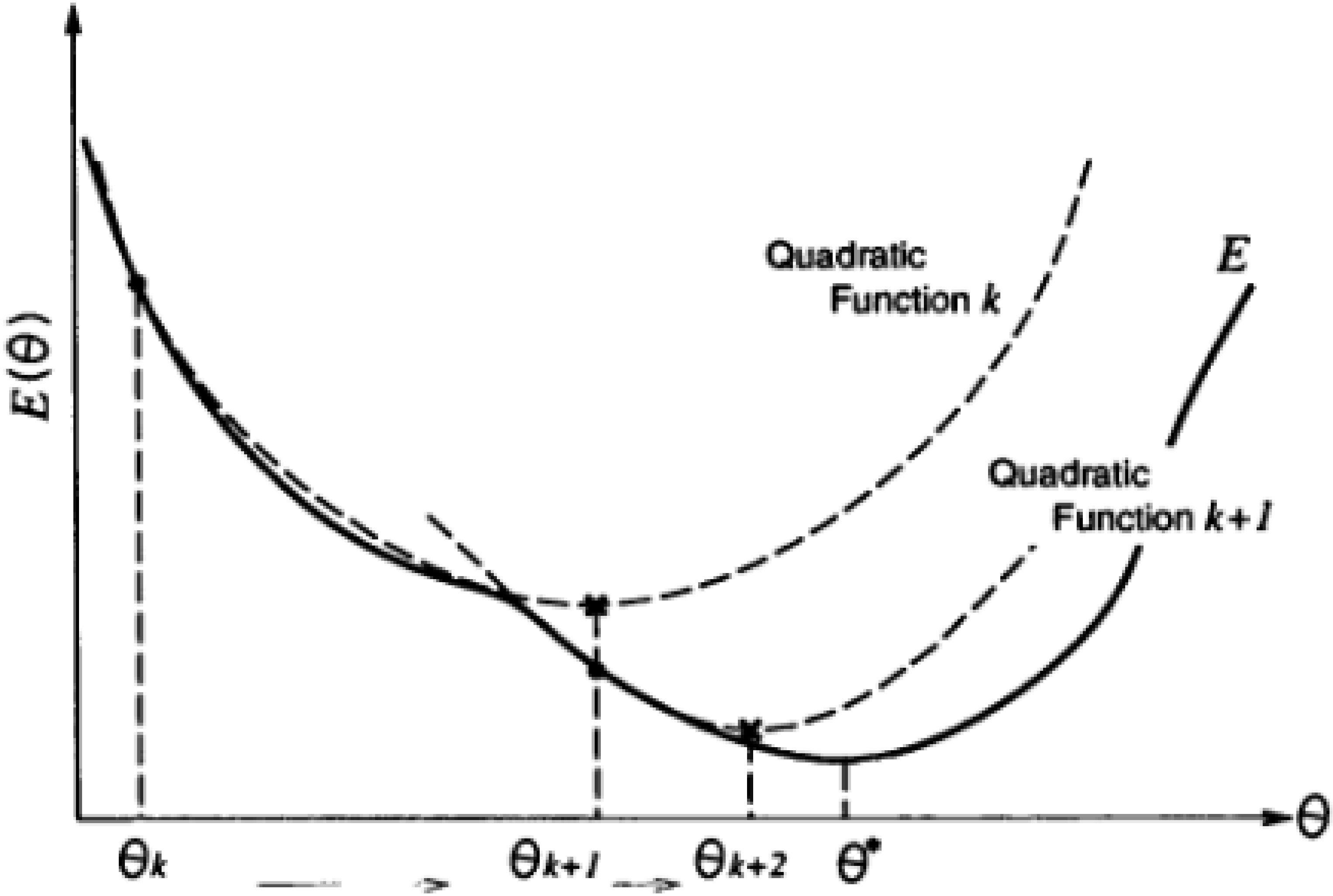
$$\theta_{k+1} = \theta_k - \eta_k \frac{\partial E}{\partial \theta_k}$$

$$(\theta_{k+1} - \theta_k) = -\left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2}\right]^{-1} \frac{\partial E(\theta_k)}{\partial \theta_k}$$

We see that $\eta_k = \left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2}\right]^{-1}$



Newton's method, also called the Newton-Raphson method



Jang, Fig. 6.3 – The Newton-Raphson method approximates $E(\theta)$ as a quadratic function

Step Size Selection

$$\theta_{k+1} = \theta_k - \eta_k \frac{\partial E}{\partial \theta_k}$$

The Newton-Raphson method

$$\eta_k = \left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2} \right]^{-1}$$

The second derivative is called the *Hessian* (Otto Hesse, 1800s)

How easy or difficult is it to calculate the Hessian?

What if the Hessian is not invertible?

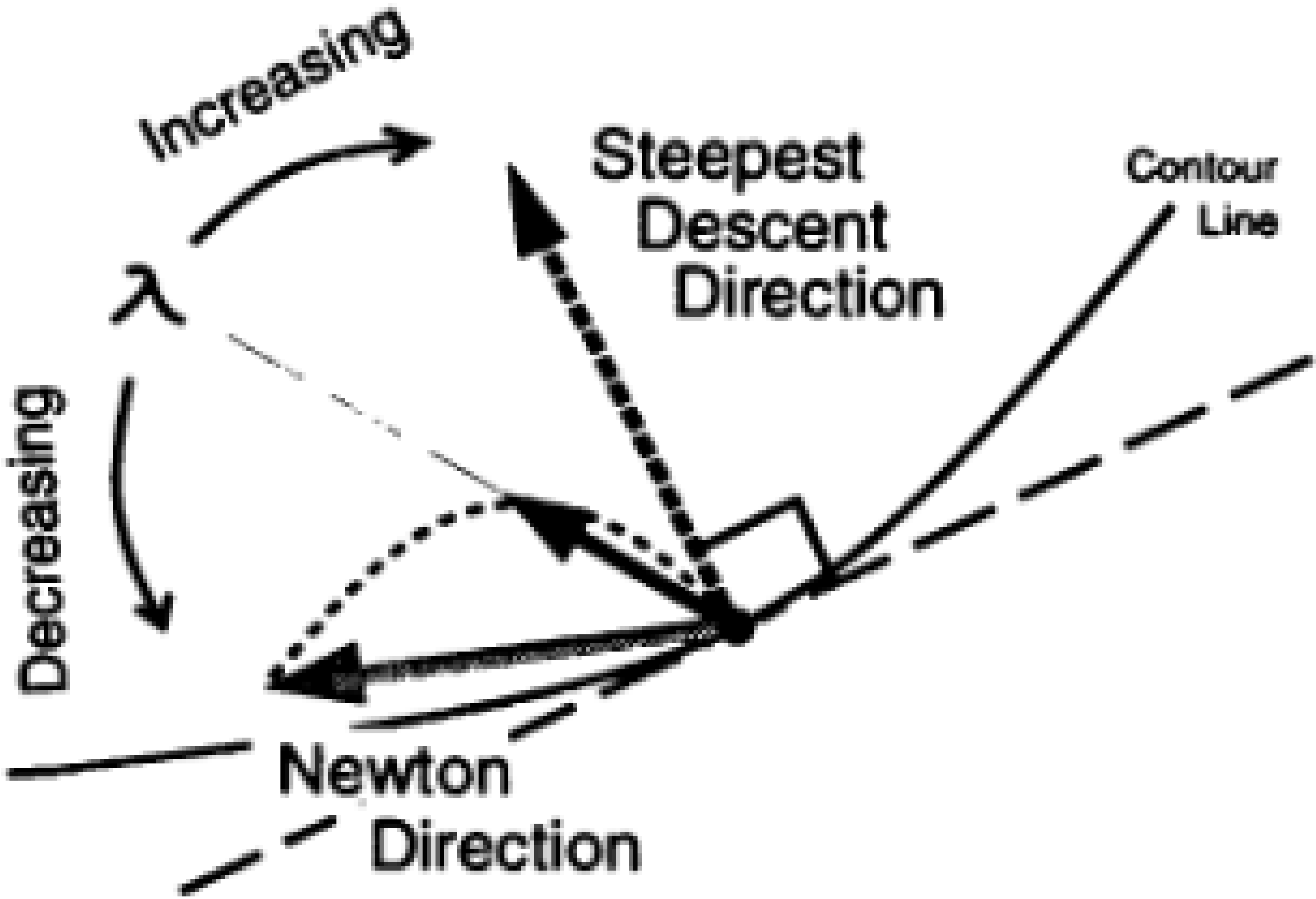
Step Size Selection

The Levenberg–Marquardt algorithm:

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2} + \lambda I \right]^{-1} \frac{\partial E}{\partial \theta_k}$$

λ is a parameter selected to balance between steepest descent ($\lambda = \infty$) and Newton-Raphson ($\lambda = 0$). We can also control the step size with another parameter η:

$$\theta_{k+1} = \theta_k - \eta \left[\frac{\partial^2 E(\theta_k)}{\partial \theta_k^2} + \lambda I \right]^{-1} \frac{\partial E}{\partial \theta_k}$$



Jang, Fig. 6.5 – Illustration of Levenberg-Marquardt gradient descent

Step Size Selection

Trust region methods: These are used in conjuction with the Newton-Raphson method, which approximates *E* as quadratic in θ:

$$E(\theta_k + \Delta\theta_k) \approx E(\theta_k) + \left(\frac{\partial E}{\partial \theta_k}\right)^T \Delta\theta_k + \frac{1}{2}\Delta\theta_k^T \left(\frac{\partial^2 E}{\partial \theta_k^2}\right)\Delta\theta_k$$

If we use Newton-Raphson to minimize *E* with a step size of $\Delta\theta_k$, then we are implicitly assuming that E(θ_k+$\Delta\theta_k$) will be equal to the above expression.

Step Size Selection

$$E(\theta_{k+1}) = \text{actual value after Newton-Raphson step}$$

$$\hat{E}(\theta_{k+1}) = \text{predicted value after Newton-Raphson step}$$

Actually, we expect the actual improvement ν_k to be slightly smaller than the true improvement:

$$\nu_k = \frac{E(\theta_k) - E(\theta_{k+1})}{E(\theta_k) - \hat{E}(\theta_{k+1})}$$

We limit the step size $\Delta\theta_k$ so that $|\Delta\theta_k| < R_k$

Our "trust" in the quadratic approximation is proportional to ν_k

Step Size Selection

ν_k = ratio of actual to expected improvement

R_k = trust region: maximum allowable size of $\Delta\theta_k$

$$R_{k+1} = \begin{cases} \mathrm{R}_k / 2 & \text{if } \nu_k < 0.2 \\ 2R_k & \text{if } \nu_k > 0.8 \\ R_k & \text{otherwise} \end{cases}$$