Lecture 2:

Multivariate Optimization: Line Searches and Steepest Descents

Week of January 17, 2023

University California, Berkeley

Machine Learning Algorithms

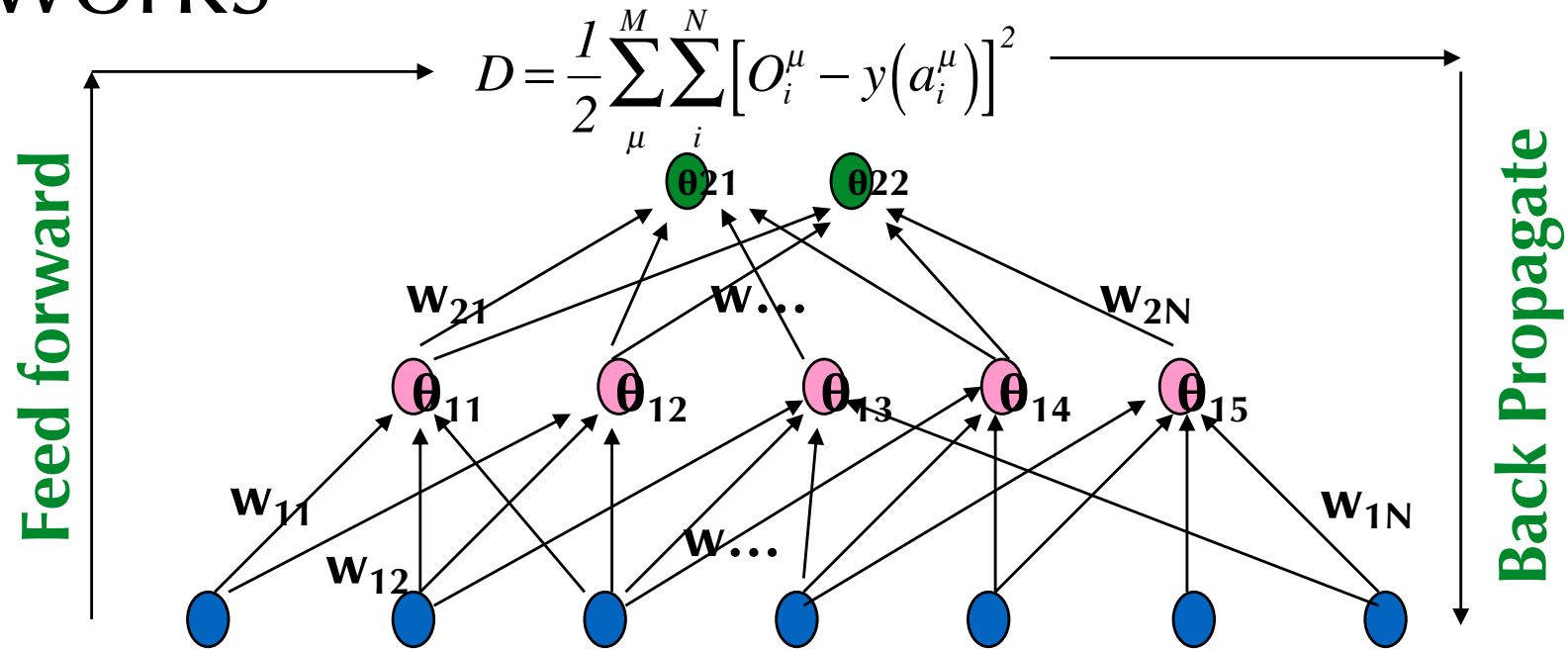
MSSE 277B, 3 Units

Spring 2023

Prof. Teresa Head-Gordon

Departments of Chemistry, Bioengineering, Chemical and Biomolecular Engineering

Supervised Learning with Neural Networks

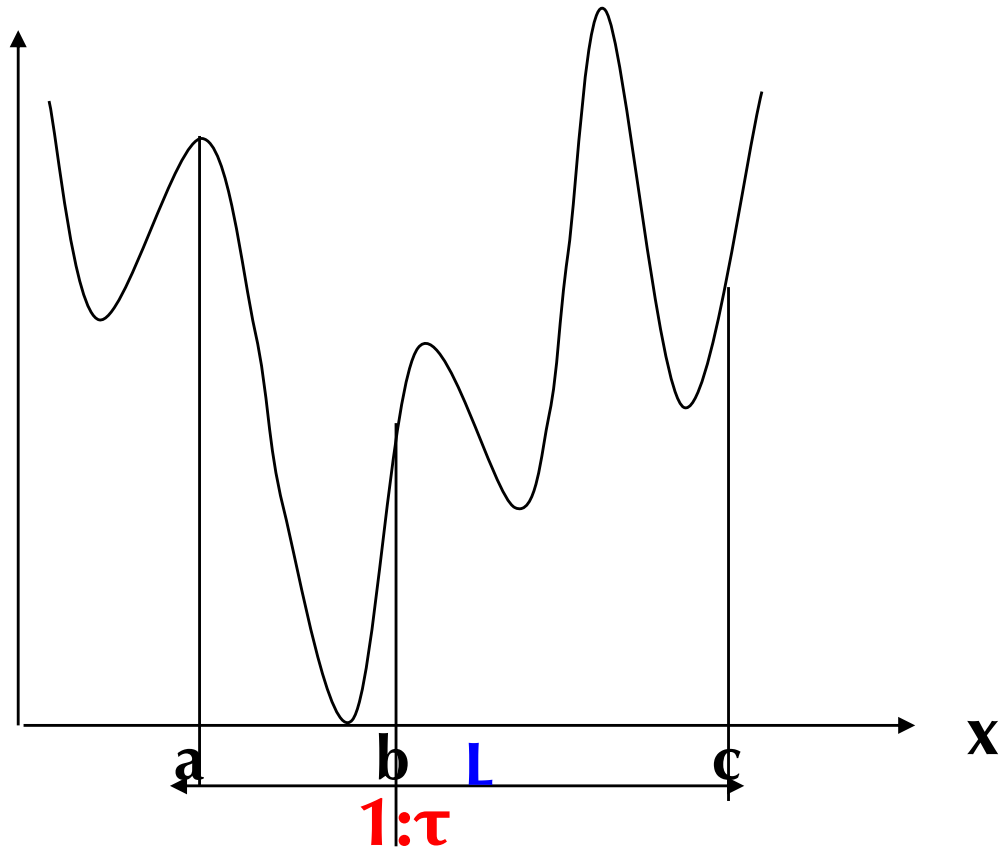


Local optimization methods are used in supervised learning to train network variables (weights, biases) to best reproduce the training data and to ultimately generalize to test or new data to make new predictions.

Adjusting weights to optimize learning via mathematical programming

$$\delta w_{ij} = -\varepsilon\frac{\partial D}{\partial w_{ij}} = \varepsilon\sum_{\mu}\left[O_i^{\mu} - y\left(a_i^{\mu}\right)\right]\frac{dy}{da_i^{\mu}}\frac{\partial a_i^{\mu}}{\partial w_{ij}}$$

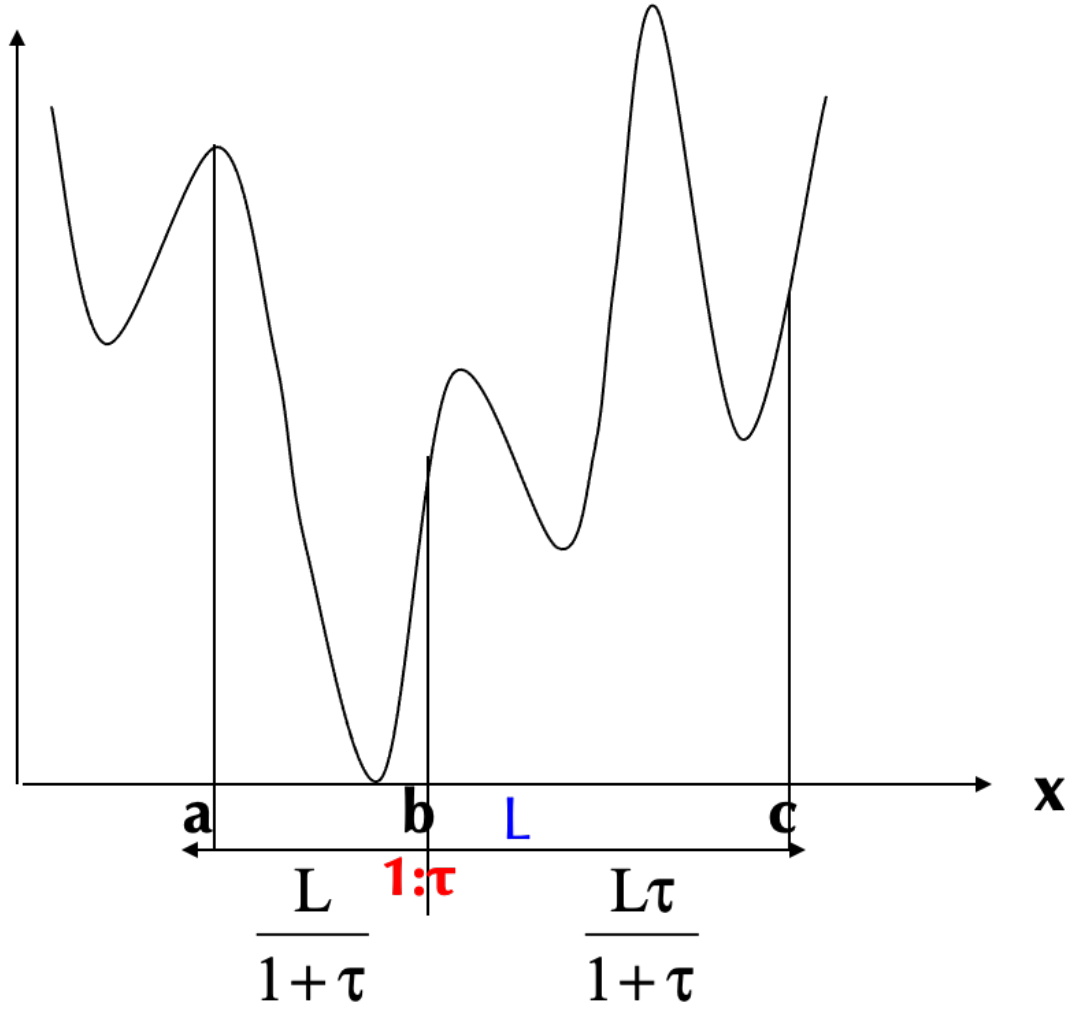
Golden Section Method: 0th Order



The faster the reduction in the size of the bracketed interval, the faster the convergence to the minimum.

Is straight bisection optimal in regards reducing the size of the original, larger interval? You will answer that in homework by comparing it to Golden Section

Golden section picks new points to keep sectioning the interval, unevenly, but optimally, through choice of intervals with ratio 1:τ.



Golden Section Method: 0th Order

Begin by laying down point b so that ratio of interval length [a,b]:[b,c] is 1:τ

$$[a,b]+[b,c]=L$$

$$x_1 L + x_2 L = L$$

$$x_1/x_2 = 1/\tau$$

Solve for x_1 and x_2

$$x_2 L/\tau + x_2 L = L$$

$$x_2(1+1/\tau)=1$$

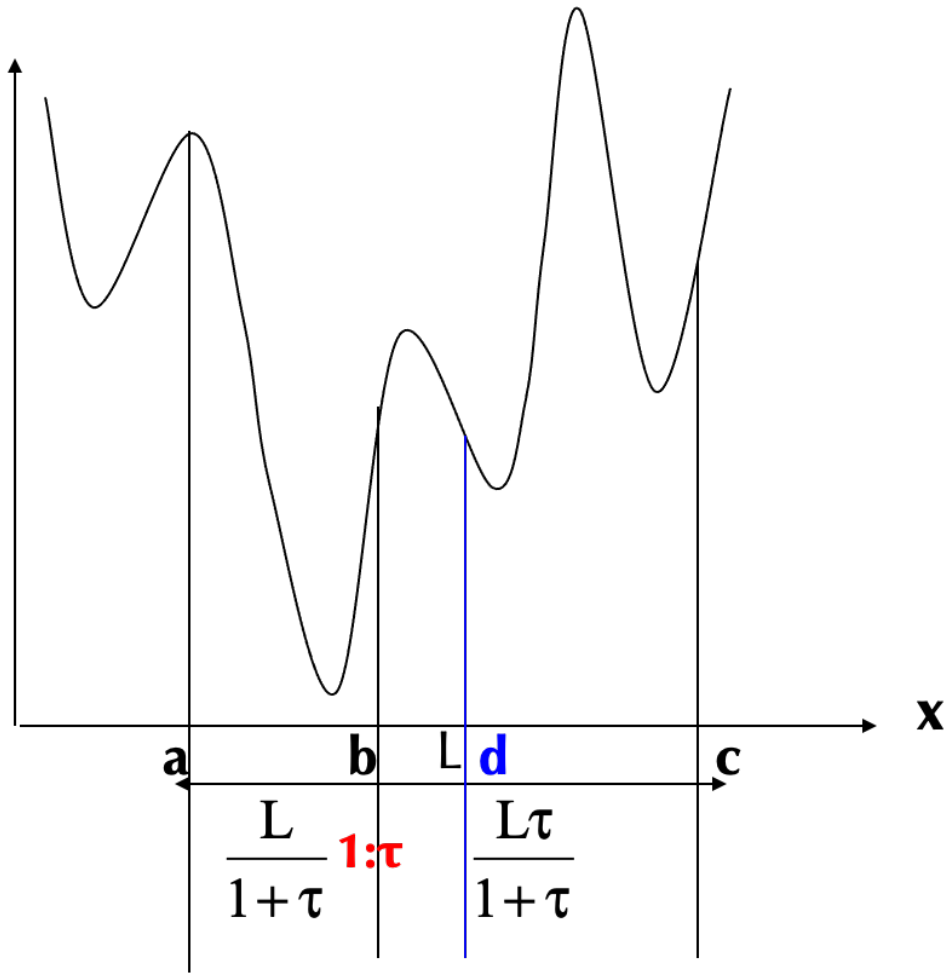
$$x_2 = \tau/(1+\tau)$$

$$x_1 = 1/(1+\tau)$$

$$[a,b] = L/(1+\tau)$$

$$[b,c] = L\tau/(1+\tau)$$

Golden Section Method

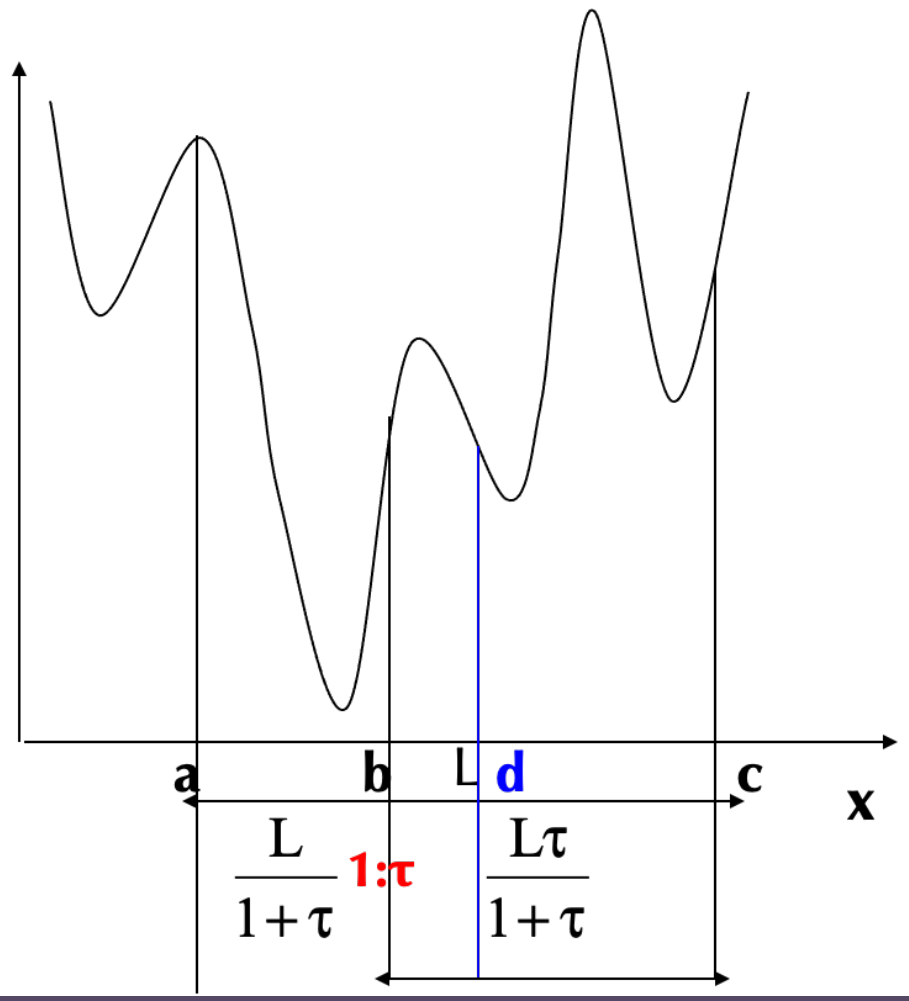


Now where, in the larger interval, do I put my next point, x=d?

It is not known which interval will be the better in bracketing the minima. Is it triplet interval [a,b,d] or [b,d,c]?

Best to define d such that length interval [a,d] is same length as [b,c]

Let's determine the τ that makes this true.



Golden Section Method

We can fix τ by making a recursive requirement in subsequent sectioning so that: [b,d]:[d,c]→1:τ

[b,d] from known information

$$[b,d] = [b,c] - [a,b]$$

$$[b,d] = \frac{L\tau}{1+\tau} - \frac{L}{1+\tau}$$

$$[b,d] = \frac{L(\tau-1)}{1+\tau}$$

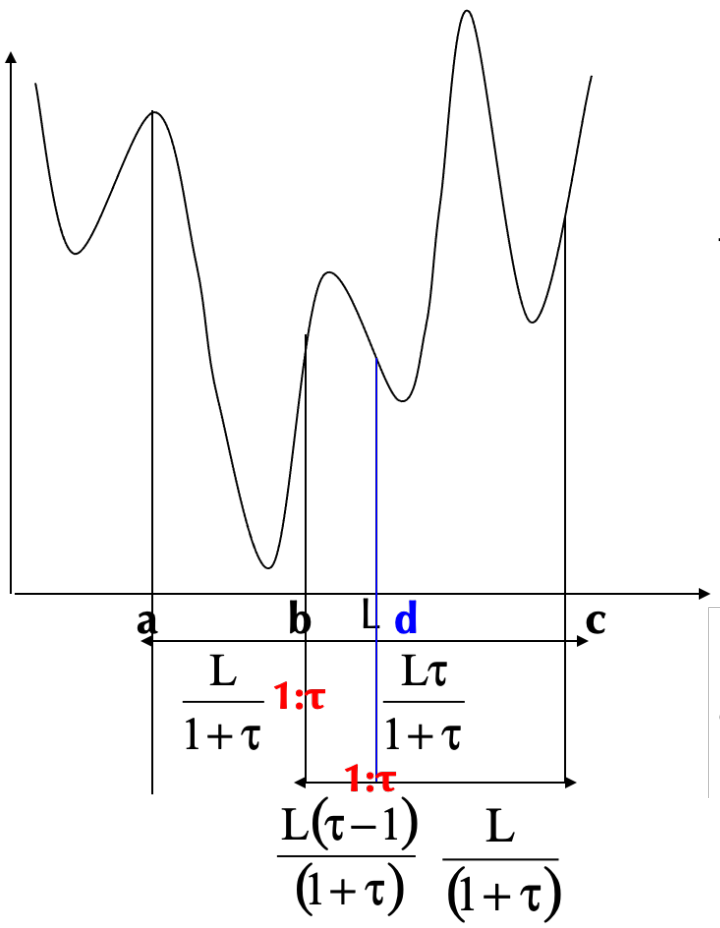
Now determine [d,c]

$$[d,c] = [b,c] - [b,d]$$

$$= \frac{L\tau}{1+\tau} - \frac{L(\tau-1)}{1+\tau}$$

$$= \frac{L}{1+\tau} = [a,b]$$

Golden Section Method



$$[b,d]=\frac{1}{\tau}[d,c]$$

$$\frac{L(\tau-1)}{(1+\tau)}=\frac{L}{\tau(1+\tau)}$$

$$\tau^2-\tau-1=0$$

x

$$\frac{-b\pm\sqrt{b^2-4ac}}{2a}\Rightarrow$$

$$\frac{1\pm\sqrt{5}}{2}\sim 1.618$$

We can fix τ by making a recursive requirement in subsequent sectioning so that:

[b,d]:[d,c]→1:τ (or [b,d]:[a,b]→1:τ)

τ=1.618…is the best way to lay down the next interval because the interval is always guaranteed to be reduced by [d,c]/[a,c]=

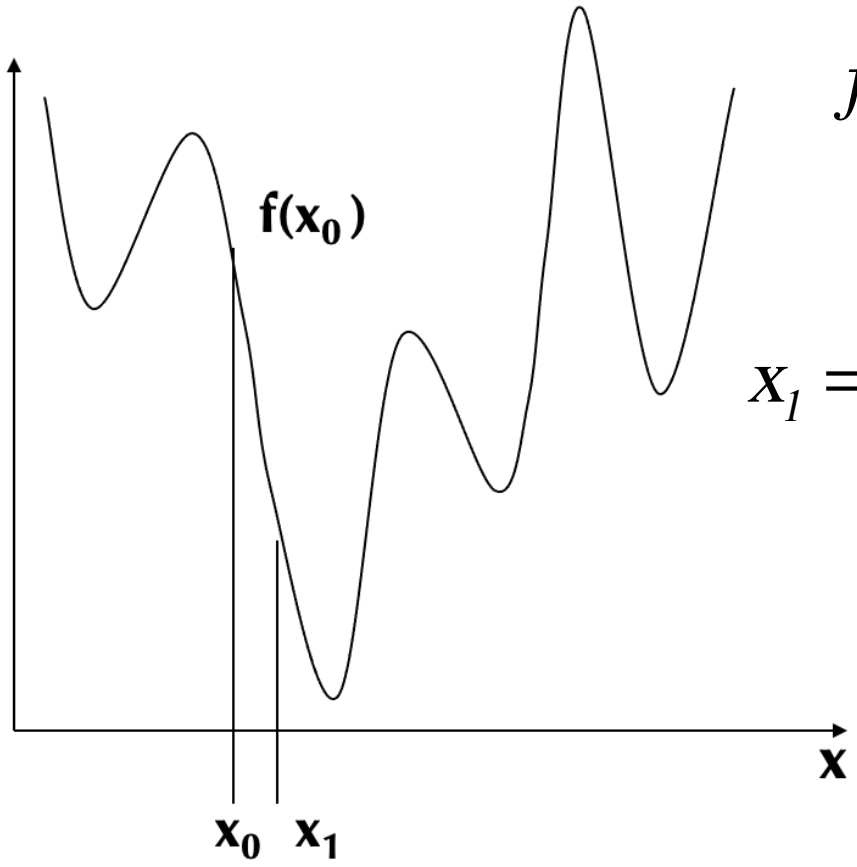
$$\frac{L}{L(1+\tau)}=\frac{1}{(1+\tau)}=0.382$$

Step 1: new interval is 0.618

Step 2: new interval is 0.382

Step 3: new interval is 0.236

Steepest Descent: 1st Order Method



$$f(x_0); \left.\frac{df}{dx}\right|_{x_0}$$

$$x_1 = x_0 - \lambda \left.\frac{df}{dx}\right|_{x_0}$$

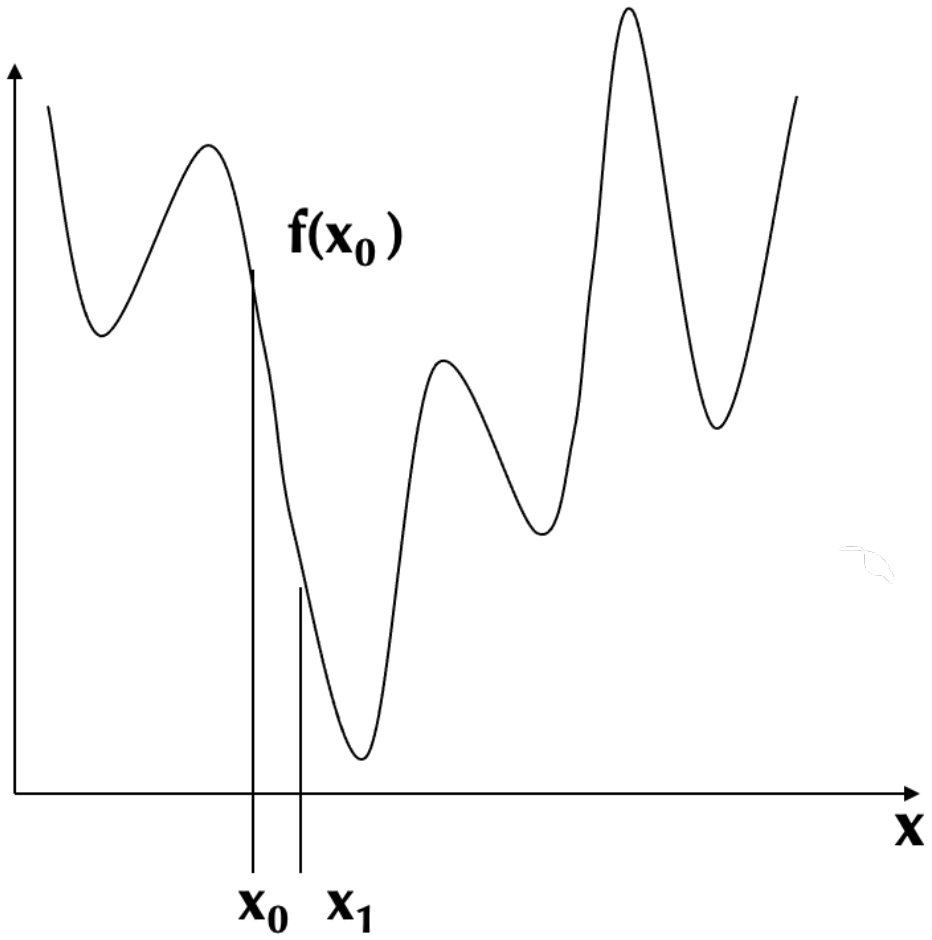
(1) Initialize variable λ=0.001

(2) Start at point x_0 evaluate function and first derivative (gradient) and define step variable λ=0.001

(3) Take a step along direction of gradient, g

(4a) Is $f(x_1) < f(x_0)$?
Yes: $\lambda = 1.2\ \lambda$
Since it was a good step (it lowered the function value), then increase λ and teturn to step 2 with new point $x_0 = x_1$

(4b) Is $f(x_1) < f(x_0)$?
No: $\lambda = 0.5\ \lambda$
Since it was a bad step (it raised the function value), then decrease λ, return to step 2 with new λ and old point x0



Steepest Descent: 1st Order Method

(5) Are we converged?

$$\left|x_{new} - x_0\right| < x_{tol}$$

$$\left|f\left(x_{new}\right) - f\left(x_0\right)\right| < f_{tol}$$

Quit when function difference or derivative < tolerance (I.e. close to zero)

Else

return

Multivariate Functions: Quadratic

Define multivariate quadratic function:

$$f(\vec{x}) = \frac{1}{2}\vec{x}^T \underline{\underline{H}}\vec{x} - \vec{b}\vec{x} + c$$

and at a minimum

$$-\vec{\nabla} f(\vec{x}_{min}) = -\underline{\underline{H}}\vec{x}_{min} + \vec{b} = 0$$

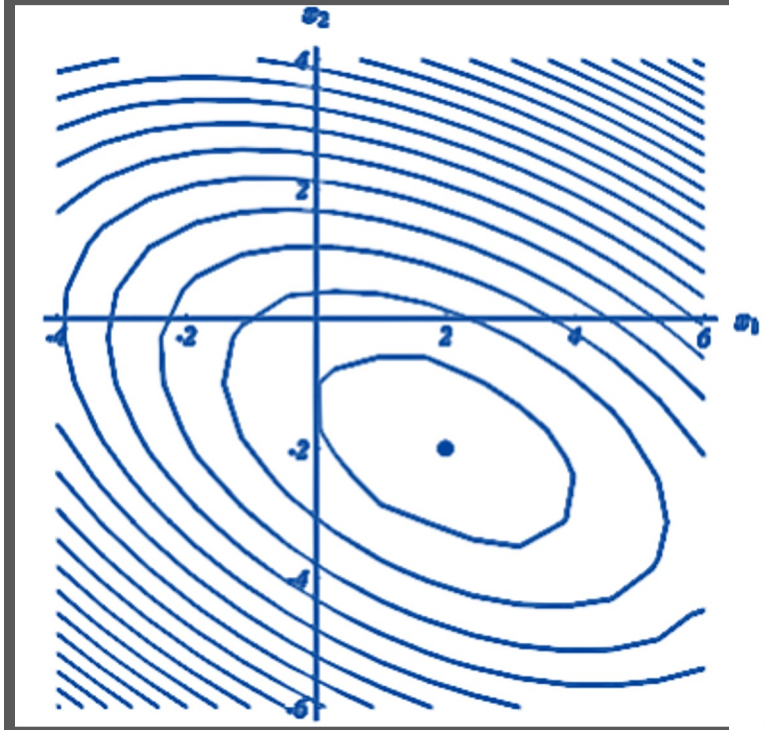
We are solving a set of linear equations,
where H (the Hessian) and b are known, but x_{min} is unknown.

The "residual" r, measures how far from b we are

$$\vec{r}_i = -\nabla f(\vec{x}_i) = \vec{b} - \underline{\underline{H}}\vec{x}_i$$

The error measures how far we are from x_{min}, and is related to residual as

$$\vec{e}_i = \vec{x}_i - \vec{x}_{min} \longrightarrow \vec{r}_i = -\underline{\underline{H}}\vec{e}_i$$



Multivariate Optimization: 1st Order

For multivariate optimization we want to minimize a function of many variables, $f(\vec{x})$, where $\vec{x}$ is an n-dimensional vector. In general, multivariate minimization using a 1st order method follows the general prescription:

(1) Define a start point vector

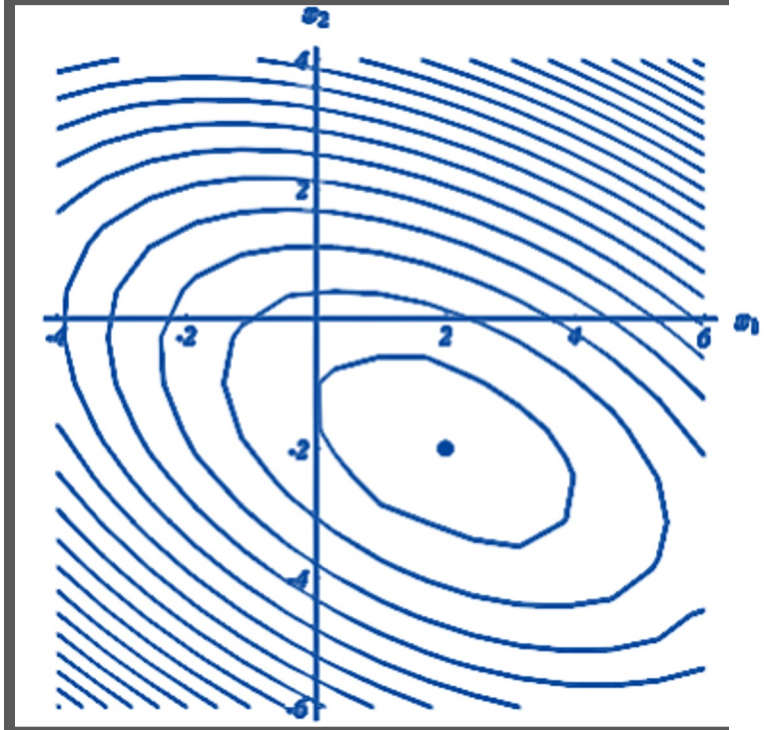
(2) Choose a search direction $d_i(\vec{x})$

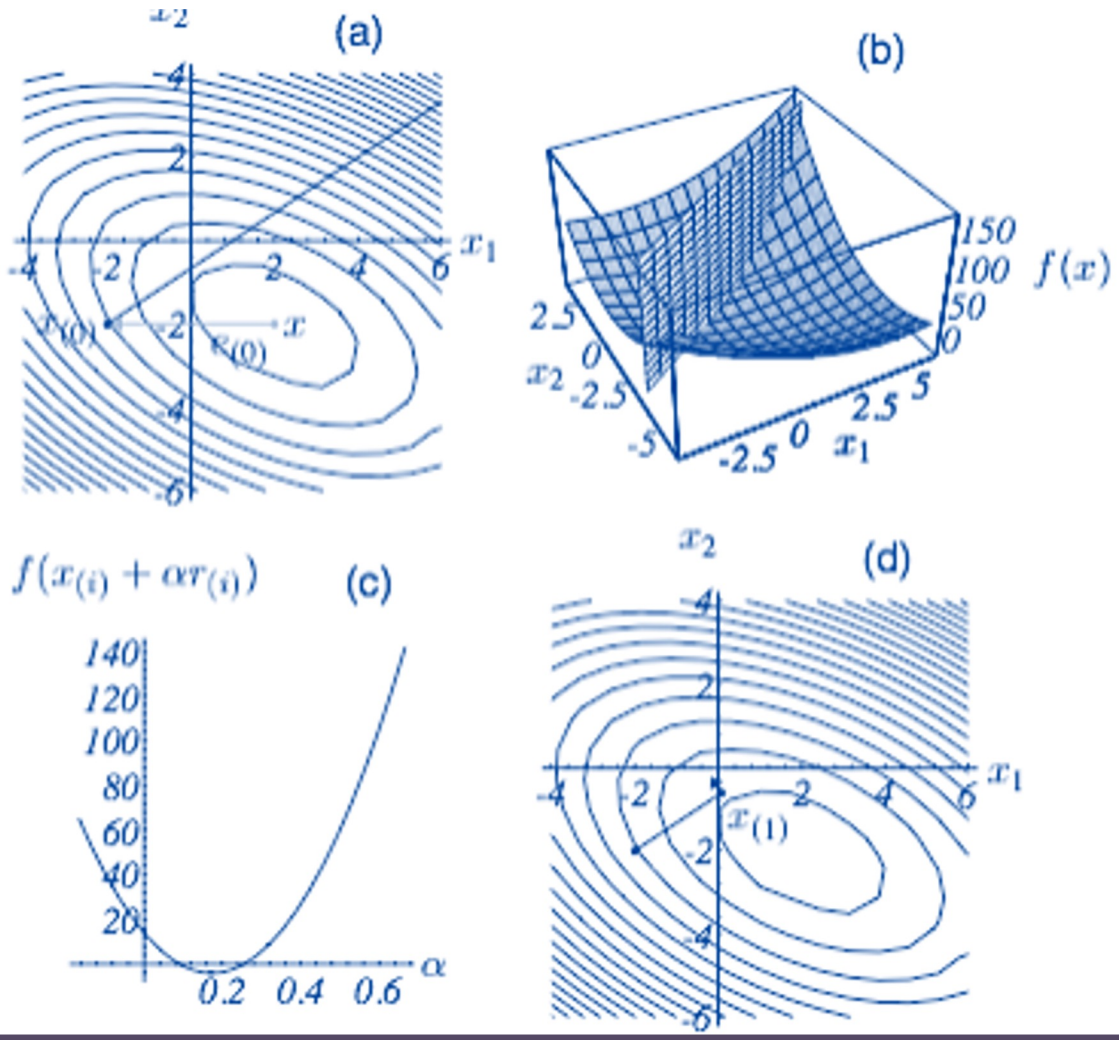
(3) Use a line search method that determines a minimum along that direction (for example, Golden Section).

$$\vec{x}_{i+1} = \vec{x}_i + \lambda_i d_i(\vec{x})$$

(4) Define if converged in multi-dimensional space

Yes: exit If no: return to (2)





Steepest Descents with line searched

A line search is a procedure that chooses λ to minimize f along solid line shown in Figure (a).

Figure (b) illustrates that we are restricted to choosing a point on intersection of vertical plane and quadratic.

Figure (c) is the parabola defined by intersection of these surfaces.

Figure (d) shows gradient vectors (solid) at various points along search line. The slope at any point along parabola is equal to magnitude of the projection of the gradient onto the line (dashed)

Steepest Descents

These projections represent the rate of increase of $f(x_1,x_2)$ as one traverses the search line. $f(x_1,x_2)$ is minimized where the projection is zero, i.e. where gradient is orthogonal to the search line.

$$\vec{x}_1 = \vec{x}_0 + \lambda_0 \vec{r}_0$$

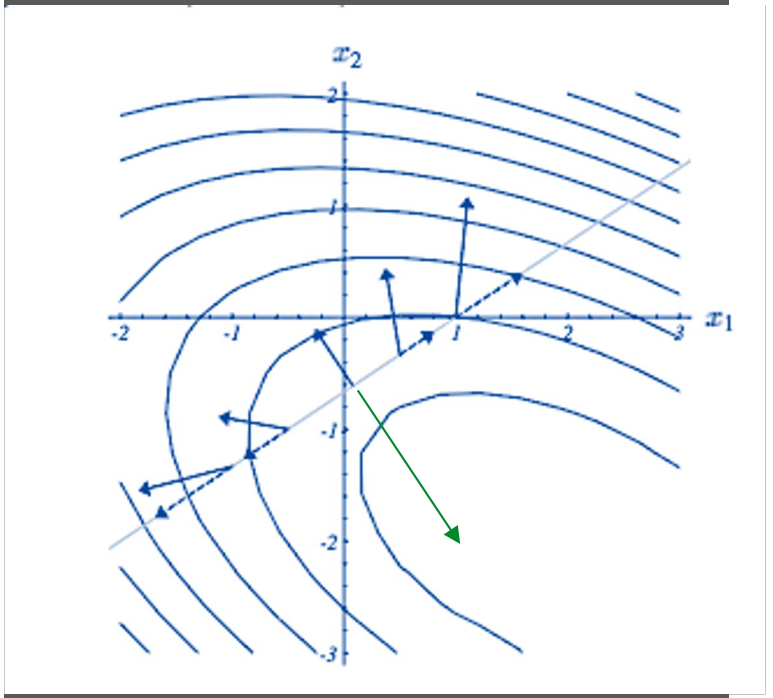
So λ that minimizes f along direction x(1)

$$\frac{df(\vec{x}_1)}{d\lambda} = 0$$

is when following is true:

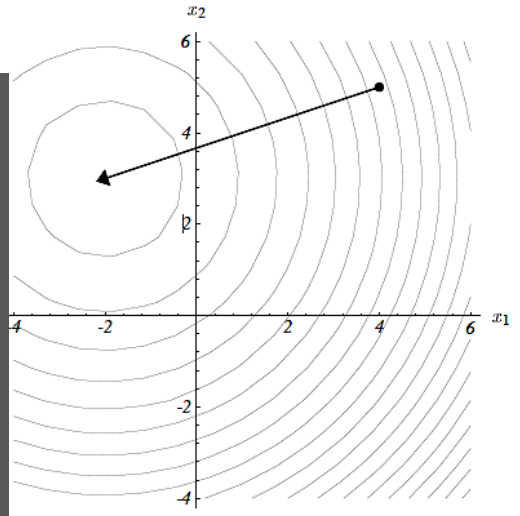
$$\frac{df(\vec{x}_1)}{d\vec{x}_1}\frac{d\vec{x}_1}{d\lambda} = 0$$

$$\nabla f(\vec{x}_1)^T \vec{r}_0 = 0 \longrightarrow \vec{r}_1^{\,T} \vec{r}_0 = 0$$



Hence next search direction is orthogonal to previous direction

Steepest Descents



Steepest descents for a homogeneous quadratic function converges in one step; hence, no matter what point we start at, the residual must point to the center of the sphere.(Top figure)

But for inhomogeneous quadratics (or arbitrary functions) the SD will take many steps, often traversing the same previous directions and thus destroying previous line minimizers!

If we are lucky enough to start at an $\vec{x}_0$ as in Figure (a), the large gradient will heavily weight the step size.

If we start at an $\vec{x}_0$ as in Figure (b), then gradients are small and step size is short. And steepest descent directions bounce back and forth from edge of trough, but not moving well along its length.