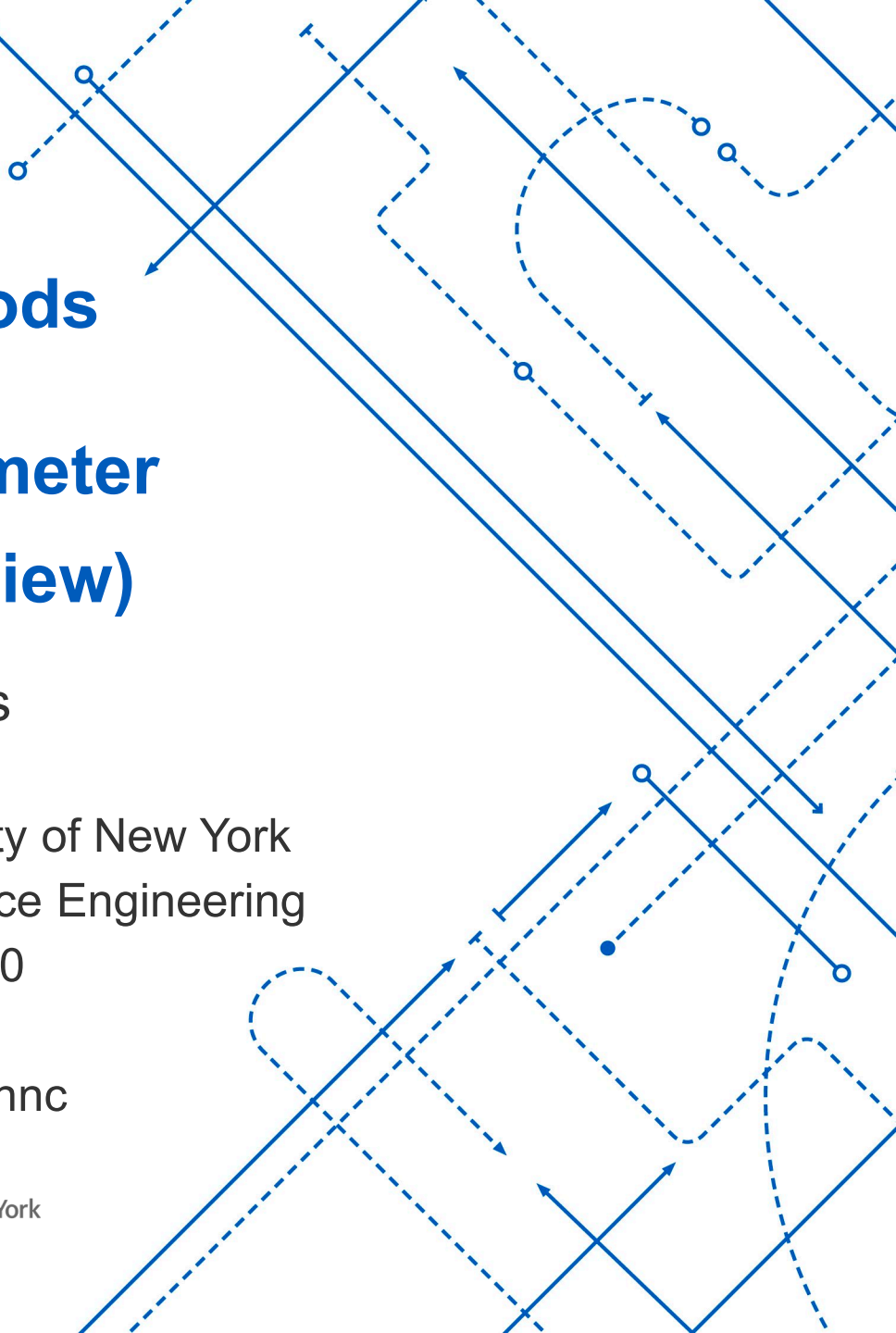


Optimal Estimation Methods

(Lecture 2 – Parameter Optimization Review)

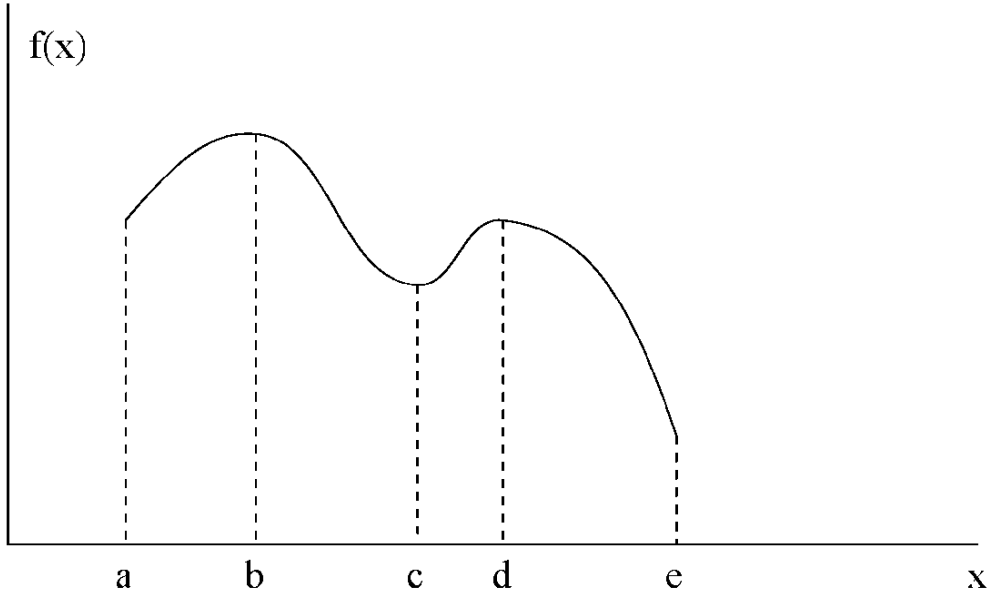
Dr. John L. Crassidis

University at Buffalo – State University of New York
Department of Mechanical & Aerospace Engineering
Amherst, NY 14260-4400
johnc@buffalo.edu
http://www.buffalo.edu/~johnc

University at Buffalo The State University of New York

Unconstrained Extrema (i)

- Extrema, whether they are local or global, can occur in three places
  1. At the boundary of the domain
  2. At a point without a derivative, or
  3. At a point where the function is exactly zero
- Consider the following function in domain [a, e]



Points b and d are local maxima and b is a global maximum, whereas a, c, and e are local minima. Note e is a global minimum (global because of the domain space)

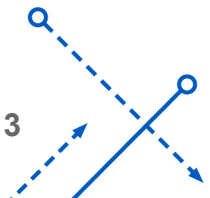
Unconstrained Extrema (ii)

- Scalar minimization (optimality conditions)
  - Suppose that $\vartheta(x)$ is a continuously differentiable function of the scalar variable x, and that, when $x = x^*$

$$\frac{d\vartheta}{dx} = 0 \quad \text{and} \quad \frac{d^2\vartheta}{dx^2} > 0$$

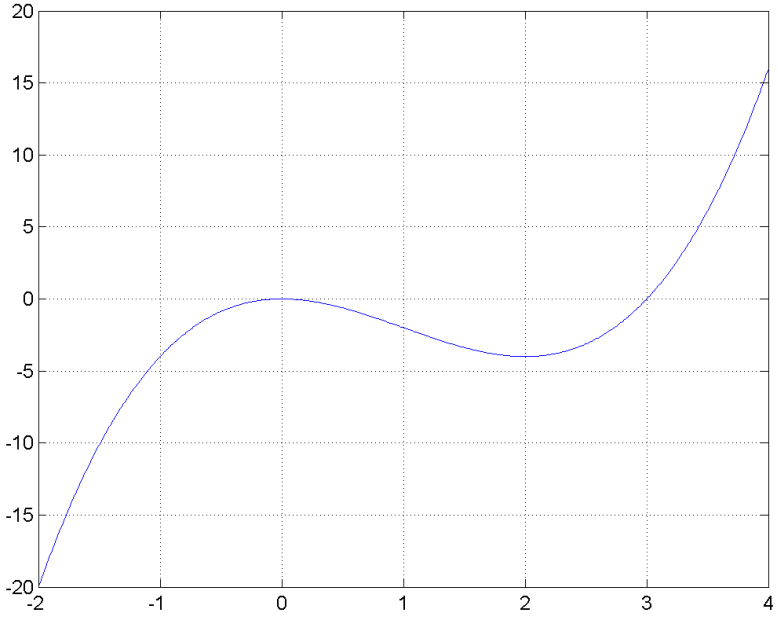
(local minimum condition)

  - The function $\vartheta(x)$ is then said to have a local minimum at x^*
    - This implies that $\vartheta(x)$ is the smallest value of ϑ in some region near x^*
    - It may also be true that $\vartheta(x^*) \leq \vartheta(x)$ for all x but the conditions do not guarantee this
  - If the conditions hold at $x = x^*$ and if $\vartheta(x^*) \leq \vartheta(x)$ for all x then x^* is said to be the global minimum
  - In practice it is usually hard to establish that x^* is a global minimum, and so we shall chiefly be concerned with methods of finding local minima



Scalar Example

- Suppose $\vartheta(x) = x^3 - 3x^2$
  - The derivative is given by $3x^2 - 6x$
    - Two stationary points given by $x_1 = 0$ and $x_2 = 2$
  - The second derivative is given by $6x - 6$
    - The first point gives a local maximum and the second point gives a local minimum (no global maximum or minimum here) (unless a domain is defined)



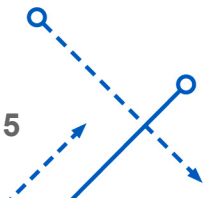
Multidimensional Case

- Optimality Conditions
  - Suppose that $\vartheta(\mathbf{x})$ is a continuously differentiable function of the vector variable $\mathbf{x}$, and that, when $\mathbf{x} = \mathbf{x}^*$

$$\nabla_{\mathbf{x}}\vartheta \equiv \frac{\partial \vartheta}{\partial \mathbf{x}} = \mathbf{0} \quad \text{and} \quad \nabla^2_{\mathbf{x}}\vartheta \equiv \left.\frac{\partial^2 \vartheta}{\partial \mathbf{x}\,\partial \mathbf{x}^T}\right|_{\mathbf{x}^*} \text{ must be positive definite}$$

for maximum

  - Now the Jacobian and Hessian are used
  - The function $\vartheta(\mathbf{x})$ is then said to have a local minimum at $\mathbf{x}^*$
  - If the Hessian matrix is negative definite, then the point is a local maximum (eigenvalues negative)
  - If the matrix is indefinite, then a saddle point exists
    - Corresponds to a relative minimum or maximum with respect to the individual components of $\mathbf{x}^*$
  - A global minimum is much more difficult to establish, though



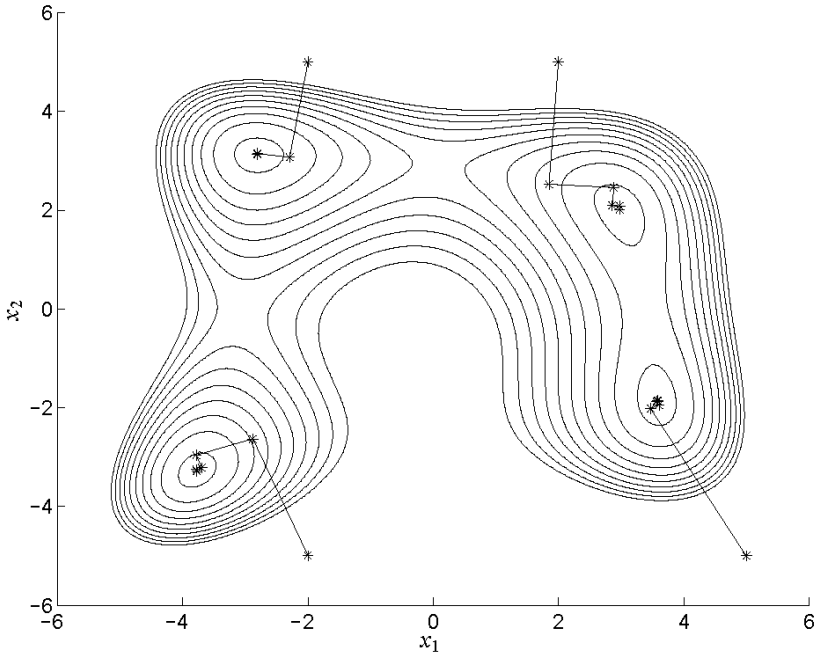
Example I

- Himmelblau's Function

$$\vartheta(\mathbf{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

  - Four stationary points which provide local minimums

- $\mathbf{x}_1^* = \begin{bmatrix} 3 & 2 \end{bmatrix}^T$, with $\vartheta(\mathbf{x}_1^*) = 0$
- $\mathbf{x}_2^* = \begin{bmatrix} -3.7792 & -3.2831 \end{bmatrix}^T$, with $\vartheta(\mathbf{x}_2^*) = 0.0054$
- $\mathbf{x}_3^* = \begin{bmatrix} -2.8051 & 3.1313 \end{bmatrix}^T$, with $\vartheta(\mathbf{x}_3^*) = 0.0085$
- $\mathbf{x}_4^* = \begin{bmatrix} 3.5843 & -1.8483 \end{bmatrix}^T$, with $\vartheta(\mathbf{x}_4^*) = 0.0011$



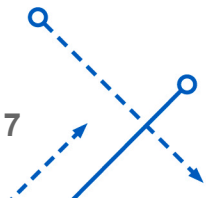
  - A numerical technique, such as the method of gradients, can converge to any one of these four points from various starting guesses

Example II (i)

- Consider the following function

$$\vartheta(\mathbf{x}) = x_1^3 + x_2^3 + 2x_1^2 + 4x_2^2 + 6$$

- In-class assignment
  - Derive the necessary conditions from the Jacobian
  - Determine the stationary points
    - Note there are four of them total
  - Derive the Hessian matrix
  - Using the Hessian to describe the nature of the stationary points
    - Are they local minimums, local maximums, or saddle points?



Example II (ii)

- Necessary Conditions

$$\frac{\partial \vartheta}{\partial x_1} = x_1(3x_1 + 4) = 0$$
$$\frac{\partial \vartheta}{\partial x_2} = x_2(3x_2 + 8) = 0$$

  - Provide the following stationary points

$$\mathbf{x}_1^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T, \quad \mathbf{x}_2^* = \begin{bmatrix} 0 & -\frac{8}{3} \end{bmatrix}^T$$
$$\mathbf{x}_3^* = \begin{bmatrix} -\frac{4}{3} & 0 \end{bmatrix}^T, \quad \mathbf{x}_4^* = \begin{bmatrix} -\frac{4}{3} & -\frac{8}{3} \end{bmatrix}^T$$

- Hessian Matrix

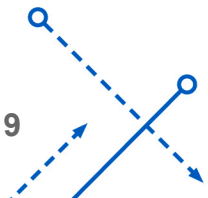
$$\nabla_{\mathbf{x}}^2 \vartheta = \begin{bmatrix} 6x_1 + 4 & 0 \\ 0 & 6x_2 + 8 \end{bmatrix}$$

Example II (iii)

- Nature of the Hessian and Values for the Loss Function

Point $\mathbf{x}_i^*$	Nature of $\nabla_{\mathbf{x}}^2 \vartheta\|_{\mathbf{x}_i^*}$	Nature of $\mathbf{x}_i^*$	$\vartheta(\mathbf{x}_i^*)$
$\mathbf{x}_1^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$	Positive Definite	Relative Minimum	6
$\mathbf{x}_2^* = \begin{bmatrix} 0 & -\frac{8}{3} \end{bmatrix}^T$	Indefinite	Saddle Point	418/27
$\mathbf{x}_3^* = \begin{bmatrix} -\frac{4}{3} & 0 \end{bmatrix}^T$	Indefinite	Saddle Point	194/27
$\mathbf{x}_4^* = \begin{bmatrix} -\frac{4}{3} & -\frac{8}{3} \end{bmatrix}^T$	Negative Definite	Relative Maximum	50/3

- The first point gives a local minimum
  - Is it the global minimum? (look at the value of the loss function)
- The next two points are saddle points
- The last point gives a local maximum



Equality Constrained Extrema (i)

- One often encounters problems that must extremize

$$\vartheta \equiv \vartheta(\mathbf{x})$$

subject to the following set of $m \times 1$ equality constraints

$$\boldsymbol{\psi} \equiv \boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}, \quad \text{with} \quad m < n$$

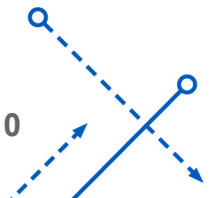
- Consider the case of $n = 2$ and $m = 1$

Suppose (x_1^*, x_2^*) locally minimizes the loss function while satisfying the constraint. If this is true, then arbitrary admissible differential variations $(\delta x_1, \delta x_2)$ in the differential neighborhood of (x_1^*, x_2^*) in the sense $(x_1, x_2) = (x_1^* + \delta x_1, x_2^* + \delta x_2)$ result in a stationary value of ϑ

$$\delta\vartheta = \frac{\partial\vartheta}{\partial x_1}\delta x_1 + \frac{\partial\vartheta}{\partial x_2}\delta x_2 = 0 \qquad (1)$$

Since we restrict attention to neighboring points that satisfy the constraint, we also require the first variation of the constraint to vanish as a condition on the admissibility of $(\delta x_1, \delta x_2)$ as

$$\delta\psi = \frac{\partial\psi}{\partial x_1}\delta x_1 + \frac{\partial\psi}{\partial x_2}\delta x_2 = 0 \qquad (2)$$



Equality Constrained Extrema (ii)

For notational convenience, we suppress the truth that all partials in Eqs. (1) and (2) are evaluated at (x_1^*, x_2^*)

- Since Eq. (2) constrains the admissible variations, we can solve for either variable and eliminate the constraint equation
- The two solutions of the constraint equations are obviously

$$\delta x_1 = -\left(\frac{\dfrac{\partial \psi}{\partial x_2}}{\dfrac{\partial \psi}{\partial x_1}}\right)\delta x_2 \quad \text{and} \quad \delta x_2 = -\left(\frac{\dfrac{\partial \psi}{\partial x_1}}{\dfrac{\partial \psi}{\partial x_2}}\right)\delta x_1$$

- Substitution of the “differential eliminations” into the differential of the loss function allows us to locally constrain the variations of ϑ and reduce the dimensionality either of two ways

Equality Constrained Extrema (iii)

- First way

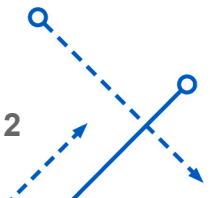
$$\delta\vartheta = \left[\frac{\partial\vartheta}{\partial x_2} - \left(\frac{\frac{\partial\vartheta}{\partial x_1}}{\frac{\partial\psi}{\partial x_1}}\right)\frac{\partial\psi}{\partial x_2}\right]\delta x_2 = 0$$

- Second way

must be zero

$$\delta\vartheta = \left[\frac{\partial\vartheta}{\partial x_1} - \left(\frac{\frac{\partial\vartheta}{\partial x_2}}{\frac{\partial\psi}{\partial x_2}}\right)\frac{\partial\psi}{\partial x_1}\right]\delta x_1 = 0$$

- It is evident that either of the equations can be used to argue that the local variations are arbitrary and the coefficient within the brackets must vanish as a necessary condition for a local minimum



Equality Constrained Extrema (iv)

- The first form of the necessary conditions is given by

$$\frac{\partial \vartheta}{\partial x_1} - \left(\frac{\dfrac{\partial \vartheta}{\partial x_2}}{\dfrac{\partial \psi}{\partial x_2}} \right) \frac{\partial \psi}{\partial x_1} = 0 \qquad (3)$$

$$\psi(x_1, x_2) = 0$$

- The second form of the necessary conditions is given by

$$\frac{\partial \vartheta}{\partial x_2} - \left(\frac{\dfrac{\partial \vartheta}{\partial x_1}}{\dfrac{\partial \psi}{\partial x_1}} \right) \frac{\partial \psi}{\partial x_2} = 0 \qquad (4)$$

$$\psi(x_1, x_2) = 0$$

Equality Constrained Extrema (v)

- Lagrange noticed a pattern in the above and decided to "automate" all possible differential eliminations by linearly combining Eqs. (1) and (2) with an unspecified scalar Lagrange multiplier λ as

$$\delta\vartheta + \lambda\,\delta\psi = \left[\frac{\partial\vartheta}{\partial x_1} + \lambda\frac{\partial\psi}{\partial x_1}\right]\delta x_1 + \left[\frac{\partial\vartheta}{\partial x_2} + \lambda\frac{\partial\psi}{\partial x_2}\right]\delta x_2 = 0$$

- While it "isn't legal" to set the two brackets to zero using the argument that (δx_1, δx_2) are independent, we can set either one of the brackets to zero to determine λ

Notice that setting the first bracket to zero and substituting the resulting equation for $\lambda = -\left(\frac{\partial\vartheta}{\partial x_1}\right) / \left(\frac{\partial\psi}{\partial x_1}\right)$ into the second bracket renders the second bracket equal to Eq. (4), whereas setting the second bracket to zero, solving for λ, and substituting renders the first bracket equal to Eq. (3)

Equality Constrained Extrema (vi)

- Thus, the following necessary generalized Lagrange form of the necessary conditions captures all possible differential constraint eliminations (only two in this case)

$$\frac{\partial \vartheta}{\partial x_1} + \lambda \frac{\partial \psi}{\partial x_1} = 0$$
$$\frac{\partial \vartheta}{\partial x_2} + \lambda \frac{\partial \psi}{\partial x_2} = 0$$
$$\psi(x_1, x_2) = 0$$

Lagrange multipliers

- It is apparent these equations are the gradient of the augmented function $\phi \equiv \vartheta + \lambda\psi$ with respect to (x_1, x_2, λ) and thus the Lagrange multiplier rule is validated

Equality Constrained Extrema (vii)

- The necessary conditions for a constrained minimum of a loss function subject to an equality constraint have the form of an unconstrained minimum of the augmented function

$$\frac{\partial \phi}{\partial x_1} = \frac{\partial \vartheta}{\partial x_1} + \lambda \frac{\partial \psi}{\partial x_1} = 0$$
$$\frac{\partial \phi}{\partial x_2} = \frac{\partial \vartheta}{\partial x_2} + \lambda \frac{\partial \psi}{\partial x_2} = 0$$
$$\psi(x_1, x_2) = 0$$

These provide three equations; all points (x_1^*, x_2^*, λ) satisfying these equations are *constrained stationary points*

Equality Constrained Extrema (viii)

- Expanding this concept to the general case results in the necessary conditions for a stationary point, which are applied by the unconstrained necessary condition to the following augmented function

$$\phi \equiv \phi(\mathbf{x}, \boldsymbol{\lambda}) = \vartheta(\mathbf{x}) + \boldsymbol{\lambda}^T \boldsymbol{\psi}(\mathbf{x})$$

  where $\boldsymbol{\lambda}$ is an $m \times 1$ vector of Lagrange multipliers
- The necessary conditions are now given by

$$\nabla_{\mathbf{x}} \phi \equiv \frac{\partial \phi}{\partial \mathbf{x}} = \frac{\partial \vartheta}{\partial \mathbf{x}} + \left[\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}}\right]^T \boldsymbol{\lambda} = \mathbf{0}$$

$$\nabla_{\boldsymbol{\lambda}} \phi \equiv \frac{\partial \phi}{\partial \boldsymbol{\lambda}} = \boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$$

- These $m + n$ equations, which define the Lagrange multiplier rule, are solved for the $m + n$ unknowns $\mathbf{x}$ and $\boldsymbol{\lambda}$

Equality Constrained Extrema (ix)

- Suppose we have a stationary point, denoted by $\mathbf{x}^*$, with a corresponding Lagrange multiplier $\boldsymbol{\lambda}^*$
- The point $\mathbf{x}^*$ is a local minimum if the following sufficient condition is satisfied

$$\nabla^2_{\mathbf{x}}\phi \equiv \left.\frac{\partial^2\phi}{\partial\mathbf{x}\,\partial\mathbf{x}^T}\right|_{(\mathbf{x}^*,\boldsymbol{\lambda}^*)} \quad \text{must be positive definite}$$

- The sufficient condition can be simplified by checking the positive definiteness of a matrix that is always smaller than the $n \times n$ matrix above
- Rewrite the loss function as

$$\vartheta(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) \equiv \vartheta(\mathbf{y}, \mathbf{z})$$

where $\mathbf{y}$ is an $m \times 1$ vector and $\mathbf{z}$ is an $p \times 1$ vector (with $p = n - m$)

Equality Constrained Extrema (x)

- The necessary conditions are still the same with $\mathbf{x} \equiv \begin{bmatrix} \mathbf{y}^T & \mathbf{z}^T \end{bmatrix}^T$
- But the sufficient condition can now be determined by checking the definiteness of the following $p \times p$ matrix

$$Q \equiv \Big\{ \left[\nabla_{\mathbf{z}} \boldsymbol{\psi}\right]^T \left[\nabla_{\mathbf{y}} \boldsymbol{\psi}\right]^{-T} \left[\nabla^2_{\mathbf{y}} \phi\right] \left[\nabla_{\mathbf{y}} \boldsymbol{\psi}\right]^{-T} \left[\nabla_{\mathbf{z}} \boldsymbol{\psi}\right] + \nabla^2_{\mathbf{z}} \phi$$

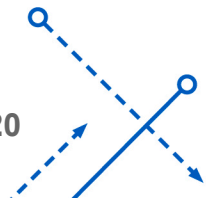
$$- \left[\nabla_{\mathbf{z}} \nabla_{\mathbf{y}} \phi\right] \left[\nabla_{\mathbf{y}} \boldsymbol{\psi}\right]^{-1} \left[\nabla_{\mathbf{z}} \boldsymbol{\psi}\right] - \left[\nabla_{\mathbf{z}} \boldsymbol{\psi}\right]^T \left[\nabla_{\mathbf{y}} \boldsymbol{\psi}\right]^{-T} \left[\nabla_{\mathbf{y}} \nabla_{\mathbf{z}} \phi\right] \Big\} \Big|_{(\mathbf{y}^*, \mathbf{z}^*, \boldsymbol{\lambda}^*)}$$

where $[\nabla_{\mathbf{z}} \nabla_{\mathbf{y}} \phi]$ and $[\nabla_{\mathbf{y}} \nabla_{\mathbf{z}} \phi]$ are $p \times m$ and $m \times p$ matrices, respectively, made up of the partial derivatives with respect to $\mathbf{y}$ and $\mathbf{z}$

- A stationary point is a local minimum (maximum) if Q is positive (negative) definite
- The inverse of an $m \times m$ matrix must be taken, which is smaller than the $n \times n$ Hessian matrix

Equality Constrained Extrema (xi)

- Physical interpretation of a Lagrange multiplier
  - It's the rate of change of the quantity being optimized as a function of the constraint variable
    - More specifically, the value of the Lagrange multiplier at the solution of the problem is equal to the rate of change in the maximal value of the objective function as the constraint is relaxed
    - Tells how the rate we are "pushing" away from the constraint
  - In Lagrangian mechanics the equations of motion are derived by finding stationary points of the action, the time integral of the difference between kinetic and potential energy
    - Force on a particle due to a scalar potential, $F = -\nabla V$, can be interpreted as a Lagrange multiplier determining the change in action (transfer of potential to kinetic energy) following a variation in the particle's constrained trajectory



Constrained Example (i)

- Consider the following loss function

$$\vartheta = 6 - \frac{y}{2} - \frac{z}{3}$$

subject to a constraint represented by an elliptic cylinder

$$\psi(\mathbf{x}) = 9(y-4)^2 + 4(z-5)^2 - 36 = 0$$

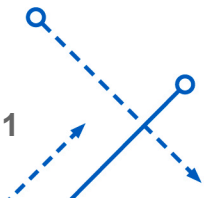
where $\mathbf{x} \equiv \begin{bmatrix} y & z \end{bmatrix}^T$

- The augmented function is given by

Free to use a minus or plus sign.
It does not change the solution.

$$\phi(\mathbf{x},\, \lambda) = 6 - \frac{y}{2} - \frac{z}{3} - \lambda \left[9(y-4)^2 + 4(z-5)^2 - 36\right]$$

- Note that there is only one Lagrange multiplier because there is only one constraint



Constrained Example (ii)

- The necessary conditions give

$$\frac{\partial \phi}{\partial y} = -\frac{1}{2} - 18\lambda(y-4) = 0$$

$$\frac{\partial \phi}{\partial z} = -\frac{1}{3} - 8\lambda(z-5) = 0$$

$$\psi(\mathbf{x}) = 9(y-4)^2 + 4(z-5)^2 - 36 = 0$$

3 equations
3 unknowns

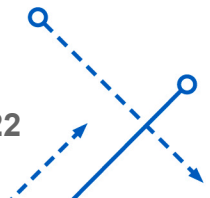
  - Solving these equations for λ gives $\lambda = \pm 1/(36\sqrt{2})$
- Therefore, the stationary points are given by

$$y^* = 4 - \frac{1}{36\lambda} = 4 \pm \sqrt{2}$$

$$z^* = 5 - \frac{1}{24\lambda} = 5 \pm \frac{3}{2}\sqrt{2}$$

$$\lambda^* = \pm \frac{1}{36\sqrt{2}}$$

Need to check whether these two solutions are local minimums or local maximums



Constrained Example (iii)

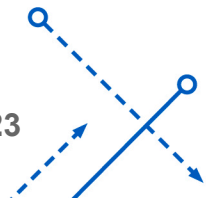
- The sufficient condition is given by

$$\nabla^2_{\mathbf{x}}\phi = -\begin{bmatrix} 18\lambda^* & 0 \\ 0 & 8\lambda^* \end{bmatrix}$$

- The matrix Q is given by

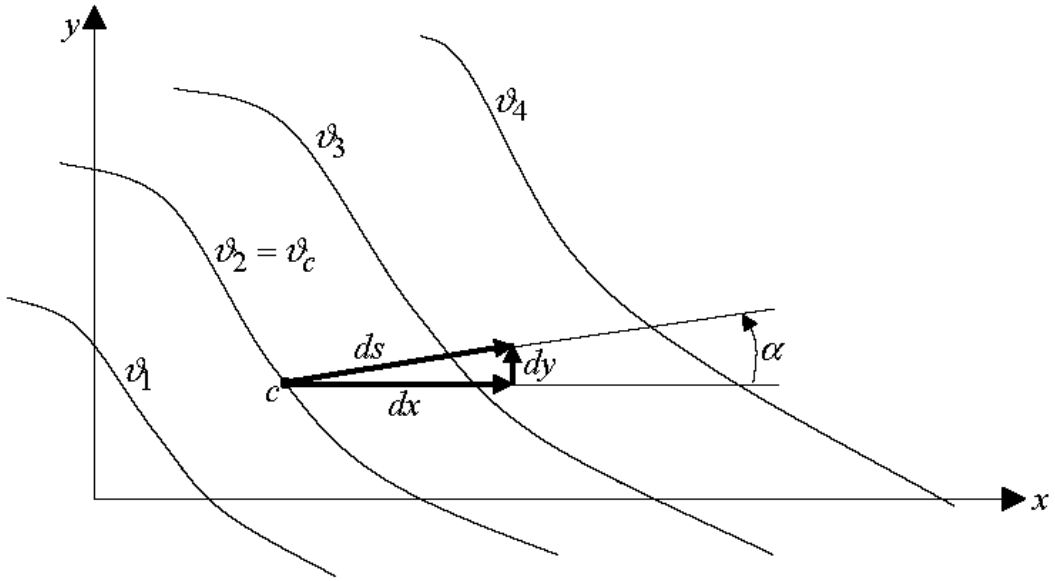
$$Q \equiv q = -8\lambda^* \left[\frac{8}{18} \frac{(z^* - 5)^2}{(y^* - 4)^2} + 1 \right]$$

Clearly, if $\lambda^* = -1/(36\sqrt{2})$, then the stationary point given by $y^* = 4 + \sqrt{2}$ and $z^* = 5 + (3/2)\sqrt{2}$ is a local minimum with $\phi = (7/3) - \sqrt{2}$. Likewise, if $\lambda^* = +1/(36\sqrt{2})$, then the stationary point given by $y^* = 4 - \sqrt{2}$ and $z^* = 5 - (3/2)\sqrt{2}$ is a local maximum with $\phi = (7/3) + \sqrt{2}$.



Nonlinear Optimization (i)

- Must use iterative methods in general
  - Several exist, but we'll only discuss two here
- Some geometrical insights with two variables



  - From the geometry of this figure we have

$$\tan\alpha = \frac{dy}{dx}, \quad \sin\alpha = \frac{dy}{ds}, \quad \cos\alpha = \frac{dx}{ds}$$

Nonlinear Optimization (ii)

- For arbitrary small displacements $(dx,\ dy)$ away from the "current" point $(x_c,\ y_c)$, the differential change in ϑ is given by

$$d\vartheta = \left.\frac{\partial \vartheta}{\partial x}\right|_c dx + \left.\frac{\partial \vartheta}{\partial y}\right|_c dy$$

- If s is the distance measured along an arbitrary line through c, then the rate of change ("directional derivative") of ϑ in the direction of the line is

$$\left.\frac{d\vartheta}{ds}\right|_c = \left.\frac{\partial \vartheta}{\partial x}\right|_c \left.\frac{dx}{ds}\right|_c + \left.\frac{\partial \vartheta}{\partial y}\right|_c \left.\frac{dy}{ds}\right|_c$$

- Making use of the definitions of $\sin\alpha$ and $\cos\alpha$ gives

$$\left.\frac{d\vartheta}{ds}\right|_c = \left.\frac{\partial \vartheta}{\partial x}\right|_c \cos\alpha + \left.\frac{\partial \vartheta}{\partial y}\right|_c \sin\alpha$$

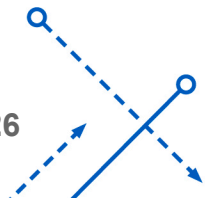
Nonlinear Optimization (iii)

- Look at a couple of interesting cases
  - Suppose we wish to select the particular line for which $(d\vartheta / ds)|_c = 0$
  - The previous equation tells us the angle $\alpha_1 = \alpha$ orienting this line is given by

$$\tan \alpha_1 = \frac{-\left.\frac{\partial \vartheta}{\partial x}\right|_c}{\left.\frac{\partial \vartheta}{\partial y}\right|_c}$$

    which gives the "contour direction"
  - Let's also find the particular direction which results in the minimum or maximum $(d\vartheta / ds)|_c$
  - The necessary condition for the extremum of $(d\vartheta / ds)|_c$ requires

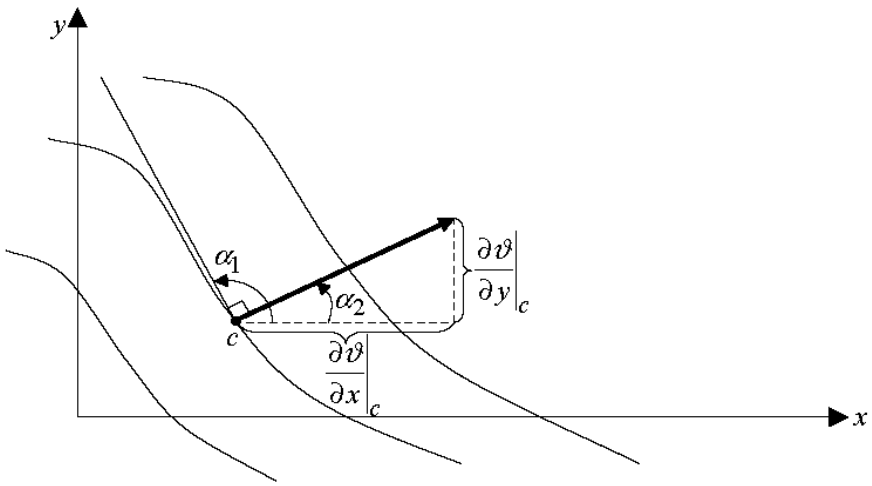
$$\frac{d}{d\alpha}\left(\left.\frac{d\vartheta}{ds}\right|_c\right) = -\left.\frac{\partial \vartheta}{\partial x}\right|_c \sin\alpha + \left.\frac{\partial \vartheta}{\partial y}\right|_c \cos\alpha = 0$$



Nonlinear Optimization (iv)

- From this equation the angle $\alpha_2 = \alpha$ that orients the direction of "steepest descent" or "steepest ascent" is given by

$$\tan \alpha_2 = \frac{\left.\frac{\partial \vartheta}{\partial y}\right|_c}{\left.\frac{\partial \vartheta}{\partial x}\right|_c}$$



which gives the "gradient direction"

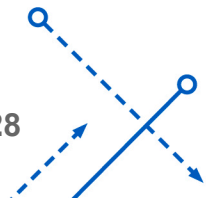
- Notice that $(\tan \alpha_1)\ (\tan \alpha_2) = -1$
- Therefore, α_1 and α_2 orient lines that are perpendicular
- So, the contour line is perpendicular to the gradient line (as shown in the figure)
- This concept can be extended to higher dimensional spaces too

Method of Gradients (i)

- From the previous analysis it is clear that (based only upon the first derivative information) the most favorable direction to take a small step toward minimizing (or maximizing) the function ϑ is down (or up) the locally evaluated gradient of ϑ
  - The "method of gradients" (also known as the "method of steepest descent" for minimizing ϑ or the "method of steepest ascent" for maximizing ϑ) is a sequence of one-dimensional searches along the lines established by successively evaluated local gradients of ϑ
  - Let the local evaluations be denoted by superscripts, e.g.

$$\vartheta^{(k)} = \vartheta\left(\mathbf{x}^{(k)}\right)$$

denotes $\vartheta(\mathbf{x})$ evaluated at the k^{th} set of $\mathbf{x}$-values



Method of Gradients (ii)

- The k^{th} one-dimensional search determines a scalar $\alpha^{(k)}$ such that

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \left[\nabla_{\mathbf{x}} \vartheta\right]^{(k)}$$

  results in

$$\vartheta^{(k+1)} = \vartheta\left(\mathbf{x}^{(k+1)}\right)$$

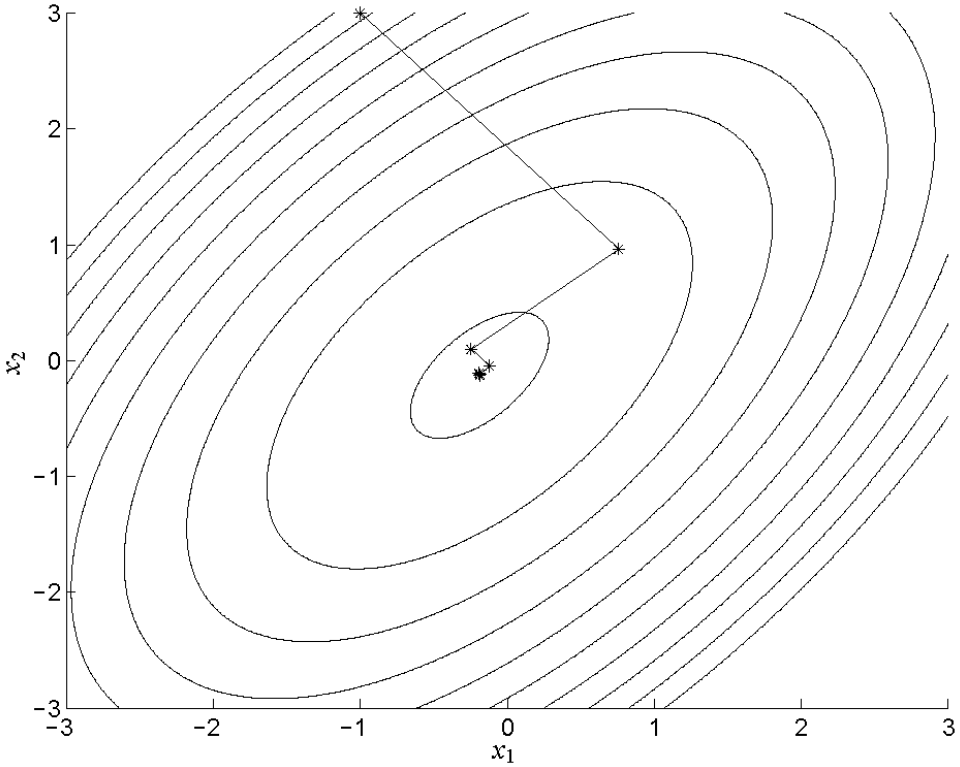
  being a local minimum or maximum
- The one-dimensional search for $\alpha^{(k)}$ can be determined analytically or numerically using various methods
- Sequences of iterations from various starting guesses for Himmelblau's function are shown previously
- The convergence of the gradient method is heavily dependent upon the circularity of the contours
- Often converges rapidly for the first few iterations (far from the solution), but is usually very poor during the final iterations

Example (i)

- Consider the following quadratic loss function

$$\vartheta(\mathbf{x}) = 4x_1^2 + 3x_2^2 - 4x_1x_2 + x_1$$

  - The starting guess is given by $\mathbf{x}(0) = [-1 \quad 3]^T$



$8x_1 - 4x_2 + 1 = 0$

$6x_2 - 4x_1 = 0$

$$\text{minimum of } \mathbf{x}^* = \begin{bmatrix} -3/16 & -1/8 \end{bmatrix}^T$$

  - This function has low eccentricity contours (well behaved)

Example (ii)

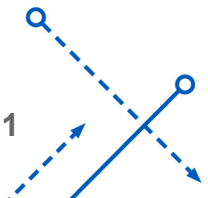
- The Hessian matrix is constant and symmetric for this function

$$\nabla^2_{\mathbf{x}}\vartheta = \begin{bmatrix} 8 & -4 \\ -4 & 6 \end{bmatrix}$$

- The eigenvalues of this matrix are all positive (well behaved, again)
- Iterations are given by

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.7576 & 0.9649 \end{bmatrix}^T$$
$$\mathbf{x}^{(2)} = \begin{bmatrix} -0.2456 & 0.1003 \end{bmatrix}^T$$
$$\mathbf{x}^{(3)} = \begin{bmatrix} -0.1192 & -0.0462 \end{bmatrix}^T$$
$$\mathbf{x}^{(4)} = \begin{bmatrix} -0.1917 & -0.1088 \end{bmatrix}^T$$
$$\mathbf{x}^{(5)} = \begin{bmatrix} -0.1826 & -0.1194 \end{bmatrix}^T$$
$$\mathbf{x}^{(6)} = \begin{bmatrix} -0.1878 & -0.1238 \end{bmatrix}^T$$
$$\mathbf{x}^{(7)} = \begin{bmatrix} -0.1871 & -0.1246 \end{bmatrix}^T$$
$$\mathbf{x}^{(8)} = \begin{bmatrix} -0.1875 & -0.1250 \end{bmatrix}^T$$

- This clearly shows the typical performance of the gradient method, where rapid convergence is given far from the minimum, but slow progress is given near the minimum
- Still, the algorithm converges to the true minimum
- This behavior is also seen from various other starting guesses



Gauss-Newton (i)

- One of the most powerful algorithms
  - Say a loss function ϑ is evaluated at a local point $\mathbf{x}^{(k)}$ by $\Delta\mathbf{x}^{(k)}$

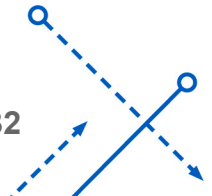
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}$$

    in such a fashion that ϑ is decreased or increased
  - The behavior of ϑ near $\mathbf{x}^{(k)}$ can be approximated by a second-order Taylor series

$$\vartheta \approx \vartheta\left(\mathbf{x}^{(k)}\right) + \Delta\mathbf{x}^T\mathbf{g}^{(k)} + \frac{1}{2}\Delta\mathbf{x}^T H^{(k)}\Delta\mathbf{x}$$

where $\mathbf{g}^{(k)} \equiv \nabla_{\mathbf{x}}\vartheta^{(k)}$ (the gradient of ϑ) and $H^{(k)} \equiv \nabla^2_{\mathbf{x}}\vartheta^{(k)}$ (the Hessian of ϑ)

  - The local strategy is to determine the particular correction vector $\Delta\mathbf{x}^{(k)}$ which minimizes (maximizes) the second-order prediction of ϑ



Gauss-Newton (ii)

- The optimality conditions are given by

necessary conditions

$$\nabla_{\Delta \mathbf{x}} \vartheta = \mathbf{g}^{(k)} + H^{(k)} \Delta \mathbf{x} = \mathbf{0}$$

sufficient condition

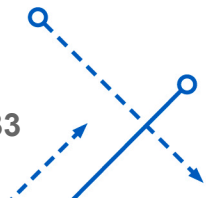
$$\nabla^2_{\Delta \mathbf{x}} \vartheta = H^{(k)} \begin{cases} \text{must be positive definite for minimum} \\ \text{must be negative definite for maximum} \\ \text{must be indefinite for saddle} \end{cases}$$

- From the necessary conditions the local corrections are

$$\Delta \mathbf{x}^{(k)} = - \left[H^{(k)} \right]^{-1} \mathbf{g}^{(k)}$$

- Then the second-order Gauss-Newton algorithm is given by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[H^{(k)} \right]^{-1} \mathbf{g}^{(k)}$$



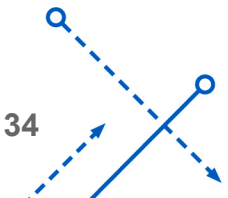
Gauss-Newton (iii)

- There is a pitfall, though!
  - If the sufficient condition is not satisfied, then the correction will be in the wrong direction!!!
  - This pitfall can be circumvented by using a gradient algorithm until the neighborhood of the solution is reached, then testing the sufficient condition and employing the second-order algorithm if it is satisfied
- Good news is that the algorithm converges in one iteration for quadratic loss functions
- Consider the previous loss function

$$\vartheta(\mathbf{x}) = 4x_1^2 + 3x_2^2 - 4x_1x_2 + x_1$$

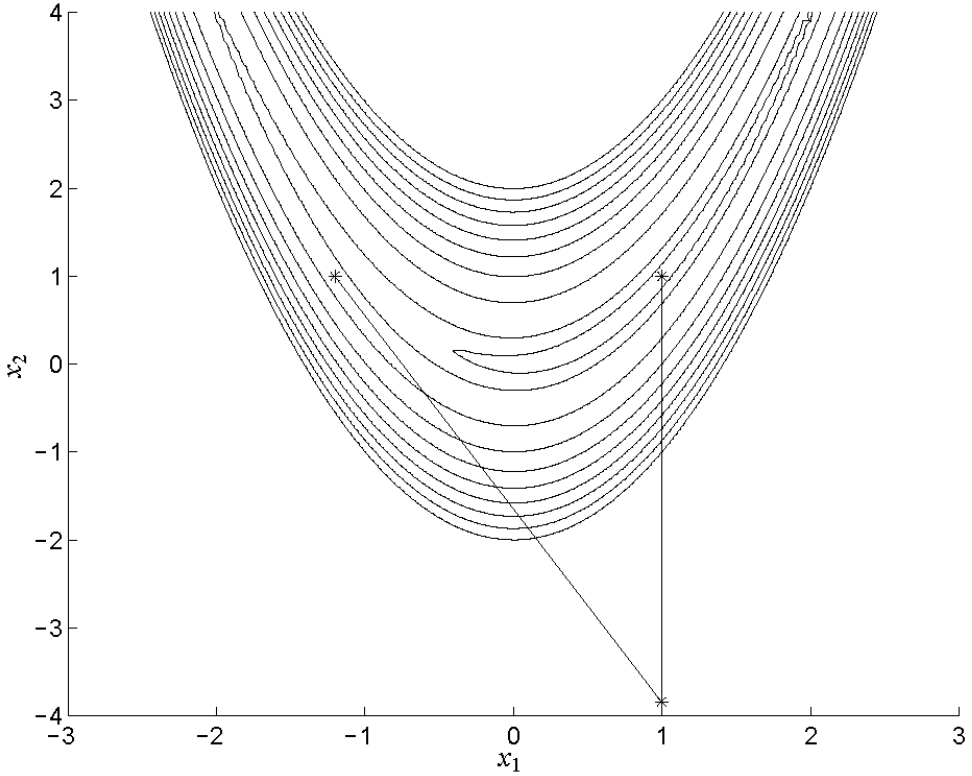
- Gauss-Newton correction after one iteration is the optimal one (!)

$$\mathbf{x}^{(1)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} - \begin{bmatrix} \frac{3}{16} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 8x_1^{(0)} - 4x_2^{(0)} + 1 \\ 6x_2^{(0)} - 4x_1^{(0)} \end{bmatrix} = -\begin{bmatrix} \frac{3}{16} \\ \frac{1}{8} \end{bmatrix}$$



Example (i)

$$\vartheta(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



- Note the highly nonlinear trenches for this function
- The starting guess is given by $\mathbf{x}^{(0)} = [-1.2 \quad 1]^T$
- For this particular problem, the gradient method does not converge to the true minimum of $\mathbf{x}^* = [1 \quad 1]^T$ even after 1,000 iterations!

fmincon - MATLAB

Example (ii)

- However, the second-order algorithm converges in just two iterations
- The iterations are given by

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1.0000 & -3.8400 \end{bmatrix}^T$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} 1.0000 & 1.0000 \end{bmatrix}^T$$

- The Hessian matrix evaluated for this function is given by

$$\nabla^2_{\mathbf{x}}\vartheta = \begin{bmatrix} -400(x_2 - x_1^2) + 800x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

which is always positive definite at all the iterations