MA2000: OTML

Nachiketa Mishra

Indian Institute of Information Technology, Design & Manufacturing, Kancheepuram

Function Example

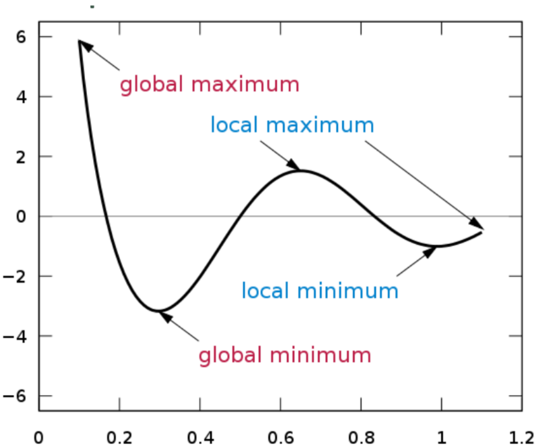
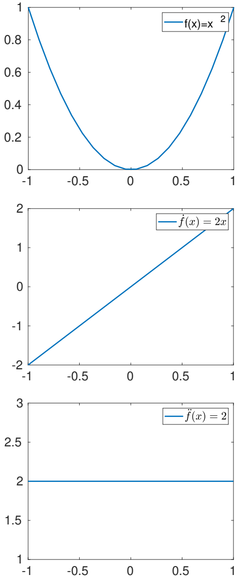


Figure: Stationary points of a function

Quickest ever review of multivariate calculus

- Derivative
- Partial Derivative
- Gradient Vector

Function and its derivatives



- Slope of the tangent line

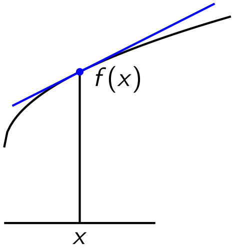
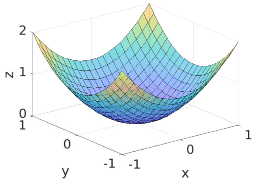


Figure: The function f is drawn in black and the tangent line to $f(x)$ is drawn in blue. The derivative of f at x is the slope of the tangent line

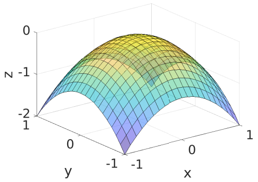
- It is easy when a function is univariate.

Partial Derivative – Multivariate Functions

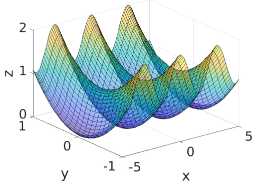
For multivariate functions (e.g two variables) we need partial derivatives – one per dimension.
Examples of multivariate functions:



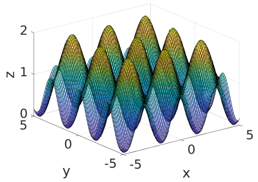
(a) $f = x^2 + y^2$



(b) $f = -x^2 - y^2$



(c) $f = \cos^2(x) + y^2$



(d) $f = \cos^2(x) + \cos^2(y)$

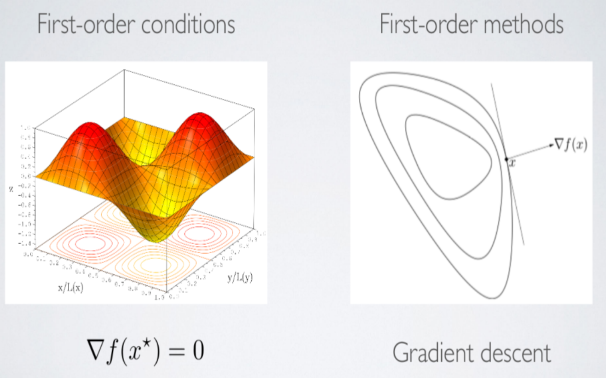
Gradient

- The gradient is the generalization of the derivative to multivariate functions.
- The gradient of f at $\mathbf{x} \in \mathbb{R}^n$ is written as $\nabla f(\mathbf{x})$ and is a vector.
- Each component of that vector is the partial derivative of f with respect to that component:

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \quad \frac{\partial f(\mathbf{x})}{\partial x_2} \quad \dots \quad \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^T .$$

- Why are we interested in gradients?
  - It captures the local slope of the function, allowing us to predict the effect of taking a small step from a point in any direction.
  - The gradient points are in the direction of the increase of a function $f(x)$, Thus $-\nabla f(x)$ gives the direction of descent. Hence, it helps in pointing to local minima.

Gradient descent method

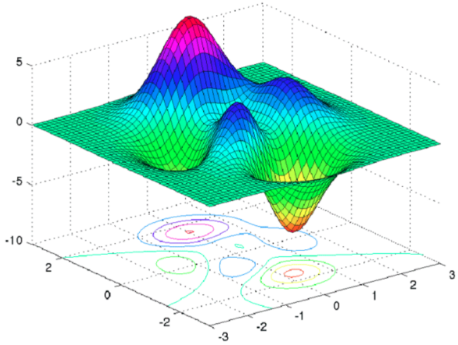


- What is a level curve?
- Where does ∇ point to?
- How does gradient help in going to minima?

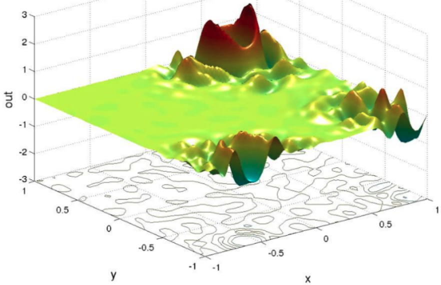
Level curve and its graphical representation

The level curve of a scalar-valued function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is defined as a set as follow

$$f_t = \{(x, y) | f(x, y) = t, t \in \mathbb{R}\}.$$

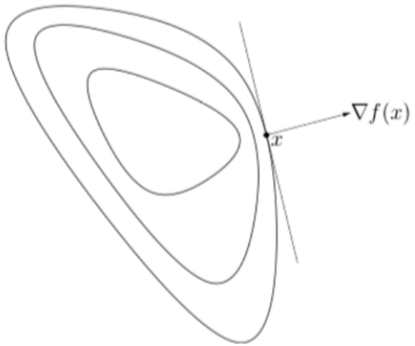


(a) Color coding is needed to distinguish maxima from minima



(b) Level curves are convenient for showing directions towards minima/maxima

Where does ∇ point to?



Let $P = (x_0, y_0, z_0)$ be a point in the level curve $f(x, y, z) = c$. Let $r(t) = (x(t), y(t), z(t))$ be a curve on the level surface with $r(t_0) = (x_0, y_0, z_0)$. We let $g(t) = f(x(t), y(t), z(t))$. Since the curve is on the level surface we have $g(t) = f(x(t), y(t), z(t)) = c$. Differentiating this equation with respect to t gives

$$\frac{dg}{dt} = \left.\frac{\partial f}{\partial x}\right|_P \left.\frac{dx}{dt}\right|_{t_0} + \left.\frac{\partial f}{\partial y}\right|_P \left.\frac{dy}{dt}\right|_{t_0} + \left.\frac{\partial f}{\partial z}\right|_P \left.\frac{dz}{dt}\right|_{t_0} = 0.$$

In vector form this is

$$\left(\left.\frac{\partial f}{\partial x}\right|_P, \left.\frac{\partial f}{\partial y}\right|_P, \left.\frac{\partial f}{\partial z}\right|_P\right) \cdot \left(\left.\frac{dx}{dt}\right|_{t_0}, \left.\frac{dy}{dt}\right|_{t_0}, \left.\frac{dz}{dt}\right|_{t_0}\right) = 0,$$

which implies

$$\nabla f|_P . r'(t_0) = 0.$$

Hence, Gradient is perpendicular to the tangent to any curve that lies on the surface and goes through P.

How does gradient help in going to minima?

- $-\nabla f(x)$ gives direction of descent.

Gradient Descent Algorithm

1. Start from initial vector: $x^{(0)}$
2. From the current position move in the direction of $-\nabla f(x^{(0)})$

$$x^{(i+1)} \longleftarrow x^{(i)} + t[-\nabla f(x^{(i)})], i = 1, 2, \cdots,$$

   where t is a parameter that tells how far to move also called as **Learning Rate.**
3. When to stop?
   - When grad is going flat, i.e., $\nabla f(x^{(i)}) \approx 0$ (at least **machine precision**)
   - Indeed, then $x^{(i)}$ won't update, so another check: $\|x^{(i)} - x^{(i-1)}\|$ is "small enough"
4. Return $x^{(i)}$ to be minima

How does gradient help in going to maxima?

- $\nabla f(x)$ gives direction of ascent.

Gradient Ascent Algorithm

1. Start from initial vector: $x^{(0)}$
2. From the current position move in the direction of $\nabla f(x^{(0)})$

$$x^{(i+1)} \longleftarrow x^{(i)} + t\nabla f(x^{(i)}), i = 1, 2, \cdots ,$$

   where t is a parameter that tells how far to move
3. When to stop?
   - When grad is going flat, i.e., $\nabla f(x^{(i)}) \approx 0$ (at least machine precision)
   - Indeed, then $x^{(i)}$ won't update, so another check: $\|x^{(i)} - x^{(i-1)}\|$ is "small enough"
4. Return $x^{(i)}$ to be maxima

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Solution:

- The gradient of f is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}^T = \begin{bmatrix} 1 + 4x_1 + 2x_2 \\ -1 + 2x_1 + 2x_2 \end{bmatrix}, S_1 = -\nabla f_1 = \nabla f(X_1) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

- Iteration 1:
  - To find X_2 , we need to find the optimal step length t_1^*. Thus, we minimize $f(X_1 + t_1 S_1) = f(-t_1, t_1) = t_1^2 - 2t_1$ with respect to t_1 .
  - Since $\frac{df}{dt_1} = 0$ at $t_1 = 1$, we get

$$X_2 = X_1 + t_1^* S_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \text{ and } \nabla f_2 = \nabla f(X_2) = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  - Thus, X_2 is not optimal.

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

- Iteration 2:
  - $S_2 = -\nabla f_2 = \nabla f(X_2) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
  - To minimize $f(X_2 + t_2 S_2) = f(-1 + t_2, 1 + t_2) = 5t_2^2 - 2t_2 - 1$ we set $\frac{df}{dt_2} = 0$, which gives $t_2^* = 1/5$
  - Thus,

$$X_3 = X_2 + t_2^* S_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \frac{1}{5} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.8 \\ 1.2 \end{bmatrix} \text{ and } \nabla f_3 = \nabla f(X_3) = \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  - Thus, X_3 is not optimal.

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

- Iteration 3:
  - $S_3 = -\nabla f_3 = \nabla f(X_3) = \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}$
  - As $f(X_3 + t_3 S_3) = f(-0.8 - 0.2t_3, 1.2 + 0.2t_3) = 0.04t_3^2 - 0.08t_3 - 1.20$, $\dfrac{df}{dt_2} = 0$, at $t_3^* = 1.0$
  - Thus,

$$X_4 = X_3 + t_3^* S_3 = \begin{bmatrix} -0.8 \\ 1.2 \end{bmatrix} + 1.0 \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix} = \begin{bmatrix} -1.0 \\ 1.4 \end{bmatrix} \text{ and } \nabla f_4 = \nabla f(X_4) = \begin{bmatrix} -0.2 \\ -0.2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  - Thus, X_4 is not optimal.
- Continuing in this way, we get the optimum point $X^* = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$

Conjugate Gradient

- Any minimization method that makes use of the conjugate directions are quadratically convergent.
- This property of **quadratically convergent** is very useful because it ensures that the method will minimize a quadratic function in n steps or less.
- Thus, convergence characteristics of the gradient descent method can be improved greatly by modifying it into a conjugate gradient method (which can be considered as a conjugate directions method involving the use of the gradient of the function).

Conjugate Gradient

- The conjugate gradient the method overcomes this issue by borrowing inspiration from methods for optimizing quadratic functions:

$$\underset{x}{\text{minimize}}\ f(x) = \frac{1}{2}x^T A x + b^T x + c,$$

  where A is symmetric and positive definite.

Conjugate Direction Vs Gradient Direction

- The conjugate direction

Algorithm of Conjugate Gradient

1. Start from initial vector: X_1
2. Set the first search direction $S_1 = -\nabla f(x^{(1)}) = -\nabla f_1$.
3. Find the point X_2 according to the relation

$$X_2 \longleftarrow X_1 + t_1^* S_1.$$

   where t_1^* is the optimal step length in the direction S_1.
4. find $\nabla f_i = \nabla f(X_i)$, and set

$$S_i = -\nabla f_i + \frac{|\nabla f_i|^2}{|\nabla f_{i-1}|^2} S_{i-1}$$

5. Compute the optimal step length t_i^* in the direction S_i, and find the new point

$$X_{i+1} = X_i + t_i^* S_i.$$

6. If X_{i+1} is optimal, stop the process. Otherwise, set the value of $i = i + 1$ and go to step 4.

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Solution:

- The gradient of f is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}^T = \begin{bmatrix} 1 + 4x_1 + 2x_2 \\ -1 + 2x_1 + 2x_2 \end{bmatrix}, S_1 = -\nabla f_1 = \nabla f(X_1) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

- Iteration 1:
  - To find X_2 , we need to find the optimal step length t_1^*. Thus, we minimize $f(X_1 + t_1 S_1) = f(-t_1, t_1) = t_1^2 - 2t_1$ with respect to t_1 .
  - Since $\frac{df}{dt_1} = 0$ at $t_1 = 1$, we get

$$X_2 = X_1 + t_1^* S_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \text{ and } \nabla f_2 = \nabla f(X_2) = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  - Thus, X_2 is not optimal.

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

- Iteration 2:
  - $\nabla f_2 = \nabla f(X_2) = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$
  - The next search direction is

$$S_2 = -\nabla f_2 + \frac{|\nabla f_2|^2}{|\nabla f_1|^2} S_1.$$

    Here $|\nabla f_1|^2 = 2$ and $|\nabla f_2|^2 = 2$.

  - Therefore, $S_2 = -\begin{bmatrix} -1 \\ -1 \end{bmatrix} + \frac{2}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

Example

Minimize $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$ starting from the point $X_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

- Iteration 2: continue
  - To find t_2^*, we minimize

$$\begin{aligned} f(X_2 + t_2 S_2) &= f(-1, 1 + 2t_2) \\ &= -1 - (1 + 2t_2) + 2 - 2(1 + 2t_2) + (1 + 2t_2)^2 \\ &= 4t_2^2 - 2t_2 - 1. \end{aligned}$$

  - As $\dfrac{df}{dt_2} = 8t_2 - 2 = 0$ at $t_2^* = \dfrac{1}{4}$.
  - Thus, the optimal value $X_3 = X_2 + t_2^* S_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$.
  - $\nabla f_3 = \nabla f(X_3) = 0$.