Tutorial on

Optimization

BITS F464 Machine Learning

8th February 2020

$$\text{optimise: } f(x) \qquad \text{OR} \qquad \begin{array}{l}\text{optimise: } f(x)\\ \text{subject to: } a \le x \le b\end{array}$$

(a) (b)

- Here *optimise* means *maximise* or *minimise*
- The optimisation problem in (a) is called *unconstrained* optimisation, and in (b) is called *constrained* optimsation
- If a constrained optimisation problem has no solution, then constraining the value of x may give a solution
  - For example, $f(x) = x$ has no finite maximum (or minimum). But if $a \le x \le b$ then the maximum (and minimum) are well-defined

Feasible Solutions

Values of x satisfying the constraints are called *feasible* solutions. The constrained optimisation problem is to find the the optimal value of $f(x)$ amongst feasible solutions

optimise: $f(x)$

subject to: $a \leq x \leq b$

(b)

Global and Local Optimum

If for some feasible x

$f(x) \leq f(x')$ — x is called a *global* minimum

$f(x) \leq f(x')$ for $x' \in Nbd(x)$ — *local* minimum

Global and Local Optimum

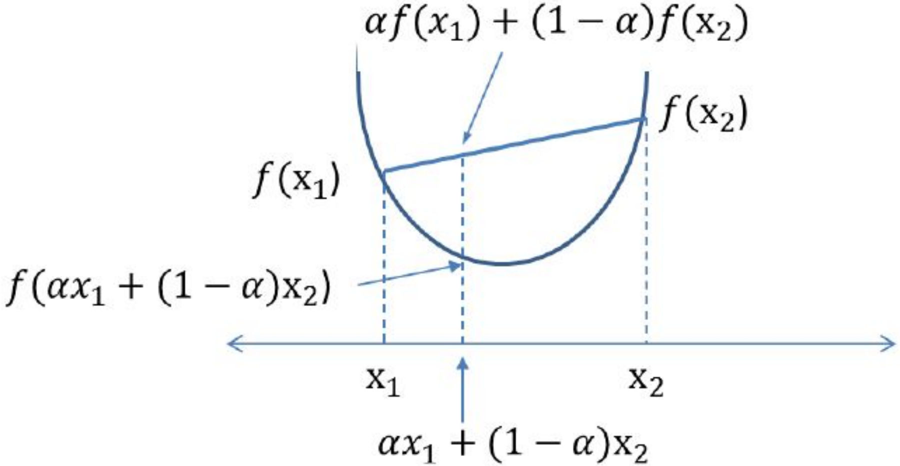
If for some feasible x

$f(x) \leq f(x')$ — x is called a *global* minimum

$f(x) \leq f(x')$ for $x' \in Nbd(x)$ — *local* minimum

Constrained Optimum != Global Optimum

Convex Functions



$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \qquad (0 \leq \alpha \leq 1)$$

Strictly Convex Functions

A function is strictly convex if the line segment is **strictly above** the function (Ex. a linear function is not strictly convex)

$$\forall x_1 \neq x_2 \in X, \forall t \in (0,1): \qquad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$

Examples of Convex Functions

Examples of convex functions are:

- Linear functions of the form ax + b (for all a, b)
- Power functions of the form $|x|^p$ for p>=1
- Exponential functions of the form e^{ax} (for all a)
- Norms like | x | or $| x |_2$
- $\max(x_1, x_2, \ldots, x_n)$ is convex

Prove that they are convex!

Important Results

- For a convex function, any local minimum is also a global minimum
- For a strictly convex function, if there is a local minimum then it is a unique global minimum

Multivariate Unconstrained Optimisation

Goal: Optimize u = f(**x**)

The results from the calculus require counterparts to the first and second-differentials

Gradient

This gradient, usually denoted ∇f, is the vector:

$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k}\right)$$

This is also denoted in matrix notation as:

$$\left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k}\right]^T$$

Gradient - Example

Let $f(x_1, x_2, x_3) = 3x_1^2 x_2 - x_2^2 x_3^3$. What is ∇f at $\mathbf{x}_0 = [1, 2, 3]^T$?

Gradient - Example

Let $f(x_1, x_2, x_3) = 3x_1^2 x_2 - x_2^2 x_3^3$. What is ∇f at $\mathbf{x}_0 = [1, 2, 3]^T$?

$$\nabla f = \begin{bmatrix} 6x_1 x_2 \\ 3x_1^2 - 2x_2 x_3^3 \\ -3x_2^2 x_3^2 \end{bmatrix}$$

$$\nabla f|_{\mathbf{x}_0} = \begin{bmatrix} 12 \\ -105 \\ -108 \end{bmatrix}$$

This is the direction of greatest increase of f at the point $\mathbf{x}_0$

Hessian

The *Hessian* matrix $\mathbf{H}_f$ associated the function $f(\mathbf{x})$ is the matrix $\mathbf{H}|_f$

$$\left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right] \qquad (i, j = 1 \ldots n)$$

We are usually interested in the value of the Hessian matrix at some value $\mathbf{x}_0$. This is denoted by $\mathbf{H}|_{f,\mathbf{x}_0}$

If the second partial derivatives of a function f are continuous at $\mathbf{x}_0$ then the Hessian $\mathbf{H}|_{f,\mathbf{x}_0}$ will be symmetric

Hessian - Example

$$f(x_1, x_2, x_3) = 3x_1^2x_2 - x_2^2x_3^3$$

What is the Hessian for f at $\mathbf{x}_0$ as above?

Hessian - Example

$f(x_1, x_2, x_3) = 3x_1^2 x_2 - x_2^2 x_3^3$

What is the Hessian for f at $\mathbf{x}_0$ as above?

$$\mathbf{H}|_f = \begin{bmatrix} 6x_2 & 6x_1 & 0 \\ 6x_1 & -2x_3^3 & -6x_2x_3^2 \\ 0 & -6x_2x_3^2 & -6x_2^2x_3 \end{bmatrix}$$

Substituting the values for $\mathbf{x}_0$ we get:

$$\mathbf{H}|_f, \mathbf{x}_0 = \begin{bmatrix} 12 & 6 & 0 \\ 6 & -54 & -108 \\ 0 & -108 & -72 \end{bmatrix}$$

Negative Definiteness

A symmetric matrix is negative definite if and only if all of its principal minors of even order are positive and all of its principal minors of odd order are negative.

Negative Definiteness

Let:

$$A_1 = \begin{vmatrix} a_{11} \end{vmatrix} \; A_2 = (-1) \times \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \; A_3 = (-1)^2 \times \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \cdots$$

Or in general:

$$A_n = (-1)^{n-1} det(A)$$

A is negative definite iff $A_1, A_2, \ldots, A_n$ are all negative and negative semi-definite iff there exists some $r < n$ s.t. the A_i for $i \leq r$ are negative, and are 0 for $i > r$.

If $f(\mathbf{x})$ has both ∇f and second partial derivates defined in some ϵ-neighbourhood around $\mathbf{x}^*$ and $\nabla f|_{\mathbf{x}}^* = \mathbf{0}$ and $\mathbf{H}|_f, \mathbf{x}^*$ is negative-definite then $f(\mathbf{x})$ has a local maximum at $\mathbf{x}^*$

Is the Hessian obtained earlier negative, or semi-negative, or neither at $\mathbf{x}_0$?

Is the Hessian obtained earlier negative, or semi-negative, or neither at $\mathbf{x}_0$?

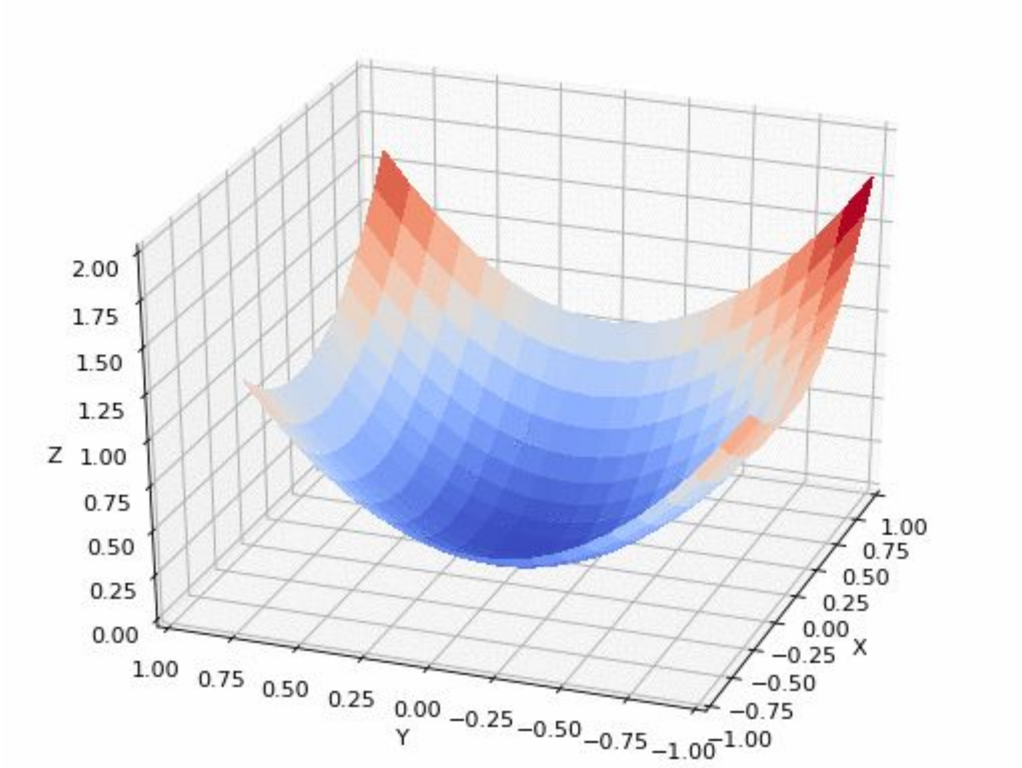
Answer. Since $A_1 = 12$ for the Hessian, it is not negative or semi-negative at $\mathbf{x}_0$.

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \quad \text{and} \quad \nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Numerical Optimization

In general, analytical expressions for optimal values of a multivariate function $f(\mathbf{x})$ are hard to obtain

Gradient Descent



Gradient Ascent

1. Start with some guess $\mathbf{x}_0$
2. Determine subsequent vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$ using the update formula:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta^* \nabla f|_{\mathbf{x}_k}$$

   where η^* is the value of a scalar η that results in the maximum value for $f(\mathbf{x}_k + \eta \nabla f|_{\mathbf{x}_k})$ (often, η^* is just taken to be a small constant)
3. Stop when $\mathbf{x}_k \approx \mathbf{x}_{k+1}$

This is a greedy search in the direction of maximal increase. Replacing the $+$ sign by $-$ in the update formula will result in a search in the direction of maximal decrease. The resulting procedure is gradient *descent*

Convex functions and Gradient Ascent/Descent

In case of convex functions, finding Local Optima is enough as it is also the global optima.

Exercise

Show that at every interation, gradient ascent at a point $\mathbf{x}_k$ moves in the direction of greatest increase of $f(\mathbf{x}_k)$

Exercise

Show that at every interation, gradient ascent at a point $\mathbf{x}_k$ moves in the direction of greatest increase of $f(\mathbf{x}_k)$

The rate of change of $f(\mathbf{x})$ at $\mathbf{x}_k$ in the direction of any unit vector $\mathbf{U}$ is:

$$\nabla f|_{\mathbf{x}_k} \cdot \mathbf{U} = |\nabla f||\mathbf{U}|cos\theta$$

This is a maximum when $cos\theta = 1$ or $\theta = 0$. That is, $\mathbf{U}$ is in the same direction as $\nabla f|_{\mathbf{x}_k}$. Any scalar multiple $\eta^* \nabla f|_{\mathbf{x}_k}$ is in this direction.

Exercise

Maximise $z = f(x_1, x_2) = -(x_1 - \sqrt{5})^2 - (x_2 - \pi)^2 - 10.$

Find the maximum for the function f above, using gradient ascent.

$$\mathbf{x}_0 = [6.597, 5.891]^T$$

Exercise

Maximise $z = f(x_1, x_2) = -(x_1 - \sqrt{5})^2 - (x_2 - \pi)^2 - 10.$

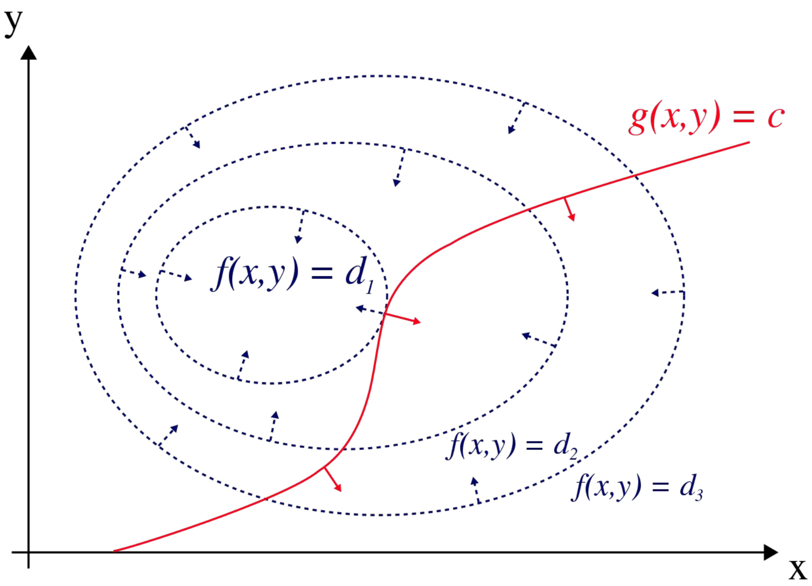
Find the maximum for the function f above, using gradient ascent.

$$\mathbf{x}_0 = [6.597, 5.891]^T$$

$x_1 = \sqrt{5}$ and $x_2 = \pi$. The value f at this point is -10, which a maximum for f

Lagrange Multipliers

Constrained Multivariable Optimization



Maximize f(x,y)
Subject to g(x,y)=0

$$
\begin{array}{lcl}
\text{minimise: } f(\mathbf{x}) & & \text{maximise: } f(\mathbf{x}) \\
\text{subject to:} & & \text{subject to:} \\
g_1(\mathbf{x}) \leq 0 & & g_1(\mathbf{x}) \leq 0 \\
g_2(\mathbf{x}) \leq 0 & \text{OR} & g_2(\mathbf{x}) \leq 0 \\
\vdots & & \vdots \\
g_m(\mathbf{x}) \leq 0 & & g_m(\mathbf{x}) \leq 0 \\
\text{(a)} & & \text{(b)}
\end{array}
$$

Provided some conditions on the partial derivatives of f and g are satisfied, then it can be shown that if for some $\mathbf{x}^*$:

$$-\nabla f|_{\mathbf{x}}^* = \lambda_i \nabla g_i(\mathbf{x}^*)$$

then $\mathbf{x}^*$ is a solution the optimisation problem (a)

The **Lagrange multiplier theorem** roughly states that at any stationary point of the function that also satisfies the equality constraints, the **gradient of the function at that point** can be expressed as a **linear combination of the gradients of the constraints at that point**, with the **Lagrange multipliers acting as coefficients**.

Similarly if:

$$\nabla f|_{\mathbf{x}}^{*} = \lambda_i \nabla g_i(\mathbf{x}^*)$$

then $\mathbf{x}^*$ is a solution to the optimisation problem (b)

We define the *Lagrangian* for (a) as the function

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

Then:

$$\nabla L = \nabla f(\mathbf{x}) - \sum_i \lambda_i \nabla g_i(\mathbf{x})$$

It is clear that for all points $\mathbf{x}^*$ s.t. $\nabla L|_{\mathbf{x}}^* = \mathbf{0}$
$\nabla f|_{mathbfx}^* + \sum \lambda_i g_i(\mathbf{x}^*) = 0$ and $\mathbf{x}^*$ is a solution to (a)

Similarly, the Lagrangian for (b) is:

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

and a similar result follows

$\nabla L = \mathbf{0}$ is a system of $n + m$ equations in $n + m$ unknowns:

$$\frac{\partial L}{\partial x_i} = 0 \qquad (i = 1, 2, \ldots n)$$

$$\frac{\partial L}{\partial \lambda_j} = 0 \qquad (j = 1, 2, \ldots, m)$$

Example

Maximise $f(x_1, x_2, x_3) = -(x_1 + x_2 + x_3)$ subject to the constraints:

$$x_1^2 + x_2 \leq 3$$

$$x_1 + 3x_2 + 2x_3 \leq 7$$

We first bring this into the standard form for the constraints:

maximise: $z = f(x_1, x_2, x_3) = -(x_1 + x_2 + x_3)$

subject to:

$x_1^2 + x_2 - 3 \leq 0$

$x_1 + 3x_2 + 2x_3 - 7 \leq 0$

The Lagrangian is the function:

$$L(x_1, x_2, x_3, \lambda_1, \lambda_2) = -(x_1+x_2+x_3)-\lambda_1(x_1^2+x_2-3)-\lambda_2(x_1+$$

The solution to the constrained maximisation problem is amongst the solutions to the equations in $\nabla L = \mathbf{0}$. That is:

$$\frac{\partial L}{\partial x_1} = -1 - 2x_1\lambda_1 - \lambda_2 = 0$$
$$\frac{\partial L}{\partial x_2} = -1 - \lambda_1 - 3\lambda_2 = 0$$
$$\frac{\partial L}{\partial x_3} = -1 - 2\lambda_2 = 0$$
$$\frac{\partial L}{\partial \lambda_1} = -(x_1^2 + x_2 - 3) = 0$$
$$\frac{\partial L}{\partial \lambda_2} = -(x_1 + 3x_2 + 2x_3 - 7) = 0$$

Solving, we get $\lambda_1 = 0.5$, $\lambda_2 = -0.5$, $x_1 = -0.5$, $x_2 = 2.75$, and $x_3 = -0.375$. This gives $z = -1.875$ as the maximum, and 1.875 as the minimum for $f(x_1, x_2, x_3)$