Machine Learning: Mathematical Background

Calculus

He Wang

Website: drhewang.com

Lecture Overview

Maths & Machine Learning [1]

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

[1] Much of this lecture is drawn from 'Mathematics for Machine Learning' by Garrett Thomas

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

Learning Outcomes for Today's Lecture

- By the end of this lecture you should be familiar with some fundamental objects in and results of **Calculus**
- For the most part we will concentrate on the statement of results which will be of use in the main body of this module
- However we will not be so concerned with the proof of these results

Lecture Overview

Derivatives

- For a function, $f : \mathbb{R} \to \mathbb{R}$, the **derivative** is defined as:

$$\frac{df}{dx} = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta} = f'(x)$$

- The **second derivative** is defined to be the derivative of the derivative:

$$\frac{d^2 f}{dx^2} = \lim_{\delta \to 0} \frac{f'(x+\delta) - f'(x)}{\delta} = f''(x)$$

Taylor Series

- For small changes, δ, about a point $x = \widetilde{x}$, any smooth function, f, can be written as:

$$\begin{aligned} f(\widetilde{x} + \delta) &= f(\widetilde{x}) + \sum_{i=1}^{\infty} \frac{\delta^i}{i!} \left(\frac{d}{dx} \right)^i f(x) \bigg|_{x=\widetilde{x}} \\ &= f(\widetilde{x}) + \delta \frac{df}{dx} \bigg|_{x=\widetilde{x}} + \frac{\delta^2}{2} \frac{d^2 f}{dx^2} \bigg|_{x=\widetilde{x}} + \dots \end{aligned}$$

Rules for Combining Functions

- **Sum Rule**
  $\forall$ functions $f, g \quad \forall \alpha, \beta \in \mathbb{R}$:

$$(\alpha f(x) + \beta g(x))' = \alpha f'(x) + \beta g'(x)$$

- **Product Rule**
  $\forall$ functions f, g:

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

- **Chain Rule**
  if $f(x) = h(g(x))$ then:

$$f'(x) = h'(g(x)).g'(x)$$

Common Derivatives

$f(x)$	$f'(x)$
x^n	nx^{n-1}
e^{kx}	ke^{kx}
$\ln x$	$\frac{1}{x}$

- Where n, k are constants.

Partial Derivatives

- For a function that depends on n variables, $\{x_i\}_{i=1}^n$, $f : (x_1, x_2, ..., x_n) \mapsto f(x_1, x_2, ..., x_n)$, then the **partial derivative** wrt x_i is defined as:

$$\frac{\partial f}{\partial x_i} = \lim_{\delta \to 0} \frac{f(x_1, x_2, ..., x_i + \delta, ..., x_n) - f(x_1, x_2, ..., x_i, ..., x_n)}{\delta}$$

  So the partial derivative wrt x_i keeps the state of the other variables fixed

Lecture Overview

Gradients

- The **gradient** of $f : \mathbb{R}^n \to \mathbb{R}$, denoted by $\nabla_{\mathbf{x}} f$ is given by the vector of partial derivatives:

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

Directional Derivative

- The **directional derivative** in direction $\widehat{\mathbf{u}}$, (where $\|\widehat{\mathbf{u}}\|_2^2 = 1$), is the slope of $f(\mathbf{x})$ in the direction of $\widehat{\mathbf{u}}$:

$$\nabla_{\mathbf{x}} f \cdot \widehat{\mathbf{u}}$$

- By the definition of angle, this can be re-written as:

$$\begin{aligned} \nabla_{\mathbf{x}} f \cdot \widehat{\mathbf{u}} &= \|\nabla_{\mathbf{x}} f\|_2 \|\widehat{\mathbf{u}}\|_2 \cos\theta \\ &= \|\nabla_{\mathbf{x}} f\|_2 \cos\theta \end{aligned}$$

  Where θ is the angle between the gradient vector and $\widehat{\mathbf{u}}$

- Thus the directional derivative is maximal when $\nabla_{\mathbf{x}} f$ and $\widehat{\mathbf{u}}$ are aligned. In other words $\nabla_{\mathbf{x}} f$ points in the direction of **steepest ascent** on a surface.

Total Derivative

- For a function that depends on n variables, $\{x_i\}_{i=1}^n$, $f : (x_1, x_2, ..., x_n) \mapsto f(x_1, x_2, ..., x_n)$, where $x_i = x_i(t)\ \forall t$, then the **total derivative** wrt t is:

$$\begin{aligned} \frac{df}{dt} &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} \\ &= \left[\frac{dx_1}{dt}, \ldots, \frac{dx_n}{dt} \right]^T \nabla_{\mathbf{x}} f \end{aligned}$$

This follows from the chain rule.

Jacobian

- The **Jacobian** matrix of a vector-valued function, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, defined such that $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})]^T$, denoted by $\nabla_{\mathbf{x}}\mathbf{f}$ or $\frac{\partial(f_1,..,f_m)}{\partial(x_1,..,x_n)}$, is the matrix of all its first order partial derivatives:

$$\nabla_{\mathbf{x}}\mathbf{f} = \frac{\partial(f_1,..,f_m)}{\partial(x_1,..,x_n)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Hessian

- The **Hessian** matrix of $f : \mathbb{R}^n \to \mathbb{R}$, denoted by $\nabla^2_{\mathbf{x}} f$ or $\mathcal{H}(\mathbf{x})$ is a matrix of second order partial derivatives:

$$\nabla^2_{\mathbf{x}} f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial {x_1}^2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Scalar-by-Matrix Derivative

- The derivative of a scalar, y, with respect to the elements of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times m}$, is defined to be:

$$\frac{\partial y}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \cdots & \frac{\partial y}{\partial A_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{n1}} & \cdots & \frac{\partial y}{\partial A_{nm}} \end{bmatrix}$$

Lecture Overview

Matrix Calculus

- Let $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$:
  - $$\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$$
  - $$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

    In particular, if $\mathbf{A}$ is symmetric:

    $$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$$

Matrix Calculus

- Let $\mathbf{a}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, and $\mathbf{B} \in \mathbb{R}^{n \times m}$:
  - $$\frac{\partial \mathbf{a}^T \mathbf{B} \mathbf{b}}{\partial \mathbf{B}} = \mathbf{a} \mathbf{b}^T$$
  - $$\frac{\partial \mathbf{a}^T \mathbf{A}^{-1} \mathbf{c}}{\partial \mathbf{A}} = - \left(\mathbf{A}^T \right)^{-1} \mathbf{a} \mathbf{c}^T \left(\mathbf{A}^T \right)^{-1}$$
  - $$\frac{\partial \log |\mathbf{A}|}{\partial \mathbf{A}} = \left(\mathbf{A}^T \right)^{-1}$$

Matrix Calculus

- Let $\mathbf{B}, \mathbf{X} \in \mathbb{R}^{n \times n}$:
  - $$\frac{\partial}{\partial \mathbf{X}} \text{tr}(\mathbf{B}\mathbf{X}\mathbf{X}^T) = \mathbf{B}\mathbf{X} + \mathbf{B}^T\mathbf{X}$$
  - $$\frac{\partial}{\partial \mathbf{X}} \text{tr}(\mathbf{B}\mathbf{X}^T\mathbf{X}) = \mathbf{X}\mathbf{B}^T + \mathbf{X}\mathbf{B}$$

Multivariate Taylor Series

- For small changes, $\boldsymbol{\delta}$, about a point $\mathbf{x} = \widetilde{\mathbf{x}}$, for a scalar function of a vector argument which is at least twice differentiable:

$$f(\widetilde{\mathbf{x}} + \boldsymbol{\delta}) \approx f(\widetilde{\mathbf{x}}) + \boldsymbol{\delta} \cdot \nabla_{\mathbf{x}} f(\widetilde{\mathbf{x}}) + \frac{1}{2}\boldsymbol{\delta}^T \mathcal{H}(\widetilde{\mathbf{x}})\boldsymbol{\delta}$$

Lecture Overview

Extrema

- For a function $f : \mathbb{R}^n \to \mathbb{R}$ and a **feasible set** $X \subseteq \mathbb{R}^n$ over which we are interested in optimising, then:
  - A point $\mathbf{x}$ is a **local minimum** (or **local maximum**) of f if $f(\mathbf{x}) \leq f(\mathbf{y})$ (or $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y}$ in some neighbourhood, $N \subseteq X$ about $\mathbf{x}$
  - If $f(\mathbf{x}) \leq f(\mathbf{y})$ (or $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y} \in X$ then $\mathbf{x}$ is a **global minimum** (or **global maximum**) of f in X

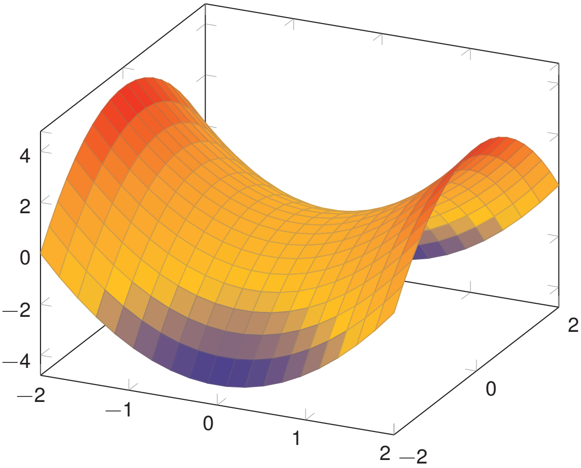
Stationary Points & Saddle Points

- Points where the gradient vanishes, i.e. where $\nabla_{\mathbf{x}} f = \mathbf{0}$, are **stationary points**, also known as **critical points**.
- It can be proved that a *necessary* condition for a point to be a maximum or minimum is that the point is stationary
- However this is not a *sufficient* condition, and points for which $\nabla_{\mathbf{x}} f = \mathbf{0}$ but where there is no local maximum or minimum are called **saddle points**

  For example:

  - $f(x_1, x_2) = {x_1}^2 - {x_2}^2 \qquad \Longrightarrow \nabla_{\mathbf{x}} f = [2x_1, -2x_2]^T$
  - $\mathbf{x} = [x_1, x_2]^T = \mathbf{0} \qquad \Longrightarrow \nabla_{\mathbf{x}} f = \mathbf{0}$
  - But at this point we have a minimum in the x_1 direction and a maximum in the x_2 direction

Saddle Points: Example



Further Conditions for Local Extrema

- Recall that by Taylor's theorem, for sufficiently small $\boldsymbol{\delta}$, and twice differentiable f about $\mathbf{x}^*$:

$$f(\mathbf{x}^* + \boldsymbol{\delta}) \approx f(\mathbf{x}^*) + \boldsymbol{\delta} \cdot \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \frac{1}{2}\boldsymbol{\delta}^T \mathcal{H}(\mathbf{x}^*)\boldsymbol{\delta}$$

- If we are at a point for which $\nabla_{\mathbf{x}} f(\mathbf{x}^*) = \mathbf{0}$, then:
  - If $\mathcal{H}(\mathbf{x}^*) \succ 0$ then $\mathbf{x}^*$ is a **local minimum**
  - If $\mathcal{H}(\mathbf{x}^*) \prec 0$ then $\mathbf{x}^*$ is a **local maximum**
  - If $\mathcal{H}(\mathbf{x}^*)$ is indefinite then $\mathbf{x}^*$ is a **saddle point**
  - Otherwise we need to investigate things further...

Conditions for Global Extrema

- These are harder to state, however a class of functions for which it is more straightforward to discern global extrema are twice differentiable **convex** functions
- These are functions where $\nabla^2_{\mathbf{x}} f \succeq 0$ globally
- Many of the learning tasks which we will perform during this module will involve optimisation over convex functions

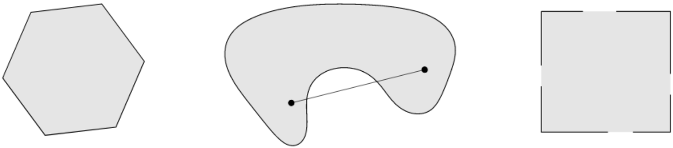
Lecture Overview

Convex Sets[2]

- **Definition:**

  A set Ω is **convex** if, for any $\mathbf{x}, \mathbf{y} \in \Omega$ and $\theta \in [0, 1]$, then $\theta \mathbf{x} + (1 - \theta)\mathbf{y} \in \Omega$.

  In other words, if we take any two elements in Ω, and draw a line segment between these two elements, then every point on that line segment also belongs to Ω.



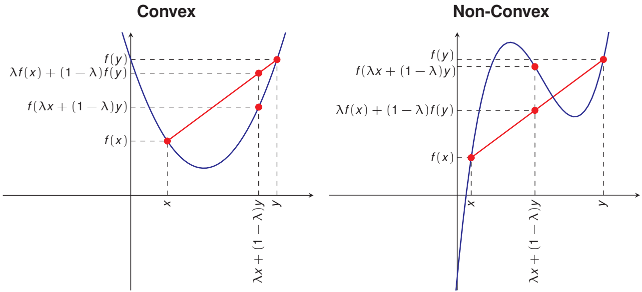
[2] Boyd & Vandenberghe, 'Convex Optimisation' [2004]

Convex Functions

- **Definition:**

  A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if its domain is a **convex set** and if, for all $\mathbf{x}, \mathbf{y}$ in its domain, and all $\lambda \in [0, 1]$, we have:

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$



Concave Functions

- **Definition:**

  A function $f : \mathbb{R}^n \to \mathbb{R}$ is **concave** if $-f$ is convex

1st & 2nd Order Characterisations of Convex, Differentiable Functions

- **Theorem A.1:**

  Suppose f is twice differentiable over an open domain. Then, the following are equivalent:

  - f is convex
  - $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \qquad \forall \quad \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
  - $\nabla^2_{\mathbf{x}} f(\mathbf{x}) \succeq 0 \qquad \forall \quad \mathbf{x} \in \mathrm{dom}(f)$

Global Optimality

- **Theorem A.2:**

  Consider an unconstrained optimisation problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to:} \quad & \mathbf{x} \in \mathbb{R}^n \end{aligned}$$

  If $f : \mathbb{R}^n \to \mathbb{R}$ is **convex**, then any point that is **locally optimal** is **globally optimal**

  Furthermore, if f is also **differentiable** then any point $\mathbf{x}$ that satisfies $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0}$ is a **globally optimal** solution

Strict Convexity

- **Definition:**

  A function $f : \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** if its domain is a **convex set** and if, for all $\mathbf{x}, \mathbf{y}, \mathbf{x} \neq \mathbf{y}$ in its domain, and all $\lambda \in (0, 1)$, we have:

  $$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- **First Order Characterisation:**

  A function f is **strictly convex** on $\Omega \subseteq \mathbb{R}^n$, if and only if:

  $$f(\mathbf{y}) > f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \qquad \forall \mathbf{x}, \mathbf{y} \in \Omega, \qquad \mathbf{x} \neq \mathbf{y}$$

- **Second Order Sufficient Condition:**

  A function f is **strictly convex** on $\Omega \subseteq \mathbb{R}^n$, if:

  $$\nabla^2_{\mathbf{x}} f(\mathbf{x}) \succ 0 \qquad \forall \mathbf{x} \in \Omega$$

Strict Convexity and Uniqueness of Optimal Solutions

- **Theorem A.3:**

  Consider an optimisation problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to:} \quad & \mathbf{x} \in \Omega \end{aligned}$$

  Where $f : \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** on Ω and Ω is a **convex set**.

  Then the **optimal solution** must be **unique**

Sums of Convex Functions

- If a $f(\cdot)$ is a **convex function**, and $g(\cdot)$ is a **convex function**, then:
  $\alpha f(\cdot) + \beta g(\cdot)$ is also a **convex function** if $\alpha, \beta > 0$.

- If a $f(\cdot)$ is a **convex function**, and $h(\cdot)$ is a **strictly convex function**, then:
  $\alpha f(\cdot) + \beta h(\cdot)$ is a **strictly convex function** if $\alpha, \beta > 0$.

- Proofs follow from an application of the definitions of convexity and strict convexity

Lecture Overview

Quadratic Functions

- Consider the following **quadratic function**, f:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}\cdot\mathbf{x} + c$$

  Where: $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, and the variable $\mathbf{x} \in \mathbb{R}^n$.

- From the *Linear Algebra* lecture note that, w.l.o.g., we can write:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x} + \mathbf{b}\cdot\mathbf{x} + c$$

  Where: $\mathbf{B} = \left(\mathbf{A} + \mathbf{A}^T\right)$, and is hence **symmetric**.

Convexity of Quadratic Functions

- $\mathbf{B} \succeq 0 \quad \Longleftrightarrow \quad$ **Convexity** of f
- $\mathbf{B} \succ 0 \quad \Longleftrightarrow \quad$ **Strict Convexity** of f

Convexity of Quadratic Functions

- **Proof:** *(for first result. Analogous proof holds for second result)*
  For any $0 \leq \lambda \leq 1$, and for any variables $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$
\begin{aligned}
\lambda f(\mathbf{x}) &= \lambda \frac{1}{2} \mathbf{x}^T \mathbf{B} \mathbf{x} + \lambda \mathbf{b} \cdot \mathbf{x} + \lambda c \\
(1-\lambda) f(\mathbf{y}) &= (1-\lambda) \frac{1}{2} \mathbf{y}^T \mathbf{B} \mathbf{y} + (1-\lambda) \mathbf{b} \cdot \mathbf{y} + (1-\lambda) c \\
f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) &= \frac{1}{2} \left(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\right)^T \mathbf{B} \left(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\right) + \mathbf{b} \cdot \left(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\right) + c \\
&= \frac{1}{2} \lambda^2 \mathbf{x}^T \mathbf{B} \mathbf{x} + \frac{1}{2} (1-\lambda)^2 \mathbf{y}^T \mathbf{B} \mathbf{y} + \lambda(1-\lambda) \mathbf{x}^T \mathbf{B} \mathbf{y} \\
&\qquad + \lambda \mathbf{b} \cdot \mathbf{x} + (1-\lambda) \mathbf{b} \cdot \mathbf{y} + c
\end{aligned}
$$

  Thus:

$$
\begin{aligned}
& f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) - \lambda f(\mathbf{x}) - (1-\lambda) f(\mathbf{y}) \\
&\qquad = (\lambda^2 - \lambda) \frac{1}{2} \mathbf{x}^T \mathbf{B} \mathbf{x} + \left((1-\lambda)^2 - (1-\lambda)\right) \frac{1}{2} \mathbf{y}^T \mathbf{B} \mathbf{y} + \lambda(1-\lambda) \mathbf{x}^T \mathbf{B} \mathbf{y} \\
&\qquad = \lambda(\lambda - 1) \frac{1}{2} \mathbf{x}^T \mathbf{B} \mathbf{x} + \lambda(\lambda - 1) \frac{1}{2} \mathbf{y}^T \mathbf{B} \mathbf{y} - \lambda(\lambda - 1) \mathbf{x}^T \mathbf{B} \mathbf{y} \\
&\qquad = \frac{1}{2} \lambda(\lambda - 1)(\mathbf{x} - \mathbf{y})^T \mathbf{B} (\mathbf{x} - \mathbf{y})
\end{aligned}
$$

  So: **Convexity** $\implies$ LHS ≤ 0 $\iff$ $(\mathbf{x} - \mathbf{y})^T \mathbf{B} (\mathbf{x} - \mathbf{y}) \geq 0$ $\implies$ $\mathbf{B} \succeq 0$

Optimisation of Quadratic Functions

- Consider the following **unconstrained** optimisation problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{1}$$

  Where:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x} + \mathbf{b} \cdot \mathbf{x} + c$$

  And: $\mathbf{B}$ is **real** and **symmetric**

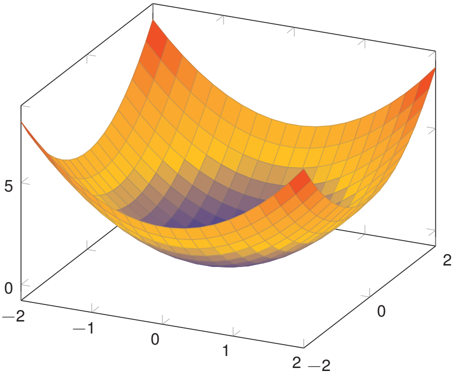
- Consider the following possible forms of f:

Strictly Convex f

- If $\mathbf{B} \succ 0$ then f is **strictly convex**
- Thus, by *Theorem A.3*, there is a **unique** solution to problem (1), $\mathbf{x}^*$:
$$\mathbf{x}^* = -\mathbf{B}^{-1}\mathbf{b}$$
- Note that this solution must exist because:
$$\mathbf{B} \succ 0 \quad \Longrightarrow \quad \det \mathbf{B} \neq 0 \quad \Longrightarrow \quad \exists\, \mathbf{B}^{-1}$$

Strictly Convex f: Example

- $f(\mathbf{x}) = x_1^2 + x_2^2$,   i.e.: $\mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \mathbf{b} = \mathbf{0}, c = 0$



Convex & Bounded f

- If $\mathbf{B} \succeq 0$, $\mathbf{B} \not\succ 0$, $\mathbf{b} \in \text{range}(\mathbf{B})$ then f is **convex** but not strictly convex, and is **bounded below**
  - **Proof:**
    Solutions to the problem involve the following condition:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{Bx} + \mathbf{b} = \mathbf{0}$$
$$\implies \mathbf{Bx} = -\mathbf{b}$$

From the *Linear Algebra* lecture, this system of equations has a solution iff it is **consistent**, i.e.:

$$\text{rank}(\mathbf{B}) = \text{rank}\left(\mathbf{B}| - \mathbf{b}\right)$$

Because $\mathbf{b} \in \text{range}(\mathbf{B})$ then $\mathbf{b}$ is some linear combination of the columns of $\mathbf{B}$, so:

$$\text{columnspace}(\mathbf{B}) = \text{columnspace}\left(\mathbf{B}| - \mathbf{b}\right)$$
$$\implies \text{range}(\mathbf{B}) = \text{range}\left(\mathbf{B}| - \mathbf{b}\right)$$
$$\implies \text{rank}(\mathbf{B}) = \text{rank}\left(\mathbf{B}| - \mathbf{b}\right)$$

Convex & Bounded f (Cont.)

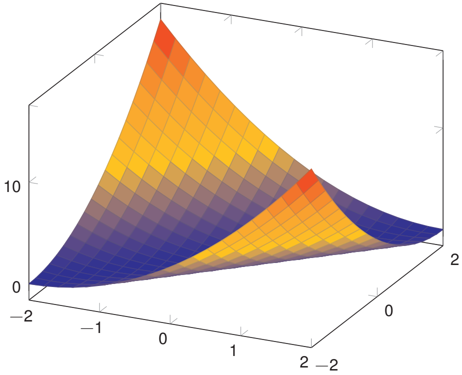
- There are **infinite** solutions to problem (1), since:

$$\begin{aligned} \mathbf{B} \succeq 0, \mathbf{B} \not\succ 0 &\implies \text{at least one eigenvalue} = 0 \\ &\implies \det \mathbf{B} = 0 \\ &\implies \mathbf{B} \text{ is not full rank} \end{aligned}$$

Thus the system of equations, $\mathbf{Bx} = -\mathbf{b}$, is **underdetermined**

Convex & Bounded f: Example

- $f(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1x_2$, i.e.: $\mathbf{B} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}, \mathbf{b} = \mathbf{0}, c = 0$



Convex & Unbounded f

- If $\mathbf{B} \succeq 0$, $\mathbf{B} \nsucc 0$, $\mathbf{b} \notin \text{range}(\mathbf{B})$ then f is **convex** but not strictly convex, and is **unbounded below**
  - **Proof:**
    As before solutions to the problem exist iff:

$$\text{rank}(\mathbf{B}) = \text{rank}\left(\mathbf{B}| - \mathbf{b}\right)$$

    But because $\mathbf{b} \notin \text{range}(\mathbf{B})$ then $\mathbf{b}$ cannot be written as some linear combination of the columns of $\mathbf{B}$, so:

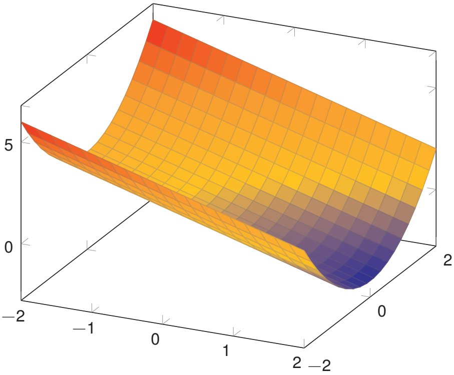
$$\begin{aligned} &\text{columnspace}(\mathbf{B}) \neq \text{columnspace}\left(\mathbf{B}| - \mathbf{b}\right) \\ \implies &\text{range}(\mathbf{B}) \neq \text{range}\left(\mathbf{B}| - \mathbf{b}\right) \\ \implies &\text{rank}(\mathbf{B}) \neq \text{rank}\left(\mathbf{B}| - \mathbf{b}\right) \end{aligned}$$

Convex & Unbounded f (Cont.)

- Thus the system of equations is **inconsistent** and **no solutions** to problem (1) exist

Convex & Unbounded f: Example

- $f(\mathbf{x}) = x_2^2 - x_1$, i.e.: $\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, c = 0$



Non-Convex f

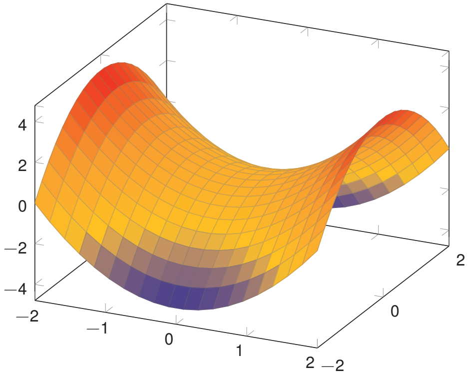
- If $\mathbf{B} \nsucceq 0$ then f is **non-convex**
- There is **no solution** to problem (1) since f is **unbounded below**
  - **Proof:**
    Consider an eigenvector of $\mathbf{B}$, $\widetilde{\mathbf{x}}$, with an eigenvalue, $\widetilde{\lambda} < 0$:

$$\begin{aligned}
&\implies \mathbf{B}\widetilde{\mathbf{x}} = \widetilde{\lambda}\widetilde{\mathbf{x}} \\
&\implies \widetilde{\mathbf{x}}^T\mathbf{B}\widetilde{\mathbf{x}} = \widetilde{\lambda}\widetilde{\mathbf{x}}^T\widetilde{\mathbf{x}} < 0 \\
&\implies f(\alpha\widetilde{\mathbf{x}}) = \frac{1}{2}\alpha^2\widetilde{\lambda}\widetilde{\mathbf{x}}^T\widetilde{\mathbf{x}} + \alpha\mathbf{b}\cdot\widetilde{\mathbf{x}} + c
\end{aligned}$$

    Thus $f(\alpha\widetilde{\mathbf{x}}) \to -\infty$ as $\alpha \to \infty$

Non-Convex f: Example

- $f(\mathbf{x}) = x_1^2 - x_2^2$, i.e.: $\mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}, \mathbf{b} = \mathbf{0}, c = 0$



Optimisation of Quadratic Functions

- Characteristic Properties of:

$$\operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x})$$

$$\text{where: } f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x} + \mathbf{b}\cdot\mathbf{x} + c$$

Parameters of f	f	$\text{argmin}_{\mathbf{x}} f(\mathbf{x})$	
$\mathbf{B} \succ 0$	**Strictly Convex**	1 solution	$(\mathbf{B}\mathbf{x} = -\mathbf{b})$ consistent & exactly deter
$\mathbf{B} \succeq 0$, $\mathbf{B} \not\succ 0$, $\mathbf{b} \in \text{range } \mathbf{B}$	**Convex & Bounded Below**	∞ solutions	$(\mathbf{B}\mathbf{x} = -\mathbf{b})$ consistent & underdetermi
$\mathbf{B} \succeq 0$, $\mathbf{B} \not\succ 0$, $\mathbf{b} \notin \text{range } \mathbf{B}$	**Convex & Unbounded Below**	0 solutions	$(\mathbf{B}\mathbf{x} = -\mathbf{b})$ inconsistent
$\mathbf{B} \not\succeq 0$	**Non-Convex**	0 solutions	f unbounded below

Example: Ordinary Least Squares

- Consider:

$$\operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w})$$

where:

$$\begin{aligned} f(\mathbf{w}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - (\mathbf{X}^T\mathbf{y}) \cdot \mathbf{w} + \frac{1}{2}\|\mathbf{y}\|_2^2 \end{aligned}$$

- So:

$$\begin{aligned} \mathbf{w} &\longleftarrow \mathbf{x} \\ \mathbf{X}^T\mathbf{X} &\longleftarrow \mathbf{B} \\ -\mathbf{X}^T\mathbf{y} &\longleftarrow \mathbf{b} \\ \frac{1}{2}\|\mathbf{y}\|_2^2 &\longleftarrow c \end{aligned}$$

Example: Ordinary Least Squares

- Also, recall from the *Linear Algebra* lecture:

$$
\begin{aligned}
& \operatorname{rank}(\mathbf{X}^T\mathbf{X}|\mathbf{X}^T\mathbf{y}) = \operatorname{rank}(\mathbf{X}^T\mathbf{X}) \\
\implies & \qquad \mathbf{X}^T\mathbf{y} \in \operatorname{range}(\mathbf{X}^T\mathbf{X}) \\
\implies & \qquad -\mathbf{X}^T\mathbf{y} \in \operatorname{range}(\mathbf{X}^T\mathbf{X})
\end{aligned}
$$

- And:

$$\mathbf{X}^T\mathbf{X} \succeq 0$$

Example: Ordinary Least Squares

- Thus the OLS problem always has at least one solution:
  - Either:

$$\mathbf{X}^T\mathbf{X} \succ 0 \qquad \Longrightarrow \qquad 1 \text{ solution}$$

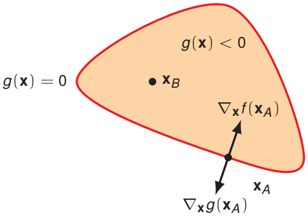
  - Or:

$$\left.\begin{aligned} \mathbf{X}^T\mathbf{X} &\succeq 0 \\ \mathbf{X}^T\mathbf{X} &\nsucc 0 \\ -\mathbf{X}^T\mathbf{y} \in \operatorname{range}(\mathbf{X}^T\mathbf{X}) \end{aligned}\right\} \qquad \Longrightarrow \qquad \infty \text{ solutions}$$

Lecture Overview

Notation[3]

- $\mathbf{x} \in \mathbb{R}^n$
- $f : \mathbb{R}^n \to \mathbb{R}$ is the function over which we wish to optimise $\mathbf{x}$
- $g(\mathbf{x}) = 0$ represents an $(n-1)$ dimensional surface constraint
- $n = 2$ dimensional illustration (with $g(\mathbf{x}_A) = 0$, and $g(\mathbf{x}_B) < 0$):



[3] Content and illustrations based on Bishop, 'Pattern Recognition & Machine Learning' [2008]

Equality Constraints: Problem

- $$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to:} \quad & g(\mathbf{x}) = 0 \end{aligned}$$

  Note that the functions f and g can be convex or nonconvex in general.

Equality Constraints: Observations

- $\nabla_{\mathbf{x}} g(\mathbf{x})$ is orthogonal to the surface defined by $g(\mathbf{x})$:
  - Because, if we denote any direction along the surface $g(\mathbf{x})$ by $\widehat{\mathbf{u}}$, then because the directional derivative along the direction of the surface must be zero 0, $\nabla_{\mathbf{x}} g(\mathbf{x}) \cdot \widehat{\mathbf{u}} = 0$.
- The optimal point, $\mathbf{x}^*$ must have the property that $\nabla_{\mathbf{x}} f(\mathbf{x}^*)$ is orthogonal to the constraint surface:
  - Because, otherwise $f(\mathbf{x})$ could decrease for movements along the surface.

Equality Constraints: Lagrange Multiplier

- Thus $\nabla_{\mathbf{x}} f(\mathbf{x})$ and $\nabla_{\mathbf{x}} g(\mathbf{x})$ must be parallel, i.e.:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \lambda \nabla_{\mathbf{x}} g(\mathbf{x}) = 0 \quad \text{for some:} \quad \lambda \neq 0$$

  Here λ is a so-called **Lagrange multiplier**

Equality Constraints: Lagrangian

- Let us define the **Lagrangian** function, $\mathcal{L}$, as follows:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

- Then:

$$\begin{aligned}
\nabla_{\mathbf{x}}\mathcal{L} &= \nabla_{\mathbf{x}} f(\mathbf{x}) + \lambda \nabla_{\mathbf{x}} g(\mathbf{x}) \\
\nabla_{\lambda}\mathcal{L} &= g(\mathbf{x})
\end{aligned}$$

Equality Constraints: Problem reformulation

- Seek stationary solutions $(\mathbf{x}^*, \lambda^*)$ which satisfy the following:

$$\nabla_{\mathbf{x}}\mathcal{L} = \mathbf{0}$$
$$\nabla_{\lambda}\mathcal{L} = 0$$

- Thus we have transformed our problem into an **unconstrained** optimisation problem
- Furthermore, we have re-phrased this problem as a well posed one involving the solution of a set of simultaneous equations

Equality Constraints: Problem reformulation

- Note that these conditions characterise a stationary point associated with the function f...
- ...But we have said nothing about whether such a point is a maximum, a minimum or a saddle point
- It can be proved that if both f and g are **convex** functions then the stationary point is a **minimum**
- And if f is **concave** and g is **convex** then the stationary point is a **maximum**

Equality Constraints: Saddle Point

- Note that the optimal solution to our constrained optimisation problem will be a **saddle point**...
- Consider the Hessian matrix for the Lagrangian:

$$\mathcal{H}(\mathbf{x}, \lambda) = \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}) + \lambda \nabla^2_{\mathbf{x}} g(\mathbf{x}) & \nabla_{\mathbf{x}} g(\mathbf{x}) \\ \nabla_{\mathbf{x}} g(\mathbf{x})^T & 0 \end{bmatrix}$$

- Now consider the quadratic form $\boldsymbol{\alpha}^T \mathcal{H}(\mathbf{x}, \lambda) \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\mathbf{a}, b]^T$, for all $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$:

$$\boldsymbol{\alpha}^T \mathcal{H}(\mathbf{x}, \lambda) \boldsymbol{\alpha} = \mathbf{a}^T \left(\nabla^2_{\mathbf{x}} f(\mathbf{x}) + \lambda g(\mathbf{x})^2 g(\mathbf{x}) \right) \mathbf{a} + 2b\mathbf{a} \cdot \nabla_{\mathbf{x}} g(\mathbf{x})$$

- Clearly if $\nabla_{\mathbf{x}} g(\mathbf{x})$ is finite then it is always possible to select $\mathbf{a}, b$ such that the second term dominates the first in magnitude and can be made either positive or negative.

Equality Constraints: Saddle Point

- Thus $\mathcal{H}(\mathbf{x}, \lambda)$ is **indefinite** and the stationary points for $\mathcal{L}(\mathbf{x}, \lambda)$ are thus saddle points

- Note that this makes the use of **gradient descent** as an optimisation procedure somewhat problematic.
  But alternatives numerical procedures exist

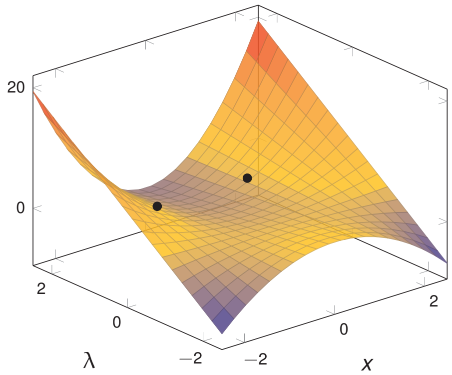
Equality Constraints: Example

- Let: $f(x) = x^2$ and $g(x) = (x^2 - 1)$
  Then:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad x^2$$
$$\text{subject to:} \quad x^2 = 1$$

- Critical points: $(x, \lambda) = (1, -1)$ and $(-1, -1)$

Equality Constraints: Example



Inequality Constraints: Problem

-

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to:} \quad & g(\mathbf{x}) \leq 0 \end{aligned}$$

- Two types of solution are possible:

Inequality Constraints: Inactive Constraint

- $\mathbf{x}^*$ lies in $g(\mathbf{x}) < 0$
- Stationary condition $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0}$
- Which is equivalent to:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0} \quad \text{with:} \quad \lambda = 0 \tag{2}$$

Inequality Constraints: Active Constraint

- $\mathbf{x}^*$ lies on $g(\mathbf{x}) = 0$
- Since the solution does not lie in $g(\mathbf{x}) < 0$ then $f(\mathbf{x})$ will only be minimal if $\nabla_{\mathbf{x}} f(\mathbf{x})$ points towards the $g(\mathbf{x}) < 0$ region. Thus:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = -\lambda \nabla_{\mathbf{x}} g(\mathbf{x}) \quad \text{for} \quad \lambda > 0$$

- Which is equivalent to:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0} \quad \text{with:} \quad \lambda > 0 \tag{3}$$

Inequality Constraints: Problem Reformulation

- Using equations (2) & (3), we can solve our problem by seeking stationary solutions $(\mathbf{x}^*, \lambda^*)$ which satisfy the following:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0}$$

$$\text{subject to:} \quad \begin{cases} g(\mathbf{x}) \leq 0 \\ \lambda \geq 0 \\ \lambda g(\mathbf{x}) = 0 \end{cases}$$

- These conditions are known as the **Karush Kuhn Tucker** (KKT) conditions.

Inequality Constraints: Complementary Slackness

- $\lambda g(\mathbf{x}) = 0$ is satisfied for both the **active** and **inactive cases**, and is known as the **complementary slackness** condition

- It is equivalent to:

$$\begin{aligned} \lambda > 0 \qquad &\Longrightarrow \qquad g(\mathbf{x}) = 0 \\ g(\mathbf{x}) < 0 \qquad &\Longrightarrow \qquad \lambda = 0 \end{aligned}$$

And, rarely, when the critical point associated with $f(\mathbf{x})$ coincides with the constraint surface:

$$\lambda = 0 \qquad \text{and} \qquad g(\mathbf{x}) = 0$$

Inequality Constraints: Problem reformulation

- Again, note that these conditions characterise a stationary point associated with the function f...
- ...But we have said nothing about whether such a point is a maximum, a minimum or a saddle point
- It can be proved that if both f and g are **convex** functions then the stationary point is a **minimum**
- Otherwise the stationary point is a **maximum**, a **minimum** or a **saddle point**
- (But note that, regardless, the KKT conditions hold)

Multiple Constraints: Problem

-

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x})$$

$$\text{subject to:} \quad \begin{cases} \{g^{(i)}(\mathbf{x}) \leq 0\}_{i=1}^{m} \\ \{h^{(j)}(\mathbf{x}) = 0\}_{j=1}^{p} \end{cases}$$

Multiple Constraints: Lagrangian

- We express the Lagrangian as:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x})$$

Where:
$\boldsymbol{\lambda} = [\lambda^{(1)}, ..., \lambda^{(p)}]^T, \{\lambda^{(j)} \in \mathbb{R}\}_{j=1}^{p};$
$\boldsymbol{\mu} = [\mu^{(1)}, ..., \mu^{(m)}]^T, \{\mu^{(i)} \in \mathbb{R}^{\geq 0}\}_{i=1}^{m};$
are Lagrange multipliers

Multiple Constraints: Problem Reformulation

- And we can solve our problem by seeking stationary solutions $(\mathbf{x}^*, \{\mu^{(i)*}\}, \{\lambda^{(j)*}\})$ which satisfy the following:

$$\nabla_{\mathbf{x}}\mathcal{L} = \mathbf{0}$$

$$\text{subject to:} \quad \begin{cases} \{g^{(i)}(\mathbf{x}) \leq 0\}_{i=1}^{m}, \{h^{(j)}(\mathbf{x}) = 0\}_{j=1}^{p} \\ \{\mu^{(i)} \geq 0\}_{i=1}^{m} \\ \{\mu^{(i)} g^{(i)}(\mathbf{x}) = 0\}_{i=1}^{m} \end{cases}$$

Duality

- It's not immediately obvious what the value of this reformulation is. We seem to have replaced one constrained optimisation problem with another...

- But actually we have trasnformed a rather opaque optimisation problem into a more familiar problem - a constrained set of simultaneous equations:

$$\nabla_{\mathbf{x}}\mathcal{L} = \mathbf{0} \qquad \text{and} \qquad \{\mu^{(i)} g^{(i)}(\mathbf{x}) = 0\}_{i=1}^{m}$$

- But there are other advantages of the Lagrangian approach, for which we will consider the concept of **duality**

Duality: Primal Problem

- The original problem is sometimes know as the **primal problem**, and its variables, $\mathbf{x}$, are known as the **primal variables**

- It is equivalent to the following formulation:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- Here the bracketed term is known as the **primal objective** function

Duality: Barrier Function

- We can re-write the primal objective as follows:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq 0} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\mathbf{x}) + \max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq 0} \left[\sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x}) \right]$$

- Here the second term gives rise to a **barrier function** which enforces the constraints as follows:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq 0} \left[\sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x}) \right] = \begin{cases} 0 & \text{if } \mathbf{x} \text{ is feasible} \\ \infty & \text{if } \mathbf{x} \text{ is infeasible} \end{cases}$$

Duality: Minimax Inequality

- In order to make use of this barrier function formulation, we will need the **minimax inequality**:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$$

- **Proof:**

$$\min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leq \phi(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y}$$

  This is true for all $\mathbf{y}$, therefore, in particular the following is true:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leq \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}$$

  This is true for all $\mathbf{x}$, therefore, in particular the following is true:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$$

Duality: Weak Duality

- We can now introduce the concept of **weak duality**:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda},\boldsymbol{\mu} \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right] \geq \max_{\boldsymbol{\lambda},\boldsymbol{\mu} \geq 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- Here the bracketed term on the right hand side is known as the **dual objective** function, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$
- If we can solve the right hand side of the inequality then we have a lower bound on the solution of our optimisation problem

Duality: Weak Duality

- And sometimes the RHS side of the inequality is an **easier** problem to solve:
  - $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is an **unconstrained** optimisation problem for a given value of $(\boldsymbol{\lambda}, \boldsymbol{\mu})$...
  - ...And if solving this problem is not hard then the overall problem is not hard to solve because:
  - $\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0} [\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})]$ is a maximisation problem over a set of affine functions - thus it is a **concave maximisation** problem or equivalently a **convex minimisation** problem, and we know that such problems can be efficiently solved
  - Note that this is true regardless of whether $f, g^{(i)}, h^{(j)}$ are nonconvex

Duality: Strong Duality

- For certain classes of problems which satisfy **constraint qualifications** we can go further and **strong duality** holds:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda},\boldsymbol{\mu} \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right] = \max_{\boldsymbol{\lambda},\boldsymbol{\mu} \geq 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- There are several different constraint qualifications. One is **Slater's Condition** which holds for **convex optimisation** problems
- Recall, these are problems for which f is convex and $g^{(i)}, h^{(j)}$ are convex sets
- For problems of this type we may seek to solve the **dual optimisation** problem:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu} \geq 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

Duality: Strong Duality

- Once again, note that the dual optimisation problem is sometimes easier to solve than the primal problem
- But, for our purposes, another interesting reason for adopting the dual optimisation approach to solving contrained optimisation problems is based on dimensionality:
- If the dimensionality of the dual variables, $(m + p)$, is less than the dimensionality of the primal variables, n, then dual optimisation often offers a more efficient route to solutions
- This is of particular importance if we are dealing with infinite dimensional primal variables

Linear Programming

- The following canonical optimisation problem is known as a **linear programme**:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to:} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \end{aligned}$$

Where:
$\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^n$.

Linear Programming

- The Lagrangian is given by:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}) &= \mathbf{c}^T\mathbf{x} + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= (\mathbf{c} + \mathbf{A}^T\boldsymbol{\mu})^T\mathbf{x} - \boldsymbol{\mu}^T\mathbf{b}\end{aligned}$$

- Taking derivatives and seeking stationarity:

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{c} + \mathbf{A}^T\boldsymbol{\mu} = \mathbf{0}$$

Linear Programming

- From which we generate the dual Lagrangian:

$$\mathcal{D}(\boldsymbol{\mu}) = -\boldsymbol{\mu}^T \mathbf{b}$$

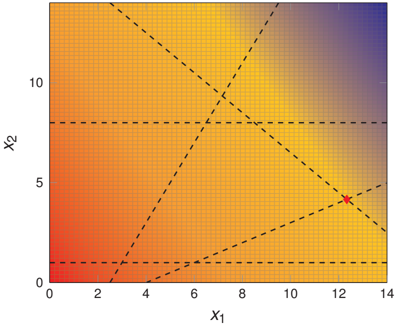
- Thus, the dual optimisation problem is:

$$\begin{aligned} \max_{\boldsymbol{\mu} \in \mathbb{R}^m} \quad & -\boldsymbol{\mu}^T \mathbf{b} \\ \text{subject to:} \quad & \mathbf{c} + \mathbf{A}^T \boldsymbol{\mu} = \mathbf{0} \\ \text{subject to:} \quad & \boldsymbol{\mu} \geq \mathbf{0} \end{aligned}$$

Linear Programming: Example[4]

- Let:

$$\mathbf{c} = -\begin{bmatrix}5\\3\end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix}2 & 2\\2 & -4\\-2 & 1\\0 & -1\\0 & 1\end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix}33\\8\\5\\-1\\8\end{bmatrix}$$



4 Example based on Deisenroth et al, 'Mathematics For Machine Learning' [2020]

Quadratic Programming

- The following canonical optimisation problem is known as a **quadratic programme**:

$$
\begin{aligned}
\min_{\mathbf{x} \in \mathbb{R}^n} \quad & \frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \\
\text{subject to:} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}
\end{aligned}
$$

Where:
$\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^n$.
$\mathbf{Q} \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, therefore the objective function is **strictly convex**.

Quadratic Programming

- The Lagrangian is given by:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

- Taking derivatives and seeking stationarity:

$$\begin{aligned} & \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{Q}\mathbf{x} + (\mathbf{c} + \mathbf{A}^T\boldsymbol{\mu}) = \mathbf{0} \\ \Longrightarrow \qquad & \mathbf{x} = -\mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\boldsymbol{\mu}) \end{aligned}$$

Quadratic Programming

- From which we generate the dual Lagrangian:

$$\mathcal{D}(\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{c} + \mathbf{A}^T\mu)^T\mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\mu) - \boldsymbol{\mu}^T\mathbf{b}$$

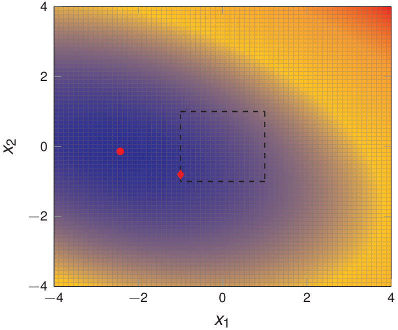
- Thus, the dual optimisation problem is:

$$\begin{aligned} \max_{\boldsymbol{\mu}\in\mathbb{R}^m} \quad & -\frac{1}{2}(\mathbf{c} + \mathbf{A}^T\mu)^T\mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\mu) - \boldsymbol{\mu}^T\mathbf{b} \\ \text{subject to:} \quad & \boldsymbol{\mu} \geq \mathbf{0} \end{aligned}$$

Quadratic Programming: Example[5]

- Let:

$$\mathbf{Q} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}; \quad \mathbf{c} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$



[5] Example based on Deisenroth et al, 'Mathematics For Machine Learning' [2020]

Lecture Overview

Integral Calculus

- Calculus can be split into two fields:
  - Differential calculus, which, as we have seen, is concerned with the instantaneous rate of change of functions
  - **Integral calculus**, which, as we will see, is concerned with the accumulation of quantities.
- We will see that integral calculus, and the operation of integration, allows us to find the area (or volume) lying beneath functions.
  - This will be crucial when we seek to make a **probabilistic** analysis of machine learning problems.

Indefinite Integral

- The **indefinite integral** or **antiderivative** of a continuous function, $f(x)$, is a function $F(x)$, written as $\int f(x)dx$, the derivative of which is f.
  Thus:
  $$F'(x) = f(x)$$

Common Indefinite Integrals

$f(x)$	$F(x) = \int f(x)dx$
$x^n \ (n \neq 1)$	$\frac{x^{n+1}}{n+1} + C$
$\frac{1}{x}$	$\ln x + C$
e^x	$e^x + C$

- Where C is an arbitrary **constant of integration**.

Definite Integral

- The **definite integral** of a continuous function, $f(x)$, between two numbers, a and b, written as $\int_a^b f(x)dx$, is defined to be the difference between $F(a)$ and $F(b)$.
  Thus:
  $$\int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a)$$

The Fundamental Theorem of Calculus

- If we divide the interval $[a, b]$ into N equal subintervals, each of length $\delta = \frac{(b-a)}{N}$, then we may seek to evaluate the **Reimann sum**:

$$\sum_{i=1}^{N} f(x_i)\, \delta$$

- This is the sum of a series of N rectangles each of which has a different height given by the values in the set $\{x_i\}_{i=1}^{N}$, but with the same width, δ.

The Fundamental Theorem of Calculus

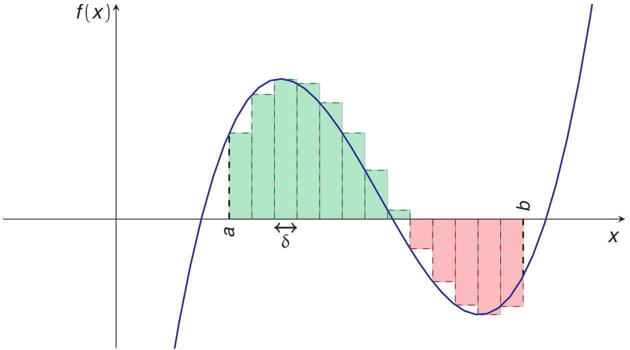
- Now, the **Fundamental Theorem of Calculus** tells us that as δ tends to zero then the Reimann sum tends to the definite integral:

$$\lim_{\delta \to 0} \sum_{i=1}^{N} f(x_i)\,\delta = \int_a^b f(x)dx$$

- So, this theorem connects the objects of integration which we have discussed with the notion of '(signed) area under the curve'.

The Fundamental Theorem of Calculus

$$\int_a^b f(x)dx = \lim_{\delta \to 0} \sum_{i=1}^{N} f(x_i)\ \delta$$



- The area of green region adds to the total of the indefinite integral, while the area of the red region subtracts from it.

Multiple Integrals: Example

- This notion generalises to higher dimensions, so that in n dimensions, where we wish to integrate a function $f(\mathbf{x})$ across a more general region of integration, we may denote the multiple integral by:

$$\int \cdots \int_V f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n = \int_V f(\mathbf{x}) d\mathbf{x}$$

- And this is equivalent to the signed hypervolume under the hypersurface $f(\mathbf{x})$, bounded by the region delineated by the region $V \subseteq \mathbb{R}^n$.

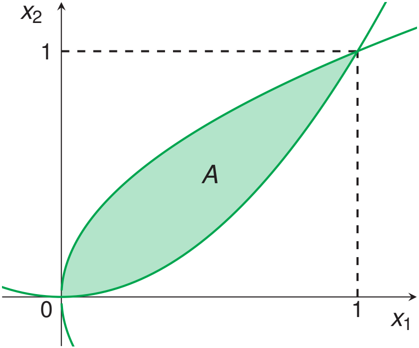
Multiple Integrals: Example

- We wish to integrate the function $g(\mathbf{x}) = x_1^2 + x_2^2$ over the region, A, bounded by $x_2 = \sqrt{x_1}$ and $x_2 = x_1^2$:

$$\iint_A (x_1^2 + x_2^2) dx_1 dx_2$$

Multiple Integrals: Example

- The region A, over which the integration takes place, is illustrated below:



Multiple Integrals: Example

-

$$
\begin{aligned}
\iint_A (x_1^2 + x_2^2) dx_1 dx_2 &= \int_0^1 \int_{x_1^2}^{\sqrt{x_1}} (x_1^2 + x_2^2) dx_1 dx_2 \\
&= \int_0^1 \left(x_1^2 x_2 + \frac{x_2^3}{3} \right) \Big|_{x_1^2}^{\sqrt{x_1}} dx_1 \\
&= \int_0^1 \left(x_1^{\frac{5}{2}} + \frac{1}{3} x_1^{\frac{3}{2}} - x_1^4 - \frac{1}{3} x_1^6 \right) dx_1 \\
&= \left(\frac{2}{7} x_1^{\frac{7}{2}} + \frac{2}{15} x_1^{\frac{5}{2}} - \frac{1}{5} x_1^5 - \frac{1}{21} x_1^7 \right) \Big|_0^1 \\
&= \left(\frac{2}{7} + \frac{2}{15} - \frac{1}{5} - \frac{1}{21} \right) \\
&= 0.171
\end{aligned}
$$

Lecture Overview

Summary

- **Calculus** is an essential tool that helps us to minimise certain (cost) functions
- We have introduced the basic machinery for calculus in one and many dimensions
- Building on this we have introduced some techniques for unconstrained and constrained optimisation that will be of direct use in machine learning