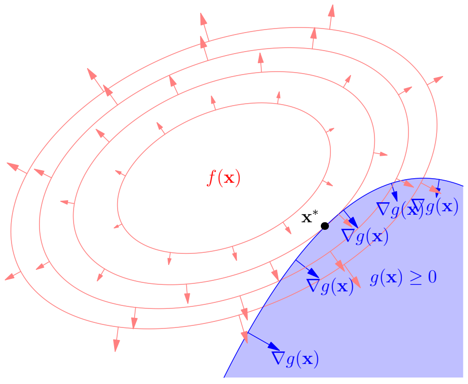
Advanced Machine Learning

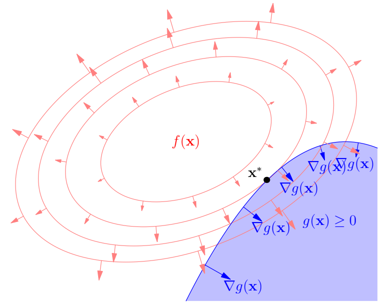
Constrained Optimisation



Lagrangians, Inequalities, KKT, Linear Programming, Quadratic Programming, Duality

Outline

1. **Constrained Optimisation**
2. Inequalities
3. Duality



Optimisation with Constraints

- There are a number of important applications where we wish to minimise an objective function subject to inequality constraints
- A prominent example of this is support vector machines
- More generally there are a large number of kernel models that involve constraints
- However, constraints are ubiquitous in machine learning (e.g. in Wasserstein GANs)

Solving Constrained Optimisation Problems

- Suppose we have a problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } g(\boldsymbol{x}) = 0$$

- A standard procedure is to define the Lagrangian

$$\mathcal{L}(\boldsymbol{x},\alpha) = f(\boldsymbol{x}) - \alpha\, g(\boldsymbol{x})$$

  where α is known as a Lagrange multiplier

- In the extended space $(\boldsymbol{x},\alpha)$ we have to solve

$$\max_{\alpha} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\alpha)$$

Conditions on Optimum

- The optimisation problem is

$$\max_{\alpha} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\alpha) \quad \text{where} \quad \mathcal{L}(\boldsymbol{x},\alpha) = f(\boldsymbol{x}) - \alpha g(\boldsymbol{x})$$

- Assuming differentiability

$$\boldsymbol{\nabla}_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\alpha) = \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) - \alpha \boldsymbol{\nabla}_{\boldsymbol{x}} g(\boldsymbol{x}) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = -g(\boldsymbol{x}) = 0$$

- The second condition is just the constraint
- But what about first condition: $\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) = \alpha \boldsymbol{\nabla}_{\boldsymbol{x}} g(\boldsymbol{x})$?

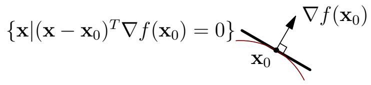
Note on Gradients

- Note that for any function $f(\boldsymbol{x})$ we can Taylor expand around $\boldsymbol{x}_0$

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{T}} \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{T}} \mathbf{H}(\boldsymbol{x} - \boldsymbol{x}_0) + \ldots$$

where $\mathbf{H}$ is a matrix of second derivative known as the Hessian

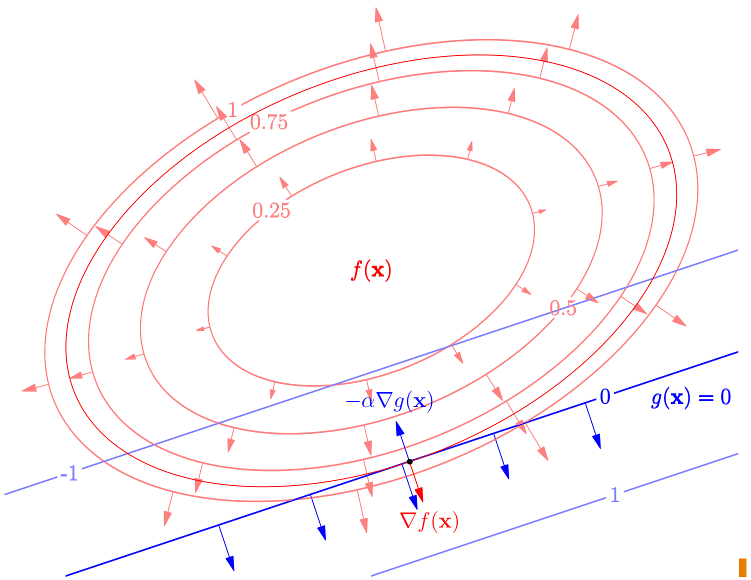
- If we consider the set of points perpendicular to $\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}_0)$ which go through $\boldsymbol{x}_0$ (the tangent plane), these will have values



$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + O(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2)$$

thus $\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x})$ is always orthogonal to the contour lines

Constrained Optima



Example

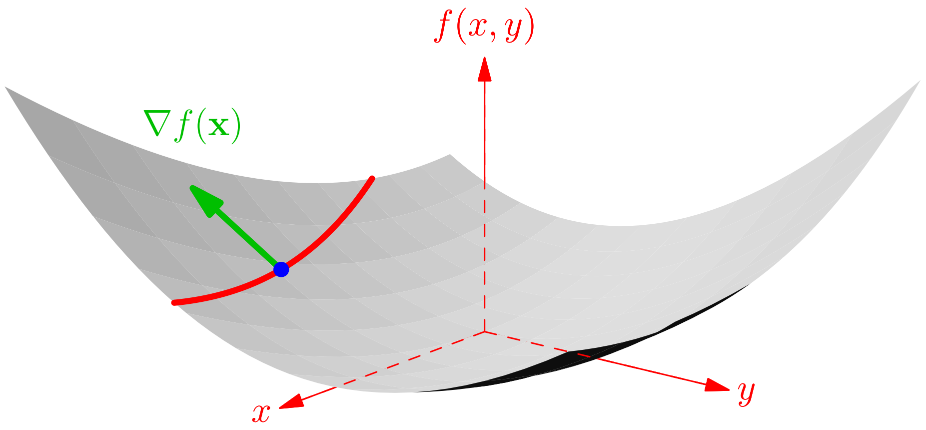
- Minimise $f(\boldsymbol{x}) = x^2 + 2y^2 - xy$
- Subject to $g(\boldsymbol{x}) = x - 2y - 3 = 0$
- Writing $\mathcal{L} = f(\boldsymbol{x}) - \alpha g(\boldsymbol{x})$
- Condition for minima is $\boldsymbol{\nabla}_{\boldsymbol{x}} \mathcal{L} = 0$

$$\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} 2x - y \\ -x + 4y \end{pmatrix} = \alpha \boldsymbol{\nabla}_{\boldsymbol{x}} g(\boldsymbol{x}) = \alpha \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

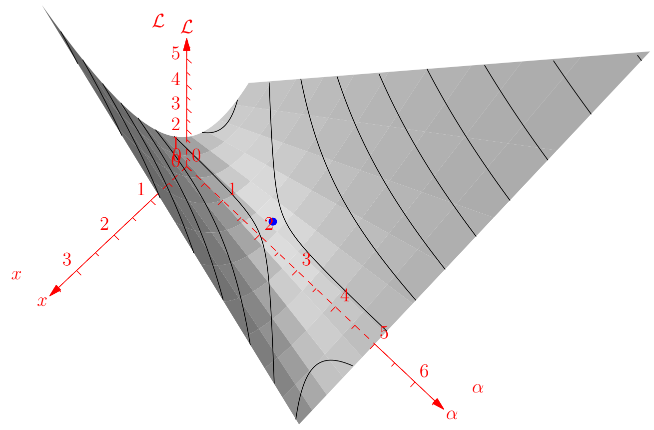
and $\frac{\partial \mathcal{L}}{\partial \alpha} = -g(\boldsymbol{x}) = -x + 2y + 3 = 0$

- Solving simultaneous equations gives minima at $(x,y) = (\frac{3}{4}, -\frac{9}{8})$ with $\alpha = \frac{21}{8}$

Surface



Saddle-Point $y=-9/8$



Multiple Constraints

- Given an optimisation problem with multiple constraints

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } g_k(\boldsymbol{x})=0 \text{ for } k=1,2,\ldots,m$$

- We introduce multiple Lagrange multipliers

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \alpha_k g_k(\boldsymbol{x})$$

- The condition for an optima is $\boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha})=0$ which implies

$$\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) = \sum_{k=1}^{m} \alpha_k \boldsymbol{\nabla}_{\boldsymbol{x}} g_k(\boldsymbol{x})$$

plus the original constraints $\frac{\partial \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha})}{\partial \alpha_k} = -g_k(\boldsymbol{x}) = 0$

Example

- Minimise $f(\boldsymbol{x}) = x^2+2y^2+5z^2-xy-xz$ subject to $g_1(\boldsymbol{x}) = x-2y-z-3=0$ and $g_2(\boldsymbol{x}) = 2x+3y+z-2=0$
- Writing $\mathcal{L}(\boldsymbol{x},\alpha) = f(\boldsymbol{x}) - \alpha_1 g_1(\boldsymbol{x}) - \alpha_2 g_2(\boldsymbol{x})$
- Condition for minima is $\boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{L}=0$ or $\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) = \sum_{k=1}^{2}\alpha_k \boldsymbol{\nabla}_{\boldsymbol{x}} g_k(\boldsymbol{x})$

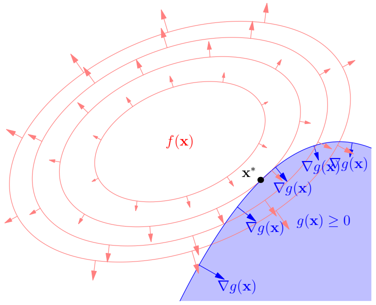
$$\begin{pmatrix} 2x-y-z \\ -x+4y \\ 10z-x \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

and $\frac{\partial \mathcal{L}}{\partial \alpha_i} = -g_i(\boldsymbol{x}) = 0$

- Solving simultaneous equations gives minima at $(\frac{37}{20}, -\frac{11}{20}, -\frac{1}{20})$ with $\alpha_1=3$ and $\alpha_2=\frac{13}{20}$

Outline

1. Constrained Optimisation
2. **Inequalities**
3. Duality



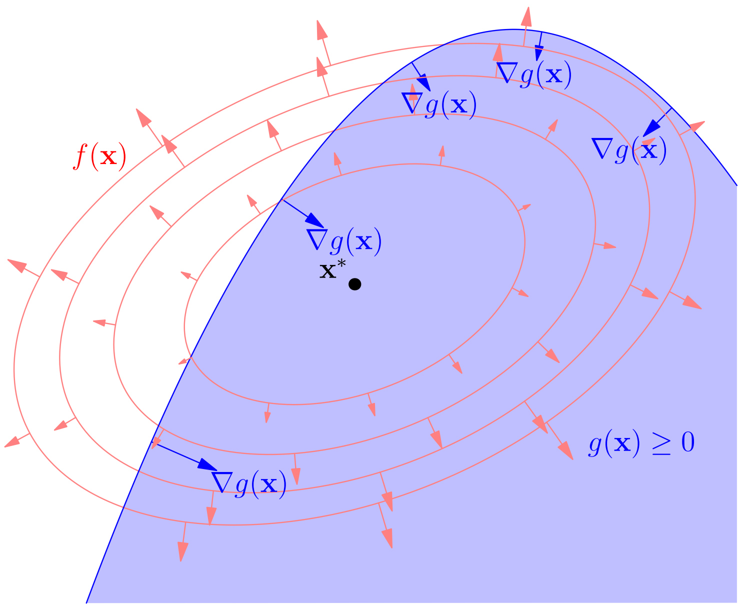
Inequality Constraints

- Suppose we have the problem

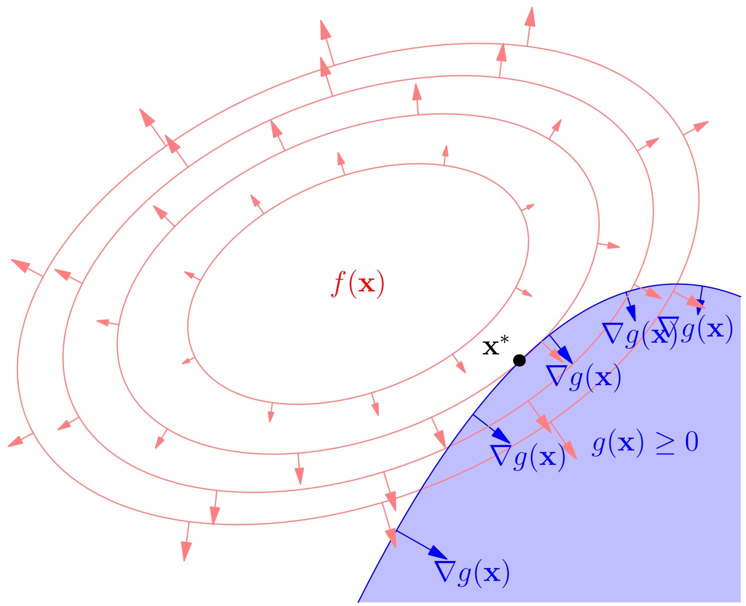
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } g(\boldsymbol{x}) \geq 0$$

- Looks much more complicated, but
- Only two things can happen
  - Either a minimum, $\boldsymbol{x}^*$, of $f(\boldsymbol{x})$ satisfies $g(\boldsymbol{x}^*) > 0$
    - We then have an unconstrained optimisation problem
  - Otherwise, it satisfies $g(\boldsymbol{x}^*) = 0$
    - We have a constrained optimisation problem

Inside Region



On the Boundary



KKT Conditions

- To minimise $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) \geq 0$

$$\mathcal{L}(\boldsymbol{x}, \alpha) = f(\boldsymbol{x}) - \alpha g(\boldsymbol{x})$$

- Then $\boldsymbol{\nabla}_{\boldsymbol{x}} \mathcal{L} = 0$ or

$$\boldsymbol{\nabla}_{\boldsymbol{x}} \mathcal{L} = \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) - \alpha \boldsymbol{\nabla}_{\boldsymbol{x}} g(\boldsymbol{x}) = 0$$

- where either
  - ⋆ $\alpha = 0$ and the solutions in the interior or
  - ⋆ $\alpha > 0$ and $g(\boldsymbol{x}) = 0$, i.e. the solution is on the boundary

- These conditions are known as the Karush-Kuhn-Tucker conditions

Many Inequalities

- Given the problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } g_k(\boldsymbol{x}) \geq 0 \text{ for } k = 1, 2, \ldots, m$$

- We introduce multiple Lagrange multipliers

$$\mathcal{L}(\boldsymbol{x}, \alpha) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \alpha_k g_k(\boldsymbol{x})$$

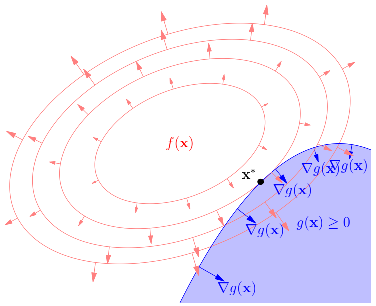
- The condition for an optima is

$$\boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) = \sum_{k=1}^{m} \alpha_k \boldsymbol{\nabla}_{\boldsymbol{x}} g_k(\boldsymbol{x})$$

- Plus the constraints that either $\alpha_k = 0$ or $\alpha_k > 0$ and $g_k(\boldsymbol{x}) = 0$

Outline

1. Constrained Optimisation
2. Inequalities
3. **Duality**



Solving the Lagrangian for x

- Consider minimising a function $f(\boldsymbol{x})$ subject to a set of constraints $g_i(\boldsymbol{x}) = 0$ or $g_i(\boldsymbol{x}) \leq 0$

- We can consider this a double optimisation problem

$$\max_{\alpha} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = \max_{\alpha} \min_{\boldsymbol{x}} \left(f(\boldsymbol{x}) + \sum_i \alpha_i g_i(\boldsymbol{x}) \right)$$

  where there would be constraints on α_i if we had an inequality constraint

Dual Problem

- If $f(\boldsymbol{x})$ and $g_i(\boldsymbol{x})$ are simple we can sometimes find a set of variables $\boldsymbol{x}^*(\boldsymbol{\alpha})$ that minimises the Lagrangian

$$\boldsymbol{\nabla}_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = 0$$

- This leaves us with the **dual problem**

$$\max_{\alpha} \mathcal{L}(\boldsymbol{x}^*(\boldsymbol{\alpha}), \boldsymbol{\alpha})$$

- If we had an inequality constraint $g_i(\boldsymbol{x}) \geq 0$ then we would have the additional constraint in the dual problem $\alpha_i \geq 0$

Linear Programming

- In linear programming we minimise a linear objective function $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}$ subject to linear constraints $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{M}\boldsymbol{x} - \boldsymbol{b} = 0$ (or $\boldsymbol{g}(\boldsymbol{x}) \geq 0$)
- The Lagrangian becomes

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{c}^\mathsf{T}\boldsymbol{x} - \boldsymbol{\alpha}^\mathsf{T}(\mathbf{M}\boldsymbol{x} - \boldsymbol{b})$$

- An equivalent way of writing the Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{x}^\mathsf{T}(\mathbf{M}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{c})$$

- An entirely equivalent interpretation is that we maximise an objective function $\boldsymbol{b}^\mathsf{T}\boldsymbol{\alpha}$ subject to constraints $\mathbf{M}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{c} = 0$ (or $\mathbf{M}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{c} \leq 0$)

Linear Programming Example

- Suppose we eat potatoes and rice and we want to ensure that we get enough vitamin A and C

	Potatoes	Rice	Daily Requirement
Vitamin A	3	5	20
Vitamin C	5	2	24
Price	5	4	

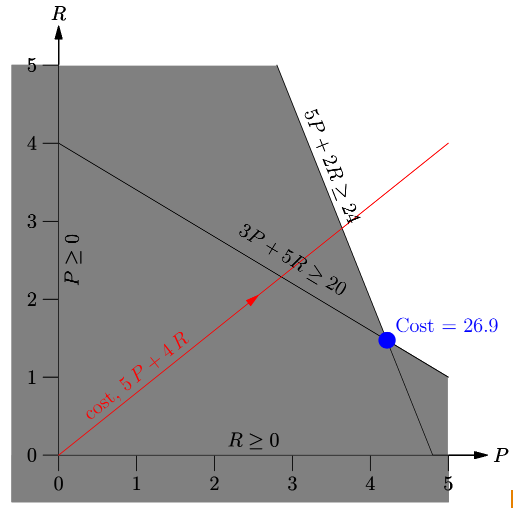
- We want to buy P kg potatoes and R kg of rice as cheaply as possible subject to fulfilling our vitamin requirement

$$\min_{P,R} 5P + 4R$$

subject to $\quad P, R \geq 0, \quad 3P + 5R \geq 20 \quad$ and $\quad 5P + 2R \geq 24$

Linear Programming

- Minimise $5P + 4R$
- Subject to
  - $3P + 5R \geq 20$
  - $5P + 2R \geq 24$
  - $P, R \geq 0$



Lagrangian

- We can write the problem as a Lagrange problem

$$\min_{P,R}\max_{A,C} \quad 5P + 4R - A(3P + 5R - 20) - C(5P + 2R - 24)$$

- subject to $P, R, A, B \geq 0$
- A and C are Lagrange multipliers for vitamin A and C
- We can rearrange the Lagrangian to obtain

$$\max_{A,C}\min_{P,R} \quad 20A + 24C - P(3A + 5C - 5) - R(5A + 2C - 4)$$

Dual Problem

- The Lagrangian

$$\max_{A,C}\min_{P,R} \quad 20A + 24C - P(3A + 5C - 5) - R(5A + 2C - 4)$$

leads to the dual problem

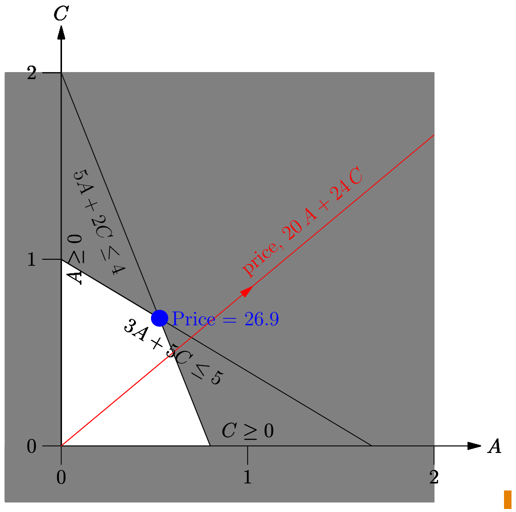
$$\max_{A,C} 20A + 24C$$

$$\text{subject to} \quad 3A + 5C \leq 5 \quad 5A + 2C \leq 4 \quad A, C \geq 0$$

- Consider someone selling vitamins A and C. They want to maximise the price of vitamins A and C, but their prices cannot exceed the price of the vitamins in potatoes or rice

Dual Linear Programme

- Maximise $20A + 24C$
- Subject to
  - ⋆ $3A + 5C \leq 5$
  - ⋆ $5A + 2C \leq 4$
  - ⋆ $A, C \geq 0$



Why?

- Why are we bothered about translating one linear programme into another?
- Sometime one form is massively easier to solve than the other
- This is because the first linear programme depends on the dimensionality of $\boldsymbol{x}$ while the second linear programme depends on the number of constraints (or dimensionality of $\boldsymbol{\alpha}$)
- This is important, for example, in Wasserstein GANs

Quadratic Programming

- A quadratic programme involves minimising a quadratic function $\boldsymbol{x}^\mathsf{T}\mathbf{Q}\boldsymbol{x}$ (with $\mathbf{Q} \succ 0$) subject to linear constraints $\mathbf{M}\boldsymbol{x} = \boldsymbol{b}$ (or $\mathbf{M}\boldsymbol{x} \leq \boldsymbol{b}$)
- We can define the Lagrangian

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = \boldsymbol{x}^\mathsf{T}\mathbf{Q}\boldsymbol{x} - \boldsymbol{\alpha}^\mathsf{T}(\mathbf{M}\boldsymbol{x} - \boldsymbol{b})$$

- Where the solution is given by $\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha})$
- If the constraints are inequality constraints then $\alpha_i \geq 0$

Solution to Quadratic Programming Problem

- Using

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = \boldsymbol{x}^\mathsf{T}\mathbf{Q}\boldsymbol{x} - \boldsymbol{\alpha}^\mathsf{T}(\mathbf{M}\boldsymbol{x} - \boldsymbol{b})$$

- Then

$$\boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = 2\mathbf{Q}\boldsymbol{x} - \mathbf{M}^\mathsf{T}\boldsymbol{\alpha}$$

- So $\boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = 0$ implies

$$\boldsymbol{x}^* = \frac{1}{2}\mathbf{Q}^{-1}\mathbf{M}^\mathsf{T}\boldsymbol{\alpha}$$

Dual Quadratic Programming Problem

- Substituting $\boldsymbol{x}^* = \frac{1}{2}\mathbf{Q}^{-1}\mathbf{M}^\mathsf{T}\boldsymbol{\alpha}$ into

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha}) = \boldsymbol{x}^\mathsf{T}\mathbf{Q}\boldsymbol{x} - \boldsymbol{\alpha}^\mathsf{T}(\mathbf{M}\boldsymbol{x} - \boldsymbol{b})$$

- We get the dual problem

$$\max_{\boldsymbol{\alpha}} -\frac{1}{4}\boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\mathbf{Q}^{-1}\mathbf{M}^\mathsf{T}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{b}$$

- If the constraints were inequality constraints then we have $\alpha_i \geq 0$
- We have exchanged one quadratic programme for another, but sometimes that very useful (e.g. SVMs)

Lessons

- A useful tool for performing constrained optimisation is the introduction of Lagrange multipliers
- This is particularly useful for problems with unique solutions (it will work when there are multiple solutions, but finding many saddle points is a pain)
- For inequality constraints we need to satisfy KKT conditions
- For simple situations (linear and quadratic programming) we can eliminate the original variables to obtain the dual problem