Convex Optimisation

Alessio Zakaria

Structure

- Convexity and Convex Optimisation
- Duality
- Solving Convex Optimisation Problems
  - Newton's Method for Equality Constraints
  - Barrier Method for Inequality Constraints

Convex Optimisation?

- What is it?
  - Finding the maxima / minima of convex functions over convex sets with respect to convex or affine constraints
- Why do we care?
  - Convex functions display theoretical properties that are suited to optimisation
  - Solving convex optimisation problems allows you to place a lower bound on non-convex optimisation problems
  - Many real world optimisation problems are convex

Convexity

A set $C \subseteq \mathbb{R}$ is convex if $\forall\, x, y \in C$, $\forall\, \theta \in [0, 1]$

$$\theta x + (1 - \theta) y \in C$$

i.e. all points on the line segment between x and y lie in C.

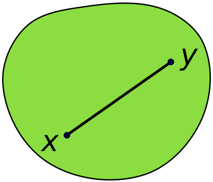

Figure: Points in a convex set

Convexity

A function $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ is convex iff C is a convex set and $\forall x_1, x_2 \in C, \forall t \in [0, 1]$ (strict if strict equality):

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

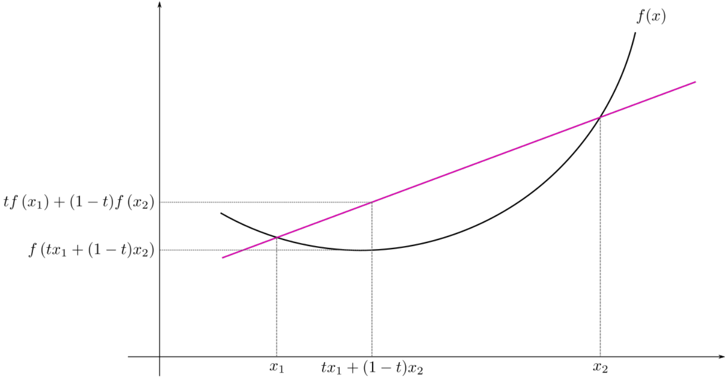

Figure: A convex function

Minima and Maxima

$f : S \subseteq \mathbb{R}^n \to \mathbb{R}$ has a global minimum (maximum resp.) $x^\star$ if:

$$\forall\, x \in S,\; f(x^\star) \leq f(x),\; (f(x^\star) \geq f(x))$$

f has a local minimum (maximum resp.) around $x^\star$ if $\exists\, R \in \mathbb{R}$ such that

$$\forall x \in B(x^\star, R),\; f(x^\star) \leq f(x),\; (f(x^\star) \geq f(x))$$

If these inequalities are strict then $x^\star$ is known as a strict maximum (minimum)

Convex Optimisation Problem

A convex optimisation problem is a minimisation problem in the following form:

$$
\begin{aligned}
\text{Find the } \min_x \ & f_0(x) \\
\text{such that } & f_1(x) \leq 0 \\
& \vdots \\
& f_n(x) \leq 0 \\
& g_0(x) = 0 \\
& \vdots \\
& g_m(x) = 0
\end{aligned}
$$

Where $f_0, \cdots, f_n$ are convex functions $\mathbb{R}^p \to \mathbb{R}$ and $g_0, \cdots, g_m$ are affine $\mathbb{R}^p \to \mathbb{R}$.

Feasible Set

The feasible set

$$C \subseteq D = \bigcap_{i=0}^{m} \operatorname{dom}(f_i) \bigcap_{j=0}^{n} \operatorname{dom}(g_j)$$

is the set of all $x \in D$ such that the constraints are satisfied.

The *optimal value* of problem is $\inf\{f_0(x) \mid x \in C\}$. If there is no lower bound the optimisation problem has optimal value $-\infty$.

The point in the feasible set, $x \in C$, that attains the optimal value under the objective is called the *optimal solution*.

Feasible Set

Some important properties of convexity are:

- The intersection of convex sets is convex
- The sublevel sets ($\{x \mid f(x) \leq 0\}$) of convex functions are convex
- The preimage of affine functions is convex

Because of this the feasible set of of a convex optimisation problem is convex.

Convexity and Optimality

The above is important because of the following key fact:

Any local minimum of a convex function over a convex set is a global minimum

This means that any algorithm for a convex optimisation problem only has to find a local minimum.

This guarantees that algorithms which are only guaranteed to find local minima for all problems provide global minima for convex problems.

Characterisations of Convexity

First Order Characterisation of Convexity: $f : \mathbb{R}^n \to \mathbb{R}$ differentiable is convex iff $\mathbf{dom}(f)$ is convex and $\forall\, x\, y \in \mathbf{dom}(f)$:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

i.e. the tangent plane underapproximates the function at every point.

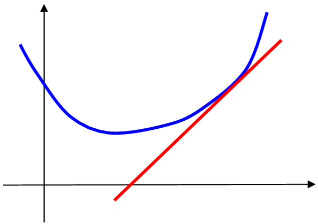

Figure: Tangent line to a convex function

Characterisations of Convexity

First Order Characterisation of Convexity: $f : \mathbb{R}^n \to \mathbb{R}$ differentiable is convex iff $\mathbf{dom}(f)$ is convex and $\forall\, x\, y \in \mathbf{dom}(f)$:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

i.e. the tangent plane underapproximates the function at every point.

Second Order Characterisation of Convexity: $f : \mathbb{R}^n \to \mathbb{R}$ twice differentiable is convex iff $\mathbf{dom}(f)$ is convex and

$$\nabla^2 f \succeq 0$$

i.e. the Hessian of f is positive semidefinite at all points.

Examples of Convex Optimization Problems

Various optimisation problems are convex optimization problems:

- Linear Programming
- Quadratic Programming with linear or convex quadratic constraints
- Least Squares

Application: Norm Approximation

An optimisation problem that occurs repeatedly in statistics is norm approximation.

$$\underset{x}{\text{minimise}} \; \|Ax - b\|$$

$A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^b$.

This problem appears in many statistical estimation tasks e.g. maximum likelihood estimation for simple linear regression. This problem is convex for all norms.

Application: Regularisation and Optimisation

Regularisation is an approach to constraining the parameters of the norm minimisation somehow using an additive term on the objective. This appears in statistics when some prior information is known about the solution to the parameters i.e. through the choice of prior in MAP estimates:

$$\underset{x}{\text{minimise}}\ \|Ax - b\| + \gamma\|x\|$$

This can be reframed as a convex constrained optimisation problem i.e. for L_2 regularisation we can see this as the constrained optimisation problem:

$$\underset{x}{\text{minimise}}\ \|Ax - b\| \qquad\qquad s.t. \|x\|^2 - c \leq 0$$

As in Tikhonov regression

Duality and the Langrangian

For any optimisation problem there is a related problem known as the dual (with the original problem known as the primal).

The solution to the dual provides some information about the solution to the primal. The dual problem is related to the Lagrangian function of the original problem.

The dual problem is always convex and therefore being able to solve convex optimisation problems allows us to solve the dual.

Duality

Consider some optimisation problem:

$$\begin{aligned}
\text{Find the } \min_x \ & f_0(x) \\
\text{such that } & f_1(x) \leq 0 \\
& \vdots \\
& f_n(x) \leq 0 \\
& g_0(x) = 0 \\
& \vdots \\
& g_m(x) = 0
\end{aligned}$$

Where $f_0, \cdots, f_n, g_0, \cdots, g_m : \mathbb{R}^p \to \mathbb{R}$

The Lagrangian

The Lagrangian, $L : \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, of this optimisation problem is the function:

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=0}^{m} \mu_j g_j(x)$$

The parameters λ and μ are known as the Lagrangian multipliers

The Dual Function

The Dual Function $L^\star : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the function

$$L^\star(\lambda, \mu) = \min_{x \in D} L(x, \lambda, \mu)$$

For any feasible point of the primal x, for all values $\lambda \geq 0$, $\forall \mu$

$$L^\star(\lambda, \mu) \leq f_0(x)$$

We call (λ, μ) dual feasible iff

$$\lambda \geq 0 \text{ and } L^\star(\lambda, \mu) > -\infty$$

The Dual Problem

Given that the values of the dual function lower bound the optimal solution for the primal we can formulate the question of what the greatest lower bound is.

This leads to the optimisation problem.

$$\begin{aligned} \text{maximise} \;\; & L^{\star}(\lambda, \mu) \\ \text{such that} \;\; & \lambda \geq 0 \end{aligned}$$

With optimal value $g^{\star}$ which is less than $f_0(x)$ for any x in the feasible set for the primal including the optimal value i.e. $g^{\star} \leq p^{\star}$.

Dual feasible points are points that are (thankfully) feasible for the dual problem.

Duality

Regardless of whether the primal problem is convex the dual problem is *always* convex.

This means that if convex optimisation problems can be solved efficiently then a lower bound can be placed on any optimisation problem that may be harder to solve.

This can be used in and of itself but can also be used to derive stopping conditions on algorithms aiming to solve the primal e.g.

For a given feasible point x we know that the solution to the primal $p^\star$ lies between $f_0(x)$ and the solution to the dual $g^\star$. We therefore know that x is at most $(f_0(x) - g^\star)$-suboptimal.

Strong Duality

If the solution to the primal is the same as the solution to the dual then we say that *strong duality* holds.

For convex problems if the feasible region has an interior point (known as Slater's condition) then strong duality holds.

If strong duality holds then the solution to the primal $x^\star$ maximises the Lagrangian with respect to the solution of the dual i.e. $\lambda^\star$ and $\mu^\star$ i.e.

$$x^\star = \min_x L(x, \lambda^\star, \mu^\star)$$

Solving Convex Optimisation Problems

One method of solving convex optimisation problems is using interior point methods.

Intuition: Reformulate the problem as an equality constrained problem which can be solved via Newton's Method

Newton's Method

Newton's method is an iterative method for solving unconstrained minimisation problems:

- For twice differentiable functions
- Uses the Hessian to take into account local curvature information

To find a minimum of some twice differentiable $f(x)$ with gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$ take iterative steps.

Newton's Method

At point x_n, set $x_{n+1} = x_n + \lambda \Delta x$. Δx is set using the second order Taylor's approximation as follows:

$$\Delta x = \underset{\Delta x \in \mathbb{R}^d}{\operatorname{argmin}}\ f(x + \Delta x) = \underset{\Delta x \in \mathbb{R}^d}{\operatorname{argmin}}\ f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

Taking the derivative of the final expression, equating to 0 and solving for Δx we obtain:

$$\Delta x = -(\nabla^2 f(x))^{-1} \nabla f(x)$$

Newton's Method for Equality Constrained Optimisation

Newton's Method can be used to find the minima of functions according to equality constraints i.e.

$$\begin{aligned} &\underset{x}{\text{minimise}}\ f(x) \\ &\text{such that } Ax = b \end{aligned}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ twice differentiable, $A \in \mathbb{R}^{n\times m}$ where Δx is calculated with the 2$^{\text{nd}}$ order Taylor Approximation:

$$\begin{aligned} \Delta x &= \underset{\Delta x \in \mathbb{R}^d}{\operatorname{argmin}}\ f(x + \Delta x) \\ &= \underset{\Delta x \in \mathbb{R}^d}{\operatorname{argmin}}\ f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T H(x) \Delta x \\ &\quad \text{such that } A(x + \Delta x) = b \end{aligned}$$

Which is a quadratic constrained minimisation problem.

Newton's Method for Equality Constrained Optimisation

If f is convex (and Slater's constraint holds) then strong duality holds assuming that a minimum exists at some $x^\star$.

Forming the Lagrangian for an equality constrained problem we obtain:

$$L(x, v) = f(x) + A^T v$$

Since $x^\star$ minimizes $L(x, v^\star)$ over x, it follows that its gradient must vanish at $x^\star$, i.e.

$$\begin{aligned} 0 &= \nabla f(x^\star) - \nabla A^T v^\star \\ \nabla f(x^\star) &= \nabla A^T v^\star \\ \nabla f(x^\star) &= A^T v^\star \end{aligned}$$

Newton's Method for Equality Constrained Optimisation

We therefore have two constraints that must be satisfied for minimum $x^\star$.

$$\begin{aligned} \nabla f(x^\star) &= A^T v^\star \text{ and} \\ Ax^\star &= b \end{aligned}$$

Replacing $x^\star$ with $x + \Delta x$ using the 2nd order Taylor approximation of the original f the constraints become:

$$\begin{aligned} \nabla^2 f(x)\Delta x + \nabla f(x) &= A^T v \text{ and} \\ A(x + \Delta x) &= b \end{aligned}$$

Which, if $Ax = b$ gives rise to the KKT system:

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ v \end{bmatrix} = \begin{bmatrix} -\Delta f(x) \\ 0 \end{bmatrix}$$

This can be analytically solved for Δx if f is strictly convex as then the Hessian of f is full rank and thus invertible.

Solving Inequality Constrained Optimisation

We now know how to deal with equality constraints in convex optimisation problems. Now it remains to deal with inequality constraints. We incorporate the inequalities into the objective using an indicator function:

$$\begin{aligned} \text{minimise} \;\; & f_0 + \sum_{i=1}^{m} \mathcal{I}_{-}(f_i(x)) \\ \text{such that } & Ax = b \end{aligned}$$

where

$$\mathcal{I}_{-}(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$$

Solving Inequality Constrained Optimisation

This objective is not differentiable and therefore Newton's method cannot be applied.

To rectify this we can use an approximation using the log barrier function.

$$\hat{\mathcal{I}}_{-}(u) = -\frac{1}{t}\log(-u)$$

$$\textbf{dom } \hat{\mathcal{I}}_{-} = -\mathbb{R}^{++}$$

Where t is a parameter that is set to adjust the accuracy of the barrier function

Log Barrier Function

The minimisation problem can be reformuated as the following:

$$\begin{aligned} \text{minimise} \;\; & tf_0 + \varphi(x) \\ \text{such that} \;\; & Ax = b \end{aligned}$$

where

$$\varphi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$$

where

$$\nabla\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x)$$

Log Barrier Function

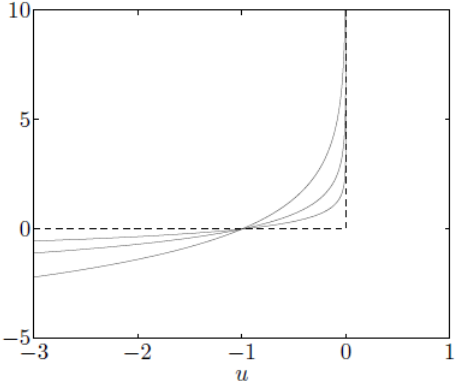

Figure: log barrier function for increasing values of t

The log barrier function is twice continuously differentiable and convex and therefore is amenable to Newton's method.

Central Path

Before applying Newton's method however, it is useful to know what value we should set t to. This can be found through analysing the *central path*.

Assuming the log barrier formalisation has a unique solution for all $t > 0$ obtainable via Newton's method, we can define the central path, $x^\star(t)$ as the *unique* solution to the minimisation problem for a given t.

From the KKT conditions a dual feasible point for every $x^\star(t)$ can be obtained which lower bounds the optimal value for the original problem.

Central Path

Let

$$\lambda_i^\star(t) = -\frac{1}{tf_i(x^\star(t))} \text{ and } v^\star(t) = \hat{v}/t$$

We claim that these are dual feasible points for the dual function

$$g(\lambda(t), v(t)) = \min_{x(t)} f_0(x(t)) + \sum_{i=1}^{m} \lambda_i(t) f_i(x(t)) + v(t)^T(Ax(t) - b)$$

and that $x^\star(t)$ is the value of $x(t)$ that minimises the above for $\lambda_i^\star(t)$ and $v^\star(t)$.

Central Path

If this is true then

$$g(\lambda^\star(t), v^\star(t)) = f_0(x^\star(t)) + \sum_{i=1}^{m} \lambda_i^\star(t) f_i(x^\star(t)) + v^\star(t)^T (Ax^\star(t) - b)$$

As $\quad Ax^\star(t) = b$ and $\lambda_i^\star(t) f_i(x^\star(t)) = -\dfrac{f_i(x^\star(t))}{t f_i(x^\star(t))} = -\dfrac{1}{t}$

$$g(\lambda^\star(t), v^\star(t)) = f_0(x^\star(t)) - \frac{m}{t}$$

This means that for a given $x^\star(t)$, $f_0(x^\star(t))$ is no more than $\frac{m}{t}$ away from the optimal value for the original primal problem.

Central Path

We claimed that $\lambda^\star(t)$ and $v^\star(t)$ were dual feasible points for the dual. To show this observe the following:

Points on the central path are characterised by the KKT conditions:

$$\begin{aligned} \exists \hat{v} \ s.t. \quad 0 &= t\nabla f_0(x^\star(t)) + \nabla\phi(x^\star(t)) + A^T\hat{v} \\ &= t\nabla f_0(x^\star(t)) + \sum_{i=1}^{m} \frac{1}{-f_i(x^\star(t))}\nabla f_i(x^\star(t)) + A^T\hat{v} \end{aligned}$$

This can be observed to be the derivative of the Lagrangian where $\lambda = \lambda^\star(t)$ and $v = v^\star(t)$. As this equals 0 this means that the dual function is bounded below for $\lambda^\star(t)$ and $v^\star(t)$.

Central Path

The other condition for dual feasibility is that $\lambda^\star(t) > 0$.

As $f_i(x^\star) < 0$ for $x^\star$ then

$$-\frac{1}{f_i(x^\star)} > 0$$

And therefore $\lambda^\star(t)$ and $v^\star(t)$ are dual feasible and the suboptimality bound holds.

Central Path

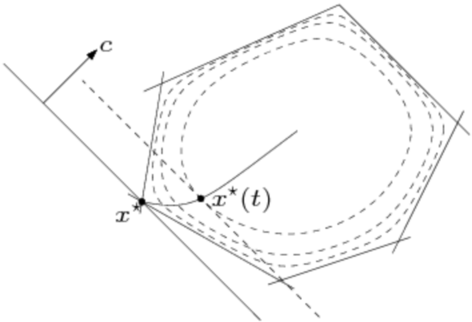

Figure: Traversing the central path to the optimal value

As $t \to \infty$ the central path solution will tend to $p^\star$.

Barrier Method

Given that an abitrary $x^\star(t)$ is $\frac{m}{t}$ suboptimal a method of finding the minimum of the original problem to a specific accuracy, ε is by solving the problem:

$$\begin{aligned} \text{minimise} \quad & \left(\frac{m}{\varepsilon}\right) f_0 + \varphi(x) \\ \text{such that } & Ax = b \end{aligned}$$

This can be solved using Newton's method however it does not have good practical performance due to numerical instability as when ε is large then the value of the objectives Hessian varies rapidly near the boundaries of the feasible set.

Barrier Method

The barrier method avoids the problems of the previous method by solving a series of minimisation problems.

The barrier method computes $x^{\star}(t)$ for increasing values of t using Newton's method until $t \geq \frac{m}{\varepsilon}$.

Barrier Method Algorithm

Algorithm 1 Barrier Method

1: **while** $m/t < \varepsilon$ **do**
2: $\quad x \leftarrow$ minimise $tf_0 + \phi$ such that $Ax = b$ starting at x
3: $\quad t \leftarrow \mu t$
4: **end while**

The method of minimisation on line 2 is assumed to be Newton's Method.

The choice of μ dictates how many of the outer non-Netwon's method loop occur. If this is too high then Newton's method might take too long to converge. If $t^{(0)}$ is too small then many outer iterations may be needed.

The Barrier Method is quite robust to the choice of these in practice.

Interior Point Methods

There exist more advanced interior point methods that are more efficient for higher accuracy than the Barrier method.

Primal-Dual methods are an active area of research for non-linear convex problems that are more efficient than Barrier methods.

Primal-Dual methods take one Newton step instead of the separation between the outer loop of the Barrier Method and updating the inner iterations of Newton's method.

Summary

In this talk we have:

- Learnt what a convex optimisation problem is
- Learnt about duality and how convex optimisation can be used to lower bound all optimisation problems
- Learnt about how Newton's Method can be adapted to solve equality constrained problems
- Learnt about how the Barrier method can be used to incorporate inequalities into the objective and iteratively apply Newton's method to solve inequality constrained problems.

Thanks

Any Questions?