Optimization

Arthur Conmy*

Part IB, Easter Term 2020

These notes are based on lectures given (virtually) by Dr M. Tehranchi in Easter term 2020. However, they are intended not as a complete set of notes but instead attempt to cover both the more interesting and challenging parts of the course.

The 'intuition focused' second year notes[1] by Neel Nanda provide inspiration for this background, though I in no way claim for these to be as insightful as those. Credit is also due to Evan Chen for the style file for these notes[2].

§1 The Lagrangian Background

We will be solving optimization problems that look like

$$\min_{x \in X} f(x) \tag{1}$$

throughout this course, where $X \subset \mathbb{R}^n$. Because of this, we shall not bolden or underline vectors or matrices since they will be used so frequently. This can lead to confusion but it simplifies statements and proofs significantly. In addition, inequalities with vectors mean these hold for *all* components, i.e. $x \geq y$ means $x_1 \geq y_1$, $x_2 \geq y_2$, $\cdots$.

§1.1 Lagrangian Sufficiency

You probably know what the Lagrangian is from IB Variational Principles, but in this course we deal with it in a way that is somewhat different, so it's worth building things up differently.

Definition 1.1 (The Lagrangian)**.** Suppose we wish to solve the problem

$$\min_{x \in X} f(x)$$

subject to the condition $g(x) = b$. Then the *Lagrangian* for this problem is the function

$$L(x, \lambda) = f(x) + \lambda^T (b - g(x)).$$

In IB Variational Principles, we interpreted this quantity as something that encoded the *regional constraint(s)* $g(x) = b$ by the fact that $\frac{\partial L}{\partial \lambda_i} = 0$ were together equivalent to that same regional constraint. For our purposes, the following more general result will be more useful:

*Please send any corrections and/or feedback to asc70@cam.ac.uk.

[1] Available here: https://dynalist.io/d/ToOhEKlz9qC2PmpjgyTpXQJz.

[2] Available here: https://github.com/vEnhance/dotfiles/blob/master/texmf/tex/latex/evan/evan.sty.

Theorem 1.2 (Lagrange Sufficiency Theorem)

Suppose we have λ and x^* with $g(x^*) = b$ such that for all x with $g(x) = b$

$$L(x^*, \lambda) \leq L(x, \lambda).$$

Then $f(x^*)$ is the answer to our problem (1).

Proof. Exercise. □

Example 1.3 (Gauss-Markov in one dimension[a])

Suppose from statistics that we have a vector of random variables Y that we expect are distributed as

$$Y_i \sim \beta x_i + \varepsilon_i$$

where the ε_i each have mean 0 and variance σ^2, and all their pairwise covariances are 0. We assume that all the x_i are known and fixed to begin with, and that we measure Y_i. We wish to estimate β by $\hat{\beta}$, an expression linear in the Y_i terms and with expectation β (i.e, find an unbiased linear estimator) that minimises $\mathrm{Var}\hat{\beta}$.

Reformulating as an optimization problem, we need find the w_i such that $\hat{\beta} = \sum w_i Y_i$, $\sum w_i x_i = 1$ (to encode the expectation) and we minimise $\sigma^2 \sum w_i^2$, i.e we should consider

$$L(w, \lambda) = \sigma^2 \sum w_i^2 + \lambda \left(1 - \sum w_i x_i\right).$$

which we can rewrite as

$$L(w, \lambda) = \lambda + \sum \sigma^2 w_i^2 - \lambda x_i w_i$$

And now we magically pick $\lambda = 2\sigma^2 \left(\sum x_i^2\right)^{-1}$, and $w_i^* = x_i \left(\sum x_i^2\right)^{-1}$, and since these w_i minimise the quadratics inside the sum and indeed satisfy $\sum w_i x_i = 1$, by the Lagrange Suffiency Theorem we've found our unbiased estimator

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

[a] Credit to https://math.stackexchange.com/a/2166172/403623. Note that the derivation in this post follows the Variational Principles approach to the Lagrangian method.

The example illustrates an important step in the Lagrangian method: by grouping all the x terms together we can choose the appropriate λ and x^* to apply LST.

Problem 1.4 (2002 4/II/14H[3])**.**

§1.2 Duality and Lagrangian Necessity

Bearing in mind LST, it is now natural to ask the question of given a fixed λ, what the value of $x \in X$ is such that $L(x, \lambda)$ is minimised. Note that if this value of x satisfies $g(x) = b$, we're immediately done! But to find such a λ, we will need to introduce the *dual problem* to (1).

[3] Available at https://www.maths.cam.ac.uk/undergrad/pastpapers/files/2002/PaperIB_4.pdf

Considering a fixed λ, we can try and minimise $L(x, \lambda)$ over X. However, in practise it is often the case that some λ choices result in an unbounded $L(x, \lambda)$, i.e we can pick x to make L approach $-\infty$. So we simply restrict ourselves to the λ that don't make this happen in the most direct way possible:

Definition 1.5. The set of *feasible Lagrange multipliers* is the set

$$\Lambda = \left\{ \lambda : \inf_{x \in X} L(x, \lambda) > -\infty \right\}. \tag{2}$$

And now we can formulate the dual problem:

Definition 1.6. The *dual problem* to (1) is the problem

$$\max_{\lambda \in \Lambda} h(\lambda). \tag{3}$$

where

$$h(\lambda) = \inf_{x \in X} L(x, \lambda). \tag{4}$$

We will now see that the Lagrangian method works precisely when we 'meet in the middle' between (1) and the dual problem. To see this, first note that $g(x) = b$ and $\lambda \in \Lambda$ imply $h(\lambda) \leq f(x)$ (write out definitions) and so that if we ever have $g(x^*) = b$ and $f(x^*) = h(\lambda)$, then $f(x^*) \leq f(x)$ for all x with $g(x) = b$, i.e x^* is a solution.

Now it turns out that we can even more precisely pin down when the Lagrangian method will work, and this will have a close connection with convexity. To do this we generalise (3) to a family of problems:

Definition 1.7 (Value Function)**.** Consider the problem (1). Then $\phi(c)$ is the answer to the problem when we set $b = c$.

and also introduce

Definition 1.8 (Supporting hyperplane)**.** Suppose ϕ is a real-valued function from $\mathbb{R}^m$. Then it has a *supporting hyperplane* at some general point b if we have some λ for all c we have that

$$\phi(b) \geq \phi(c) + \lambda^T (b - c). \tag{5}$$

This is easy to visualise in the case where $m \leq 2$ and ϕ is differentiable: we're just saying that ϕ lies above its derivative at b everywhere. In fact we can be more general than this, and note that convex functions have supporting hyperplanes everywhere, by repeating the proof of Jensen's inequality from 1A Probability[5].

With this setup, we can now prove the other, harder, half of the Lagrange Sufficiency Theorem.

Theorem 1.9 (Lagrange Necessity Theorem)

The Lagrangian method works (i.e we can find a feasible λ so that (1) and the dual problem agree) iff the value function ϕ has a supporting hyperplane at b.

Proof. The idea is to reformulate the condition (5) as a condition involving inf and note that our choice of notation was no accident: $\lambda^T(b - g(x)) + \lambda^T(g(x) - c) = \lambda^T(b - c)$. See lecture notes for a full proof. □

[5]Credit to Daniel for this: http://db808.user.srcf.net/Optimisation.pdf

With these foundations established, we can now approach the linear programming scenario. While I've tried to make things interesting so far, this is higher maths so unfortunately there is even more terminology that we'll need use to establish before getting there. To keep things condensed and hopefully more efficient to get through, I've gathered all the terminology needed in the next section. Note that this includes the definitions of slack variables and complementary slackness - I promised the notes would be brief!

§1.3 Glossary

To apply further results, it is worth establishing some definitions and notation. More comprehensive notes should be referred to in order to, but the table below collects all terms I shall use going forward:

Glossary	
Regional constraint	The condition $x \in X$ in (1). X will often be closed and/or compact so that we can not only take the min but also argmin.
Functional constraint	The condition $g(x) = b$ in (1). Unlike the regional constraint, we general relax this condition when we construct the Lagrangian.
Feasible	When referring to x, this means $x \in X$ satisfying the functional constraint. When referring to λ this means $\lambda \in \Lambda$ (see (2).
Objective function	$f(x)$ in (1). $h(\lambda)$ is the dual objective function.
Convex	A convex set is a set where line segments between points in the set lie entirely in the set. Convex functions are functions which lie entirely below lines between points on their graph.
Slack variables	Variables introduced to handle inequality conditions in functional constraints by turning these into regional constraints; if $x \in X$ and $g(x) \leq b$, then we can consider $(x, z) \in X \times \mathbb{R}_{\geq 0}$ with obvious generalisations to higher dimensions, and cases where some functional contstraints are inequalities and some aren't.
Complementary slackness	This is best illustrated with an example, and is problem dependent. When we introduce slack variables, we get a λz term. When we then take the infimum with respect to x, we will need this to be zero, so either $\lambda = 0$ or $z = 0$, since only non-negative λ will be feasible.

§2 Linear Programming Theory

We're going to be developing methods to approach linear programs, problems where both our objective function and all our regional constraints are linear in all our variables, and hence are both convex (remember convex means something similar, but of course different for functions and sets).

Specifically, we'll use the following form of a linear program:

Definition 2.1 (Fundamental Theorem of Linear Programming (FTLP) Form of a Linear Program)**.** The FTLP form of a linear program with n variables and n conditions is the optimization problem

$$\max_{x \geq 0} c^T x \tag{6}$$

subject to $Ax = b$.

The important result concerning this form is the form of the dual problem:

Problem 2.2. Check that the dual problem to (6) is

$$\min_{\lambda \geq 0} b^T \lambda \tag{7}$$

subject to $A^T \lambda \geq c$.

And now another (sigh) derivation shows that the dual problem of this dual problem is the same primal problem. This means that if the λ dependent term

$$\lambda^T (b - Ax) \tag{8}$$

that should have arisen in solving (2.2).

Recall that the Lagrangian method works if we can get the objective function to agree with its dual counterpart. This means we've established one direction of the *fundamental theorem of linear programming*, that if we have some feasible x and λ[8] and they are optimal for their respective problems, this being equivalent to

$$0 = \lambda^T (b - Ax) = x^T (A^T \lambda - c) \tag{9}$$

then we have the optimal solution to the linear program. A sanity check of this last condition is that it implies the objective function and its dual counterpart agree (why?).

What about the other direction? We want to use the Lagrangian Necessity theorem, and while this seems difficult, note that 'everything is convex': since both our objective function and regional constraint are linear. See the example sheet for how to use this to prove that our value function is convex.

§3 The Simplex Algorithm

§4 More Linear Programs

With a lot of theory developed, we can now use our hammer to find a bunch of nails!

§4.1 The Transportation Problem

This problem has a very 'real world' setup:

Definition 4.1 (The Transportation Problem)**.** Suppose there are m suppliers of a product, and n different destinations they need supply these products to. Suppose also that the ith supplier has a supply S_i and that the jth destination has a demand D_j and that a cost of d_{ij} is incurred when supplying destination j from supplier i. Then what is the optimal way to organise the supplying?

[8] x feasible for the primal problem is actually equivalent to x being feasible when interpreted as the Lagrange multiplier for the dual problem. I don't think understanding this is necessary to establish this direction of FTLP.

But this is just a linear program:

$$\min_{x_{ij} \geq 0} \sum_{i,j} d_{ij} x_{ij} \tag{10}$$

subject to $S_i = \sum_j x_{ij}$ for all i, and $D_j = \sum_i x_{ij}$, and the additional assumption that the total supply $\sum_i S_i$ equals the total demand $\sum_j D_j$.

Of course, we could end the story here since we can now apply the simplex algorithm. But in this case we have two algorithms that will be far more practical in application.

To derive these, consider the Lagrangian

$$L(x, \lambda) = \sum_{i,j} (d_{ij} - \lambda_i - \mu_j) x_{ij} + \sum_i \lambda_i S_i + \sum_j \mu_j D_j \tag{11}$$

which means that the feasible (λ, μ) are those that satisfy $d_{ij} - \lambda_i - \mu_j$ for all i, j. We can now apply FTLP and the BFS theory established: if we have some feasible Lagrange multipliers and a set of x_{ij} values of which exactly $m + n - 1$ are non-zero[10] then this is an optimal solution to the transportation problem.

We can turn this into an algorithm by writing the d_{ij} values in a table, assigning $m + n - 1$ of the x_{ij} values to non-zero values and then performing simplex algorithm pivots, that will be much easier to perform than in the simplex table.

§4.2 The NW Algorithm

[10] The minus one comes from the fact that our equal supply-demand assumption means one of our equations is implied by the others