Recreating Optimisation: The Lagrangian Dual

Ang Ming Liang

December 2021

> What I cannot create, I do not understand.
>
> *Richard Feynman*

1 Preface

I hope that this blog would be useful at conveying and stringing together the high level intuitions and ideas often lost during the busyness of semester. The concept of mathematics is often simple but non-trivial. Often times in an attempt to cover all the content during the semester, we might have missed out the simplicity underlying the topic. It is my hope that someone out there might find this blog enlightening and helpful to them in their studies.

2 Introduction

The field of optimisation is central to machine learning. The task of learning is often formulated through the lens of optimisation (e.g. Empirical Risk Minimisation). Therefore, it makes a lot of sense to try to think carefully and analyse optimisation algorithms to gain new insight into the nature of learning itself.

To aid us in this endeavour, let us first start with a simple but common optimisation problem we might see in practice and see if we can generalise and develop a more robust understanding of optimisation from there. This comes from my own experience studying mathematics major. I often find it is beneficial to think of simple and familiar ideas in simple settings like $\mathbb{R}$ first before generalising these intuitions to more abstract settings like a compact Hausdorff space. This allows the brain to seek the familiar and let these simple intuition become a helpful tour guide through the more abstract and general settings.

Coming back to our central discussion, the problem I have in mind to begin our discussion is something any secondary school/high school student taking

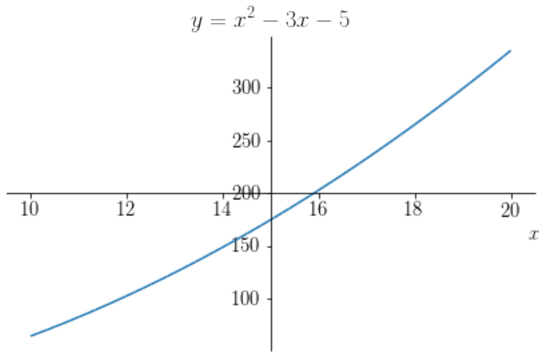

Figure 1: Quadratic function on $[10, 20]$

calculus would know how to solve: how to find the minimum of a $y = f(x) = x^2 - 3x - 5$ on $\mathbb{R}$ i.e

$$\min_{x \in \mathbb{R}} f(x) = x^2 - 3x - 5, \tag{1}$$

To solve this, all we need is to find the turning points and use either the first or second derivative test to check if the turning point is the minimum or maximum point. For the case of the quadratic function, if the turning point we found is a minimum, we are done as it is also the global minimum. Otherwise, there is no point in $\mathbb{R}$ that is the global minimum.

Let us generalise this problem. Now consider trying to find the minimum of a quadratic function on a closed-bounded interval, for example, $[0, 1]$ or $[-12, 3]$ but not $(0, 1)$ or $[1, \infty)$ instead of $\mathbb{R}$. Is looking only at the turning points sufficient to solve our minimisation problem? The answer is no. We also need to consider the endpoints.For example for

$$\min_{x \in [10,20]} f(x) = x^2 - 3x - 5, \tag{2}$$

there are no turning points to be found in $[10, 20]$. However, there is clearly a minimum located at the endpoints of the interval $[10, 20]$.

This need to account for the endpoints is the crucial difference between unconstrained and constraint optimisation. In fact, because of this, mathematicians and computer scientists prefer to work with unconstrained rather than constraint optimisation as there are simply fewer things to consider, especially in higher dimensions. However, we often need to consider constraint optimisation in practice, so how do we convert a constraint optimisation problem to the more familiar unconstrained setting?

3 The Lagrangian

What if we could approximate the unfamiliar with the familiar? This is a running theme in both applied and pure math. In measure theory, you approximate the nasty function you want to integrate using a sequence of simple functions quite literally[1]. In our problem, we create an unconstrained problem that approximates the constraint problem accurately.

How might we go about doing something like that ? Well naively, we could to define a new function that penalises x whenever it goes out of bound i.e no longer follows the constraint. We might do this by first representing our constraint $[10, 20]$ as 2 functions,

$$x \in [10, 20] \implies 10 \leq x \leq 20 \implies h_1(x) = x - 20 \leq 0 \text{ and } h_2(x) = 10 - x \leq 0$$

Hence, our problem of finding $f(x^*)$ which is the minimum of $f(x)$ where $x \in [10, 20]$ can be written as

$$\begin{aligned} &\min_{x \in \mathbb{R}} f(x) = x^2 - 3x - 5 \\ \text{s.t. } &h_1(x) = x - 20 \leq 0 \\ &h_2(x) = 10 - x \leq 0 \end{aligned} \tag{3}$$

Writing it in the above way makes things a lot easier to work with as it will make the algebra simpler, trust me. We can then define a new function

$$Q(x) = \max_{\mu_1 \geq 0, \mu_2 \geq 0} L(x, \mu_1, \mu_2) = \max_{\mu_1 \geq 0, \mu_2 \geq 0} f(x) + \mu_1 h_1(x) + \mu_2 h_2(x).$$

The function $L(x, \mu_1, \mu_2) = f(x) + \mu_1 h_1(x) + \mu_2 h_2(x)$ is also known as the Lagrangian. Now if x is no longer in $[10, 20]$, then $h_1(x) > 0$ and $h_1(x) > 0$. Since we are maximising the Lagrangian over μ_1, μ_2 where $\mu_1 \geq 0, \mu_2 \geq 0$, then we can set $\mu_1 = \mu_2 = \infty$ so $Q(x) = \infty$. If x is in $[10, 20]$, then $h_1(x) \leq 0$ and $h_1(x) \leq 0$. Then we can set $\mu_1 = \mu_2 = 0$ so $Q(x) = f(x)$. Thus,

$$Q(x) = \begin{cases} f(x) & x \in [10, 20] \\ \infty & x \notin [10, 20] \end{cases}.$$

Hence, $Q(x)$ can be considered as a very strict penalising function. Any value of x that steps out of bound for even the slightest is immediately given infinite cost or punishment. It is clear that any minimiser of $Q(x)$ also solves the constraint optimisation problem of $f(x)$. This can be shown by contradiction, suppose that the minimiser of $Q(x)$ is not the solution to the constraint optimisation problem of $f(x)$. Let $x^* \in \arg\min Q(x)$, then there exist $y \in [10, 20]$ such that

[1] A simple function is a linear combination of characteristic/indicator functions. This approximation is the critical insight behind the Lebesgue integral.

$f(y) < Q(x^*)$. Since $Q(x) = f(x)$ when $x \in [10, 20]$, thus $Q(y) = f(y) < Q(x^*)$. Hence, $x^* \notin \arg\min Q(x)$. This is a contradiction. Hence, minimiser of $Q(x)$ also minimises $f(x)$. This shows that heuristic approach is a good one. That is great and all but how do we even minimise $Q(x)$?

Well let us first take and approximation again and see where that leads us, let us approximate then approximate our original optimisation problem by switching the max and min around as such

$$\min_{x\in\mathbb{R}} \max_{\mu_1\geq 0,\mu_2\geq 0,} L(x,\mu_1,\mu_2) \approx \max_{\mu_1\geq 0,\mu_2\geq 0,} \min_{x\in\mathbb{R}} L(x,\mu_1,\mu_2)$$

Why do we want to do this? Because we want to move our unconstrained minimisation problem inwards before doing the slightly trickier constraint problem, which makes solving our new optimisation problem slightly more straightforward than our original primal problem. Our initial problem is then called the primal problem, and our approximation is called the dual problem. In the above, we claim that the solutions to the primal and dual problems approximate each other well.

This begs the question of how well our solution of the dual problem approximates the optimal solution for the primal solution. Before we answer, let us first consider this other question, what is the relationship between the two optimisation problems. To answer that, we need to use the famous min-max inequality[2]

$$\min_{x\in\mathbb{R}} \max_{\mu_1\geq 0,\mu_2\geq 0,} L(x,\mu_1,\mu_2) \geq \max_{\mu_1\geq 0,\mu_2\geq 0,} \min_{x\in\mathbb{R}} L(x,\mu_1,\mu_2)$$

We see from this inequality that the dual problem solution lower bounds the primal solution. This is known as weak duality. The difference between the primal and dual solutions is the duality gap. If the primal and the dual solution are equal, i.e. there is no duality gap, we call that strong duality. Therefore, it is of interest for us to check if strong duality holds because it means that our approximation is not just good but exact. One such condition for strong duality to occur is Slater's condition. Slater's condition states that for the following convex optimisation problem

$$\begin{aligned} &\min_{x\in\mathbb{R}} f(x) \\ \text{s.t. } & h_i(x) \leq 0 \ , \ i = 1, \ldots, m \\ & g_j(x) \leq 0 \ , \ j = 1, \ldots, k \end{aligned} \tag{4}$$

where $f(x)$, $h_i(x)$ and $g_j(x)$ are convex functions, strong duality holds if there exists an x^* such that x^* is strictly feasible i.e there exist a feasible point that doesn't lie on the edge of any constraint. There are many such conditions, but

[2] I am fully aware that in actuality, we are dealing with sup and inf but for the sake of some non-mathematical readers, let us assume the regularity conditions to use min and max hold.

I won't cover the rest of them. My goal is to understand how to interpret and understand instead of solving such optimisation problems. In part because I believe there are many optimisation packages and blogs that tell you how to do the latter but not how to do the former. Nonetheless, I will complete the initial optimisation problem I posed before delving into the interpretation of weak duality.

Now lets look at our initial problem again in equation (3), we see something very interesting. The first is that $f(x)$, $h_1(x)$ and $h_2(x)$ are all convex. It is also fairly clear that $12 \in [10, 20]$ such that $h_1(x) < 0$ and $h_2(x) < 0$ i.e 12 is strictly feasible. Thus, Slater condition holds. That means strong duality holds. Hence, to solve the dual problem we first solve the unconstrained problem

$$\min_{x\in\mathbb{R}} L(x, \mu_1, \mu_2) = x^2 - 3x - 5 + \mu_1(x - 20) + \mu_2(10 - x) \tag{5}$$

We can solve this by first noticing it is a strict convex unconstrained optimisation problem and the turning point is $x = \frac{\mu_2-\mu_1+3}{2}$.Then substituting this value into back into the equation above, now we need to maximise the following

$$\begin{aligned}
&\max_{\mu_1\geq 0 \mu_2\geq 0} \min_{x\in\mathbb{R}} L(x, \mu_1, \mu_2) \\
&= \max_{\mu_1\geq 0 \mu_2\geq 0} \left(\frac{\mu_2 - \mu_1 + 3}{2}\right)^2 - 3\left(\frac{\mu_2 - \mu_1 + 3}{2}\right) - 5 \\
&+ \mu_1\left(\left(\frac{\mu_2 - \mu_1 + 3}{2}\right) - 20\right) + \mu_2\left(10 - \left(\frac{\mu_2 - \mu_1 + 3}{2}\right)\right)
\end{aligned} \tag{6}$$

This is not very easy to solve as you have 2 variables, so the algebra gets a little hairy. It is much more easier to usually let a computer algebra system like sympy solve this. However, if we want to do the algebra ourselves we can make our lives a bit easier by only considering a single constraint initially i.e solve the following first

$$\min_{x\leq 20} L(x, \mu) = x^2 - 3x - 5 + \mu(10 - x) \tag{7}$$

Then, considering what happens if $x = 20$ and $x < 20$. We see that if $x = 20$

$$\begin{aligned}
&\max_{\mu\geq 0} \min_{x\in\mathbb{R}} L(x, \mu) \\
&= \max_{\mu\geq 0} (20)^2 - 3\,(20) - 5 \\
&+ \mu\,(-10) \\
&= 335 \text{ as } \mu = 0 \text{ since we are dealing with a linear function}
\end{aligned} \tag{8}$$

Similarly, if $x < 20$ then the solution for (7) is $x = \frac{\mu+3}{2}$ by considering the turning points as before, thus

$$\begin{aligned}
&\max_{\mu\geq 0}\min_{x\in\mathbb{R}} L(x,\mu)\\
&=\max_{\mu\geq 0}\left(\frac{\mu+3}{2}\right)^2-3\left(\frac{\mu+3}{2}\right)-5\\
&+\mu\left(10-\left(\frac{\mu+3}{2}\right)\right)\\
&=\left(\frac{17+3}{2}\right)^2-3\left(\frac{17+3}{2}\right)-5\\
&+\mu\left(10-\left(\frac{17+3}{2}\right)\right)\\
&\text{this problem can be solved using calculus or using properties of quadratics}\\
&=(10)^2-3\,(10)-5\\
&=65
\end{aligned}\tag{9}$$

Thus, the optimal value of x is $x = 10$.

This feels like a lot more work. In fact, it is a lot more work than just considering the endpoints and turning points for our simple problem. The beauty and utility of this approach come when considering more than one dimension and numerous constraints where the number of endpoints grows exponentially with the number of constraints we consider. In that setting, converting using a Lagrangian dual approach makes more sense. However, hopefully by working through this exercise you see how to solve constraint optimisation using the Lagrangian dual setup. You can try practising this by considering other quadratics or convex functions, different constraints or even high dimensions.

Besides making life slightly easier by converting a constraint optimisation to a more familiar unconstrained problem or a form where we can do algebra a lot easier, Lagrangian duals are more often used in research to provide a different perspective of the same problem. For example, in Support Vector Machines (SVMs) the primal problem is

$$\begin{aligned}
&\min_{\boldsymbol{\theta},\theta_0,\zeta}\frac{1}{2}\left\|\boldsymbol{\theta}\right\|^2+C\sum_{t=1}^{n}\zeta_t\\
\text{s.t. }\;&y_t(\boldsymbol{\theta}^T\boldsymbol{x}_t+\theta_0)\geq 1-\zeta_t,\\
&\zeta_t\geq 0
\end{aligned}\tag{10}$$

but the dual problem is

$$\begin{aligned} &\max_{\alpha} \sum_{t=1}^{n} \alpha_t \frac{1}{2} \sum_{s=1}^{n} \sum_{t=1}^{n} \alpha_s \alpha_t y_s y_t \boldsymbol{x_s}^T \boldsymbol{x_t} \\ \text{s.t. } &\alpha_t \in [0, C] \quad \forall t \in \{1, \dots, n\}, \\ &\sum_{t=1}^{n} \alpha_t y_t = 0. \end{aligned} \tag{11}$$

this allows us to see that there is a dot product taking place and we can replace that dot product with a kernel allowing us to generalise to infinite dimensional feature space. This is the essence of the kernel trick used in SVMs.

Another way to gain insight from the dual formulation of constraint optimisation is by interpreting the coefficient λ mean. In microeconomics, when maximising profits, λ is the shadow price where a particular regulation such as a quota is the constraint. In thermodynamics, when maximising the entropy of a given system, λ is the temperature where the total energy is the constraint of the system. If strong duality holds, we can typically use λ to tell how changing the constraint will affect the system. This is known as sensitivity analysis.

Furthermore, interpreting the duality gap can also be a fruitful endeavour. The max-min formulation I gave above would remind a reader familiar with game theory of saddle-points in a min-max game. Indeed such an interpretation is, in fact, valid, and one can think of the Lagrangian as a min-max game. From that perspective, the duality gap is a player's advantage of going first. If there is no advantage of going first, there is no duality gap.

Lastly, the idea of Lagrangian duality is not restricted to constraint optimisation problems. It can also be used for unconstrained optimisation problems by adding new dummy variables and equality constraints, allowing us to find the dual to the new constraint optimisation problem. This approach is used in LASSO Dual algorithm.

4 Conclusion

These examples would show that thinking about the dual of an optimisation problem is a fruitful line of thinking, whether it is to approximate an intractable or complex to solve optimisation problem or trying to understand the importance of certain constraints of the system. Duality is a valuable tool in a researcher toolbox to uncover new insights and generalise algorithms to solve problems. Hopefully, you learnt something from reading this.