Basic Mathematical Optimisation in Economics

By Shrimray Bikash Saikia, Raffles Institution

Abstract

It is not money that makes the world go around, but optimisation. In this essay, we introduce the reader to the powerful technique of mathematical modelling before expounding more specifically on univariate calculus-based optimisation. Then, we explore two applications of such basic optimisation in economics. Throughout the essay, analogies and graphical illustrations are used to explain difficult concepts more easily to readers while the appendices contain further reading for interested advanced readers.

(72 Words)

Introduction

Have you ever asked yourself if you were spending your money in the best way possible? This fundamental question transcends many different contexts in economics: be it microeconomic or macroeconomic. For instance, big businesses constantly wish to decrease costs and increase revenue so as to increase profits. Students wish to maximise grades and decrease effort taken to improve quality of life.

Such a problem, where we want to maximise a value and minimise another to produce the "best" outcome, are termed optimisation problems. Optimisation problems are among the most important in Economics, but also the hardest to solve – examples include the famous Travelling Salesman Problem.

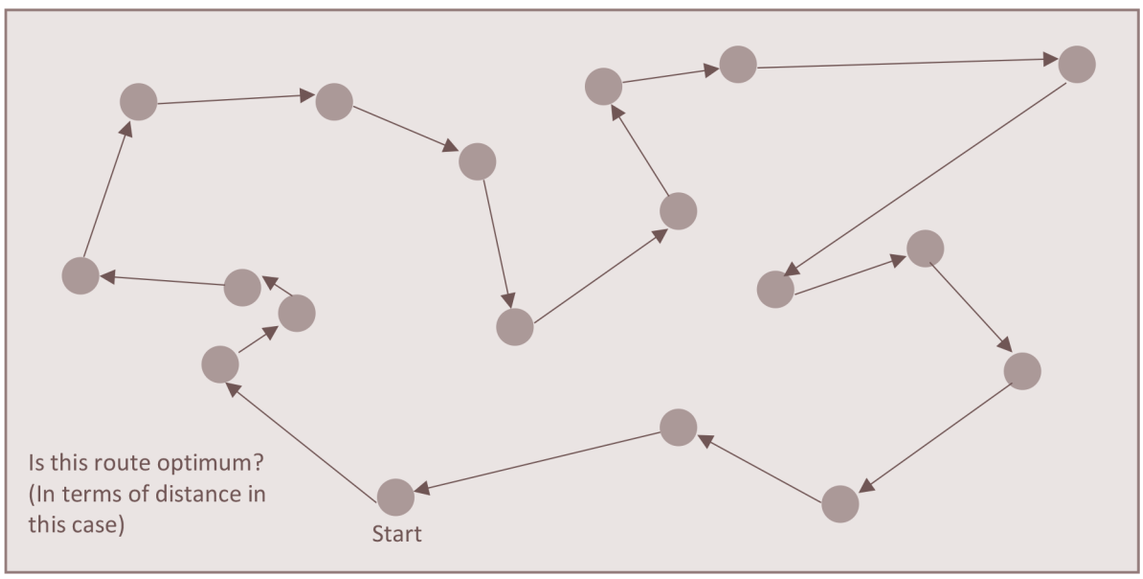

Fig 1: Illustration of the Travelling Salesman problem. How do you find the shortest path that visits each of the cities, represented here as red dots, for any n cities? This optimisation problem has applications in many fields such as Logistics, Genetics and Astronomy. This instance of n=18 is left as an exploration for the reader.

More generally, we can optimise an n-variate function to find its maximum/minimum by in turn classifying candidate solutions by finding the determinant of its Hessian matrix as shown below:

$$H = \begin{vmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{vmatrix}$$

Linking back to the discussion of marginal quantities in the essay earlier, companies often have to quantify intangible quantities when making decisions – which our earlier approach could not mitigate.

For example, a production manager might have to approximate how much each tonne of waste gas their factory emits, costs them in terms of employee health and environmental detriment. It can also be used to estimate how much a business can at most pay for a good with limited supply—such as one hour of overtime from fully-worked employees.

The "shadow price" as such a quantity is called in economics is simple the Lagrange multiplier obtained by using the method of Lagrange multipliers for constrained optimisation.

Stated simply, this technique checks values of the objective function at pre-defined boundaries obtained from the constraints. For understanding, suppose we wish to minimise an objective function $y = 4 - x^2$ subject to $-3 \leq x \leq 4$. Then we would draw two straight lines in addition to the graph: namely $x = -3$ and $x = 4$. These are the boundaries of our problem. Then, simple inspection shows the solution to be $y = -12$ occurring at $x = 4$.

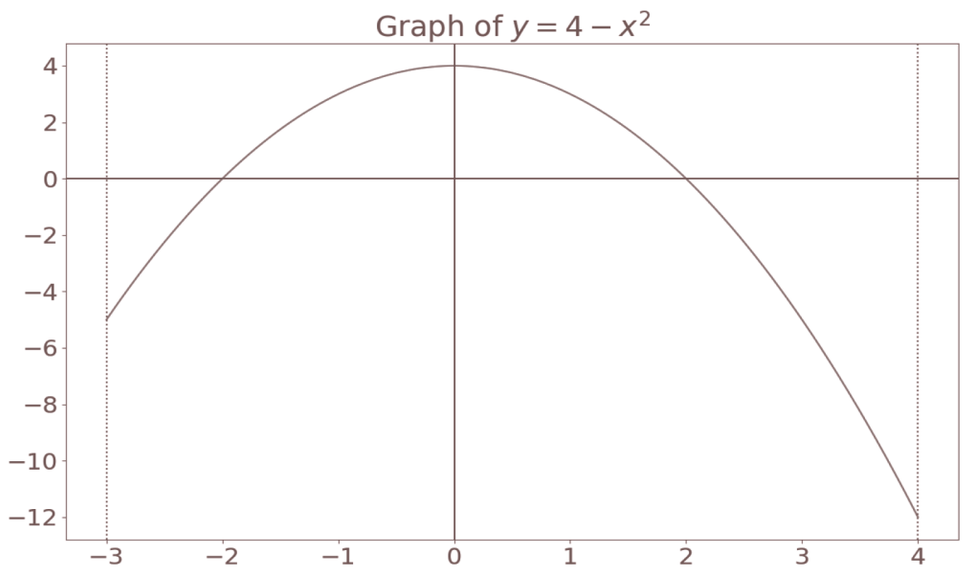

Fig 11: Graph for above discussion.

More specifically, suppose we implicitly define a boundary curve as $g(x,y) = k$ for some real constant k. Then to optimise $f(x,y)$ subject to $g(x,y)$, we instead consider the Lagrangian Function.

$$F(x,y,\lambda) = f(x,y) + \lambda[k - g(x,y)]$$

$$Then\ solve\ \nabla F(x,y,\lambda) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

To determine if a solution to the above equations is a local maximum or minimum, we use the determinant of the bordered Hessian, similar to but not to be confused with the Hessian earlier. A significant advantage of this method is that it simplifies many difficult constrained optimisation problems at higher dimensions into a form much easier to compute and solve.

Interestingly, the Lagrangian method can be considered as a generalisation of isoquants, as discussed earlier, to higher dimensions. Similar to how maximising level of production for a given budget involved finding the isoquant tangent to the budget line, the method of Lagrange multipliers involves finding the contour line of a function tangent to the boundary curve.

Visually speaking, one can consider the scenario as shown below. The contour line of a function can be considered analogous to contour lines in Geography. Just as how a contour line in Geography denotes a line of equal height in terrain, a contour line of the objective function represents the locus of points yielding the same output value subject to the objective function $f(x,y)$.

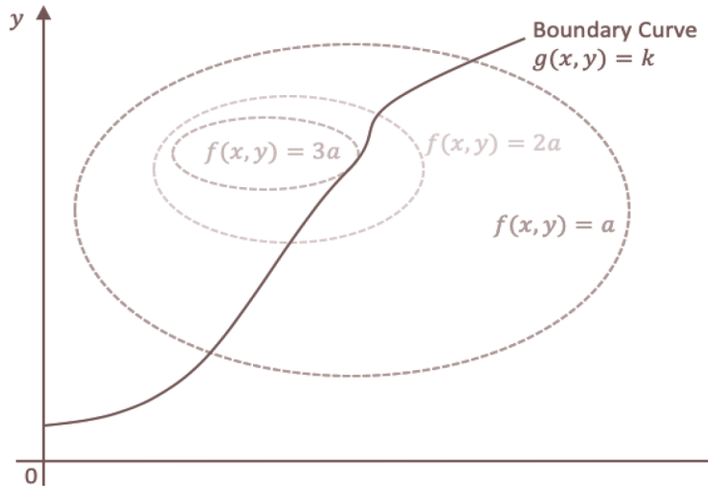

Fig 12: A sketch to represent this discussion. The dotted lines represent the contour lines. The solid black line represents the boundary curve. In this case, the intersection between the blue contour line $f(x,y) = 3a$ and boundary curve represents a potential local maximum. In this case where a is a real positive number, the function is actually concave because for a fixed increase in value of a the distance between successive contour lines decreases.

Building on this analogy, one can also notice how convexity and concavity are also indicated by the trend in distances between contour lines – just as how a convex or concave slope in terrain is indicated by Geographical contour lines.

Appendix B: Non-Concavity

Another constraint we placed on the content in the essay was that we only considered convex optimisation. However, only a small fraction of functions are ever convex or concave entirely, just as a very small number of functions are monotonic (constantly increasing or constantly decreasing). For instance, even a simple non-zero cubic function is non-concave.

One option to apply concave/convex analytical solving methods to such non-concave functions would be to simply split the graph up into regions that are purely convex or concave and then solve. While this technique is still feasible at smaller dimensions, at higher dimensions that we cannot represent easily in our 3-Dimensional world, this method simply fails.

Furthermore, it has been shown that finding out if a function is concave or non-concave, for functions with integral exponents, is NP-hard (Ahmadi, Olshevsky, Parrilo & Tsitsiklis; 2011); where NP-hard means that this problem is at least as hard as an NP problem where NP stands for Non-deterministically polynomial.

Such problems' solutions can be verified easily but finding a solution is difficult – similar to how verifying Mathematical proofs is relatively easy compared to actually writing a proof. Thus, even the most advanced supercomputers cannot solve an NP-hard problem such as this one in a reasonable amount of time (unless it is a true quantum computer). Hence, such a method to solve non-concave optimisation problems is clearly not feasible.

However, if we allow for a certain degree of error in the answer, we can also use heuristics. Similar to how a Lagrange multiplier value can give an approximation of an intangible quantity's value, a heuristic such as gradient descent can allow us to find the solutions to a non-concave optimisation problem in a reasonable amount of time.

Simply speaking, gradient descent involves iteratively changing the input parameters so as to move in the direction of steepest descent of the function (ie. antiparallel of the gradient). However, teething problems with this method are defining the step size – how much to iteratively change the x-value by every time; and determining if the found solution is really the global extremum (especially in the non-concave case).

Hence, much more advanced methods such as cubic regularisation and random stochastic gradient descent have been developed over the years. Without going too much into detail, we refer you to the figures on the next page if you wish for a pictorial understanding of gradient descent.

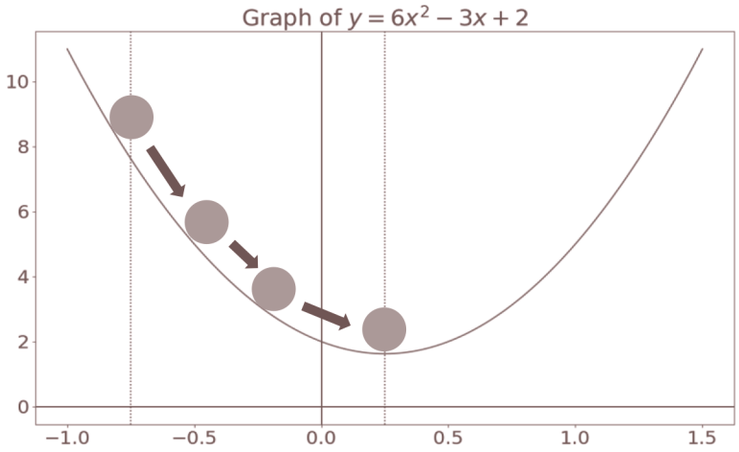

Fig 13: Imagine placing a ball at $x = -0.5$ on a ramp shaped like this function. Supposing the ramp is infinitely long, eventually, the ball will roll down to $x = 0.25$ (the minima) under the force of gravity. The gradient descent function also works in a similar way by "moving down" under the influence of gravitational pull.

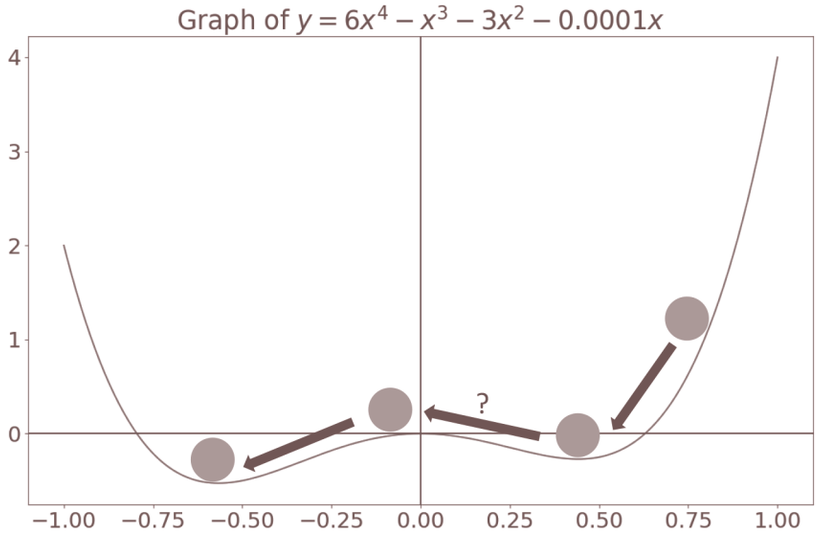

Fig 14: However, if we start our ball at x=0.75 in this case, it might not climb over the local maximum at x=0 to reach the global minimum to the left of the origin. Instead, it might get stuck at the local minimum to the right of the origin depending on the step size. However, a possible solution is to define the ball to have a certain "momentum" so it can reach the global minimum if approaching from the right. Of course, fine-tuning this momentum and step size is still key to successful utilisation of Gradient Descent.

Interestingly, as testament to the versatility of Mathematics and particularly Mathematical Modelling, similar gradient descent methods form the basis of Machine Learning. In such applications, the objective function is to minimise the error between the model's output and expected output, as expressed in terms of the model's parameters such as weight.

In such a context, fine-tuning the learning rate (step size) is still crucial to an effective gradient descent and hence effective model.

References/Credits

Ahmadi, A. A., Olshevsky, A., Parrilo, P. A., & Tsitsiklis, J. N. (2011). NP-hardness of deciding convexity of quartic polynomials and related problems. *Mathematical Programming, 137*(1-2), 453-476. doi:10.1007/s10107-011-0499-2

Dowling, E. T., & Dutch, K. (2006). *Mathematical economics ; based on Schaums Outline of theory and problems of introduction to mathematical economics, by Edward T. Dowling.* New York: McGraw-Hill.

Fig 10 was drawn using Python code adapted from online source.
https://commons.wikimedia.org/wiki/File:Saddle_point.svg

Mathematical Modelling

Mathematical Modelling is a powerful method we can use to solve optimisation problems by expressing them mathematically. First, we translate key quantities (e.g. price, cost, profit) in our problem into algebraic variables. Next, we form the objective function – the quantity we wish to maximise/minimise expressed in terms of these variables. Then, we form constraints – boundaries/relations among the variables in the form of equalities/inequalities. Finally, we solve the objective function (try to maximise/minimise it) according to the constraints.

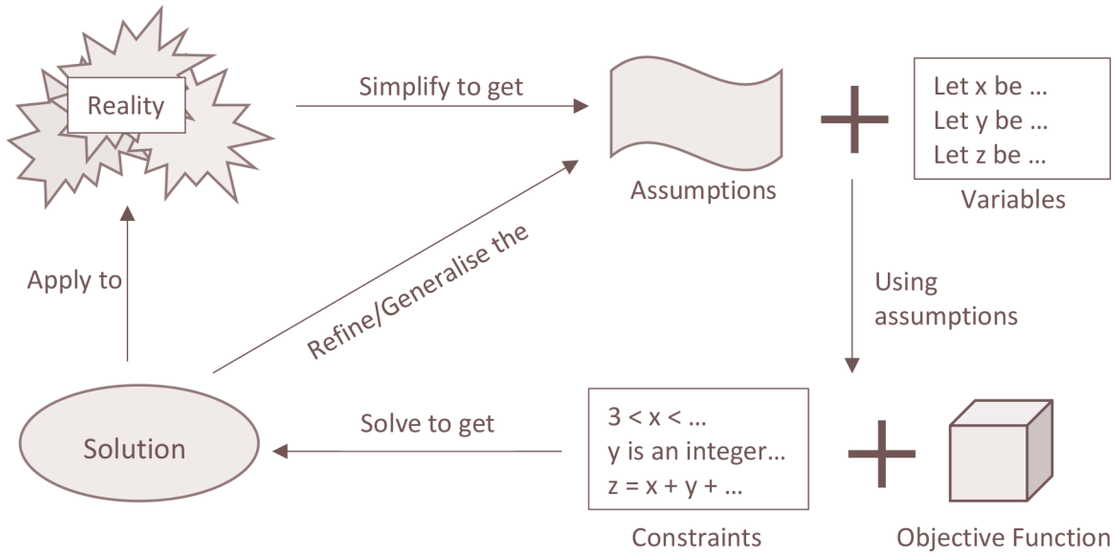

Fig 2: A diagrammatic summary of the Mathematical modelling process for optimisation.

When we solve the objective function, there are 3 distinct possibilities:

1. Feasible. There exists a solution satisfying all constraints.
2. Infeasible. There does not exist any such solution.
3. Unbounded. There is no one most optimal solution

Even if we do not find a solution, we could try simplifying the problem. For instance, we could hold one variable constant or refine our problem by adding another constraint and then generalising our solution from there.

George B. Dantzig and Leonid Kantorovich first initiated the field of modern optimisation in the 1940-1950s with linear programming where the above-mentioned objective functions and constraints were all purely linear.

However, just as a single colour cannot paint a masterpiece, linear functions themselves cannot represent diverse and complex relations between variables. Hence, with improved computational power, non-linear programming has become more widely used since it offers greater fidelity in mathematical models.

Concavity

In this essay, we deal with a special class of non-linear programming: convex optimisation. Consider some real, differentiable, continuously defined function $y = f(x)$. We deem this function convex over the interval $[a, b]$ if $f''(x) > 0$ and concave if $f''(x) < 0$.

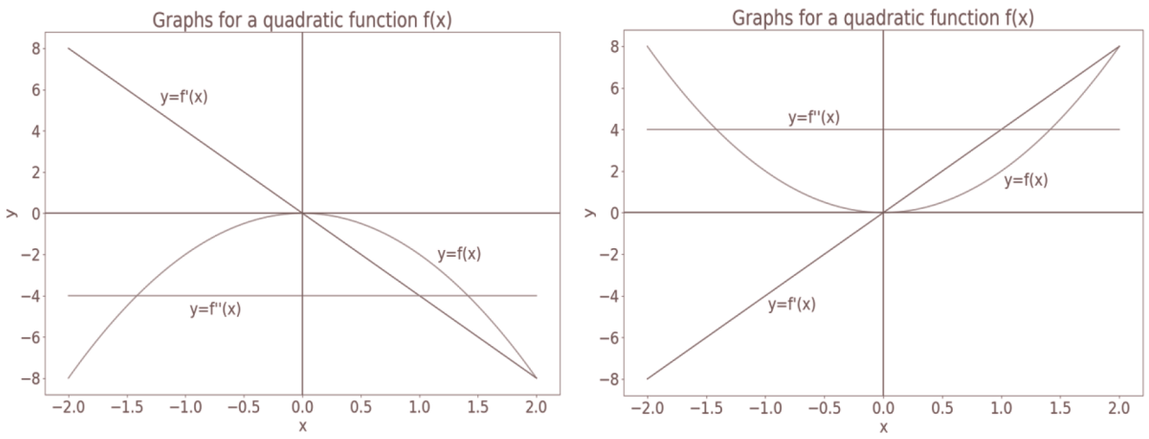

Fig 3a (left) and 3b (right): The graph in 3a is concave and the graph in 3b is convex over the real interval [-2,2]. In fact, a quadratic graph is convex over the real domain if the coefficient of x^2 is positive and concave if the aforementioned coefficient is negative.

Thus, if a function is convex/concave over the whole real domain, there is only one global minimum/maximum respectively. More commonly however, a function has both regions of convexity and concavity with multiple relative minima/maxima.

To determine the minima/maxima of such a function, we use the necessary not sufficient condition of finding points where the first derivative of the function is 0 or undefined. These are critical points which at $x = a$ each can be a:

1. Relative minimum if $f''(a) > 0$
2. Relative maximum if $f''(a) < 0$
3. Inflection point where concavity of the function changes (ie. from concave to convex or vice versa); if and only if $f''(a) = 0$ or $f'(a)$ is undefined
4. Undefined critical point otherwise

We can then find the maximum/minimum of greatest magnitude which is the global maximum/minimum, the optimal solution. Interestingly, the duality principle holds that a convex minimisation problem has an equivalent concave maximisation form, subject to certain conditions.

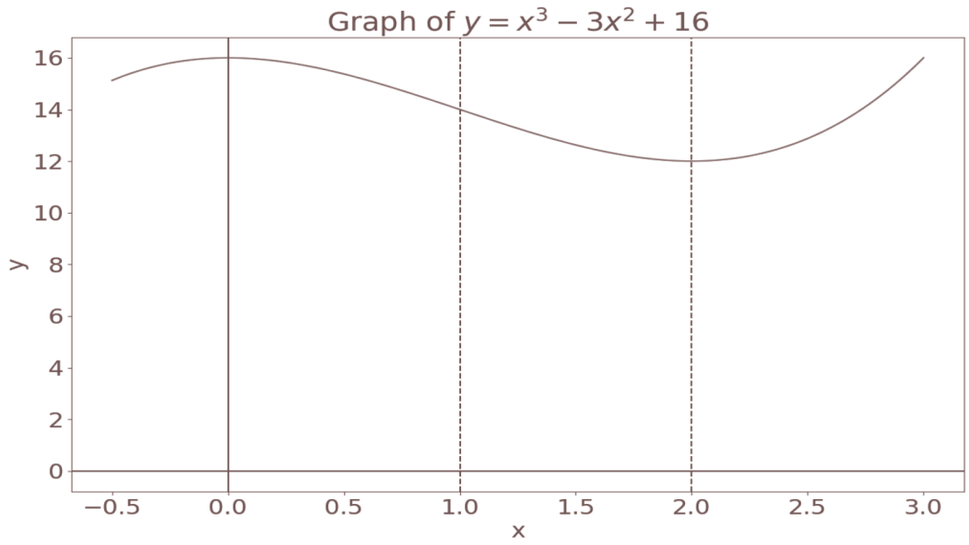

Fig 4: Considering $y = x^3 - 3x^2 + 16$ over the real interval $[-0.5,3]$, we find that a relative maximum occurs at $x = 0$, an inflexion point occurs at $x = 1$ and a relative minimum occurs at $x = 2$.

Law of Diminishing Returns

Many decisions we make boil down to simple cost-benefit analyses, for example, we may choose to not eat an ice-cream if the health detriments of doing so outweigh any mental pleasure we derive from it.

Businesses can also use a similar framework to decide on how to change production to maximise their revenue. Marginal cost (MC) and marginal revenue (MR) are defined as the extra cost and revenue of an additional unit of production, respectively. If the business had the means to, production should be increased when $MR > MC$ as this results in a net increase in revenue; and decreased when $MC > MR$.

If we define TC (Total Cost) and TR (Total Revenue), to be functions in x, the number of products manufactured, we have:

$$MC = \frac{d}{dx}TC$$

$$MR = \frac{d}{dx}TR$$

$$Marginal\ Return = MR - MC$$

$$MR > MC \iff MR - MC > 0$$

Clearly, to maximise our revenue, we should keep on increasing production (if we have the means) until the point where $MC > MR$. In other words, we are finding the inflection point of the graph of $y = TC - TR$ or expressed differently, $y = TP$ where TP is total profit.

Suppose $TP(x) = x^3 - 8x^2 + 17x - 3$. Then, $MP(x) = \frac{dTP(x)}{dx} = 3x^2 - 16x + 17$. Consider mean product (AP) as well where $AP(x) = \frac{TP(x)}{x} = x^2 - 8x + 17 - \frac{3}{x}$. Graphing these functions over the real interval $[0.19\,, 4.5]$ we have:

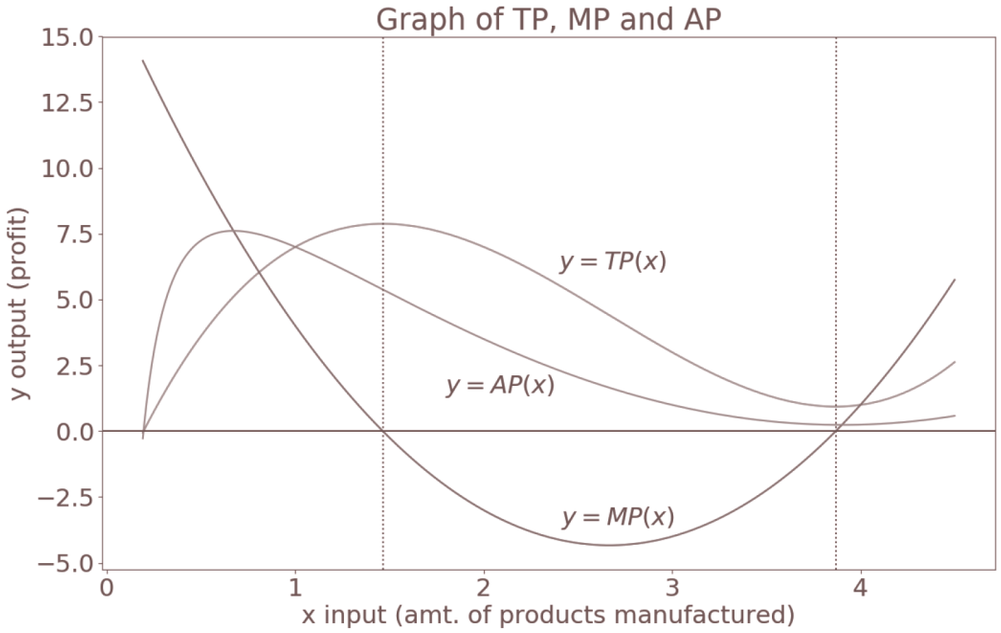

Fig 5: Graph of aforementioned functions over $[0.19,4.5]$

Observe that when MP crosses the x-axis (ie. $MP = 0$), TP reaches a relative maximum/minimum while AP is negligibly close to an inflexion point.

More importantly, observe how as x increases from 0.19 to $\frac{8}{3}$, the marginal profit actually decreases. This is an important law in economics known as the law of diminishing returns which states that at some point, increasing input might decrease marginal profit rather than increase it. For instance, if more workers are added to an already crowded factory, there will only be more congestions and inefficiencies compared to extra production. Thus, it is crucial for businesses to be wary of this region and aim to increase input in a big step so as to avoid this region altogether.

Cost Minimisation using Isoquants

For a business deciding how two or more factors affects a given output, for instance labour and cost affecting production, the isoquant is a simple optimisation method. The isoquant comes from “iso” meaning same; and “quant” meaning quantity. It is a locus of points which each represent a particular combination of input A and input B resulting in the same amount of output – for instance labour and capital affecting production.

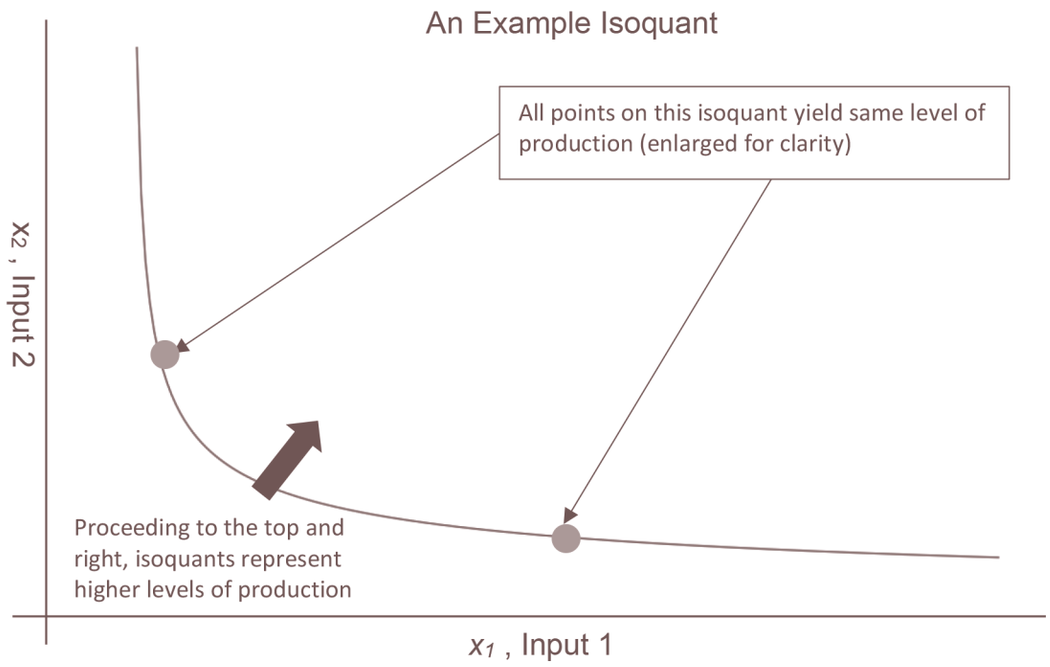

Fig 6: An example isoquant. Note the captions in the figure.

When we optimise production for a given budget, we aim to maximise output for a given amount of input we can use. Graphically speaking, we want to find a point on the budget line which lies on the isoquant of highest possible production. Considering a factory, suppose due to physical limitations the two inputs must fulfil the constraint $x_1 + x_2 \leq a$ for some constant a. A rational firm would consider the case $x_1 + x_2 = a$ so as to maximise the level of production. Then the budget line is simply the locus of all points fulfilling this equation.

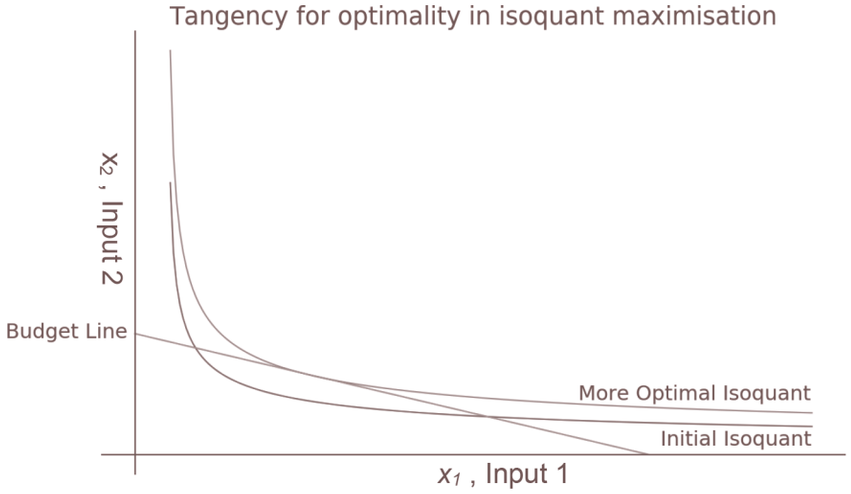

Fig 7: Graphically showing that for an isoquant to give the maximal level of production for a given budget line, the budget line should be tangent to the isoquant.

Investigating, we find that at such a point, the budget line must be tangent to the isoquant. Else, we could choose a point between the two intersections which is on a higher isoquant. Note that this method generally only works for the case of convex isoquants.

In fact, the convexity of an isoquant is related to how easily the business can substitute one input for the other, keeping the production amount the same. If both inputs are perfect substitutes, the isoquant is a straight line and using one is as good as using the other, production-wise (see Fig 8). There could also be a case where the inputs simple cannot be substituted; and the isoquant would simply be one point (see Fig 9).

Mathematically, how well one input can be substituted for the other is known as elasticity of substitution – related to how steeply the isoquant curves.

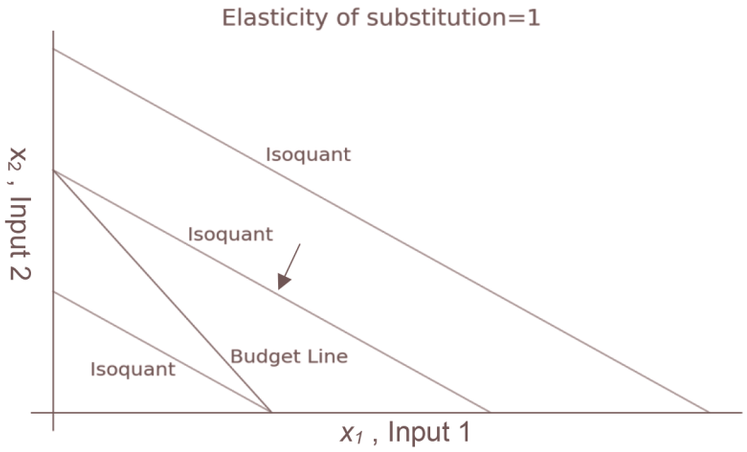

Fig 8: Isoquants when elasticity of substitution is 1. Please note that diagram is not to scale. In this case, we choose the isoquant labelled with the arrow to maximise the level of production as it is the highest isoquant the budget line intersects in the first quadrant.

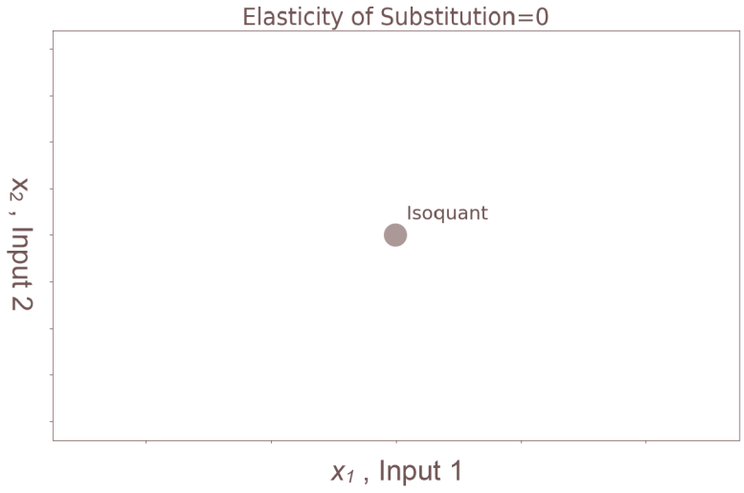

Fig 9: Isoquant (or point) when elasticity of substitution is 0. This is because the two inputs cannot be substituted at all, thus you cannot have any other point yielding the same level of production. Point has been enlarged for clarity.

Another important economic insight is returns to scale. As consumers, we might be familiar with economies of scale – when we buy a product in bulk, price per unit product is lower. Similarly, but from the perspective of the firm, increasing returns to scale means that a given increase in inputs (e.g. labour and capital) results in a proportionally larger increase in output.

We can observe if a long-term increase in input will cause an increasing, decreasing or constant return to scale from the isoquant map (isoquants of different production levels drawn on the same set of axes).

Increasing the production level by a fixed step every time, if the distance between successive isoquants decreases, then, there is a long-term increasing return to scale since a smaller successive increase in budget is needed to increase the production by the same amount. Following a similar line of reasoning, increasing distance between successive isoquants indicates a long-term decreasing return to scale and constant distance between successive isoquants indicates constant return to scale.

Ideally, businesses would aim for increasing returns to scale which can be achieved, in turn through optimisation of production methods, so that marginal revenue is greater than marginal cost of production at any point – put more informally, even companies love a “buy one, get one free” offer.

Conclusion

In the short span of around 1000 words, we have thus covered the basics of mathematical optimisation. However, we have barely scratched the surface of this vibrant and relevant topic. Economics may be considered a social science, but Mathematical modelling lies at its heart, just like many other diverse fields ranging from Physics and Machine Learning all the way to Genetics and Engineering.

In ending off, the author hopes the readers enjoyed reading this essay just as much as he enjoyed writing it and that the marginal utility of reading the appendices behind outweighs the marginal cost. Thank you.

Appendix A: Multivariate Optimisation

In the essay, we considered only univariate functions where one independent variable affects one dependent variable only. However, multivariate functions offer greater fidelity since they are able to show how, many dependent variables might affect the output. For instance, a downturn in the tertiary hospitality industry could be due to a fuel shortage which hurt both primary and secondary industries, creating a vicious cycle of negative feedback.

We can use a similar approach for multivariate functions as we did for univariate functions – we take the first-order partial derivatives of the functions over their respective independent variables and equate them to 0. In differentiating multivariate functions, we treat non-target variables as constants before following basic calculus rules.

Given that the gradient of a function returns us a vector pointing in the direction of steepest change, we could equivalently rephrase this process as finding all points where the gradient is a zero vector. Canonically, $\nabla f(x_1, x_2, \ldots, x_n) = 0$ for an n-variate function.

Note that this is a necessary, not sufficient condition for a point to be maximal/minimal, according to the objective function. In fact, in the bivariate case, in addition to the classes of points discussed earlier (inflection point, relative maximum etc.) there is also the unique case of the saddle point.

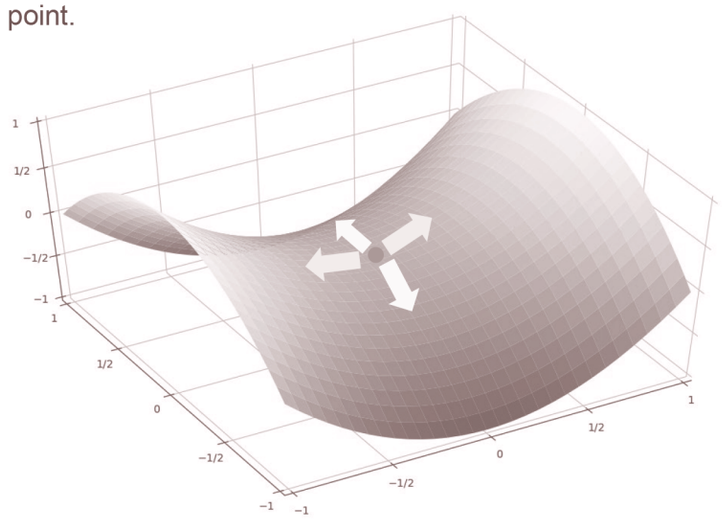

Fig 10: Observe the saddle point highlighted in red, enlarged for clarity. If we consider values in the directions of the green arrows only, the red point seems like a minimum. However, if we consider values in the directions of the yellow arrows, the red point seems like a maximum. Thus it is called a saddle point.

The graph around a saddle point is quite literally in the shape of a saddle – it is not maximum/minimum only, in both dimensions. Again, to differentiate between such points we use the second partial derivative test analogous to the second derivative test for univariate functions as shown below:

$$H = f_{xx} \cdot f_{yy} - \left(f_{xy}\right)^2$$

Where $f_{xx} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)$ for a real differentiable function $f(x, y)$. If $H > 0$ and $f_{xx} < 0$ then the point is a local maximum, else if $H > 0$ and $f_{xx} > 0$then the point is a local maximum, or if $H < 0$ then the point is a saddle point. The test is inconclusive if $H = 0$.