Optimization Foundations for Support Vector Machines

BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in

innovate achieve lead

BITS Pilani
Pilani Campus

Session 6
Date – 13/02/2022
Time – 10 to 12

Text Book(s)

T1	Christopher Bishop: Pattern Recognition and Machine Learning, Springer International Edition
T2	Tom M. Mitchell: Machine Learning, The McGraw-Hill Companies, Inc..

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Tom Mitchell and many others who made their course materials freely available online.

Topics to be covered

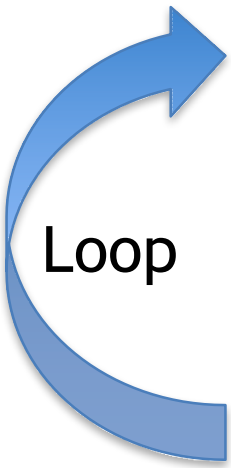
Module 6 : Optimization Foundations for Support Vector Machines

- Constrained and Unconstrained Optimization
- Lagrange Multiplier
- Primal and Dual of an optimization problem
- Quadratic Programming
- KKT conditions

Quick Recap

- KNN
- Naïve Bayes
- Logistic Regression
- Decision Tree

ML in Practice



- Understand domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learn models
- Interpret results
- Consolidate and deploy discovered knowledge

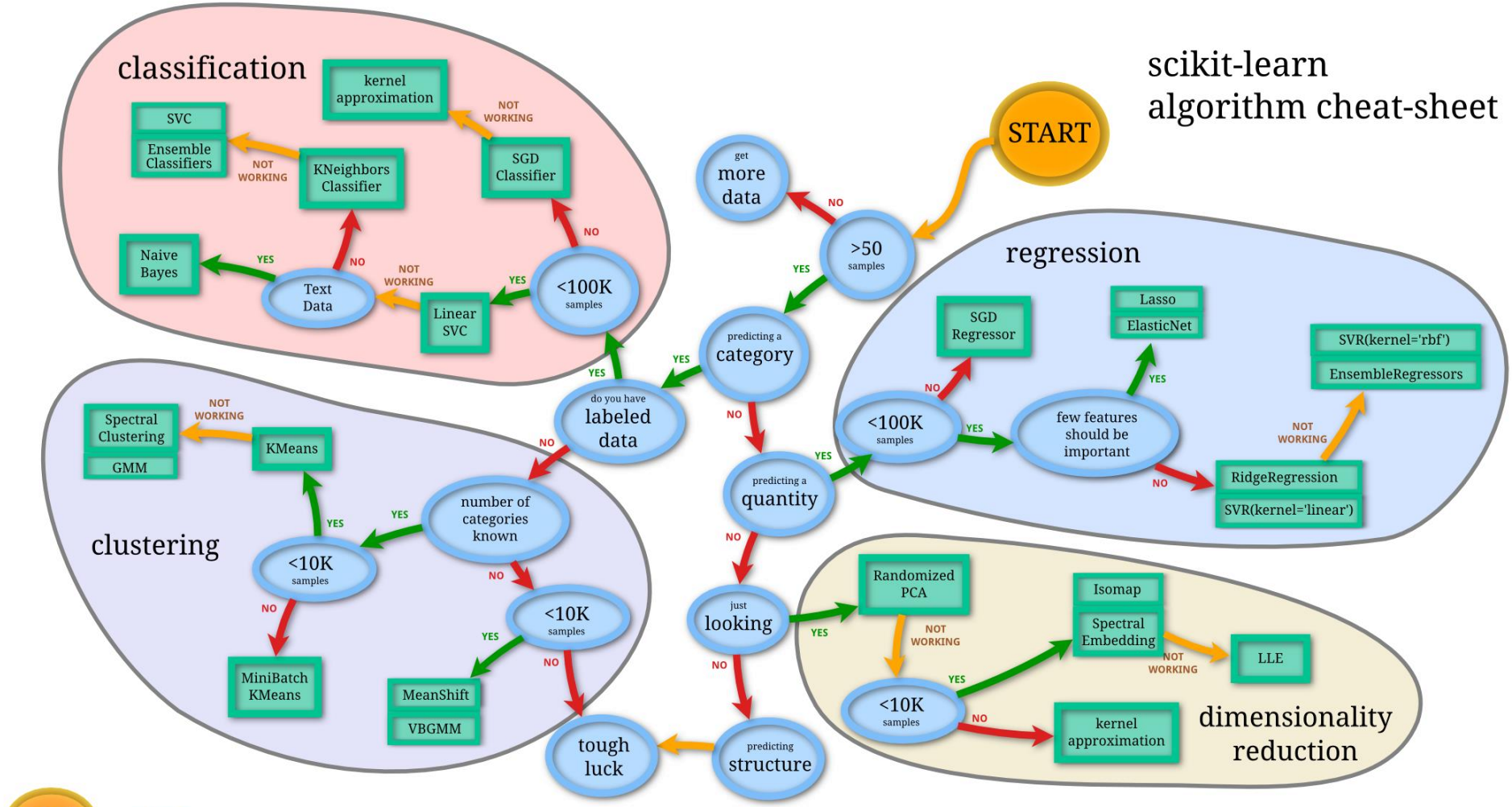
Based on a slide by Pedro Domingos

ML algorithm selection

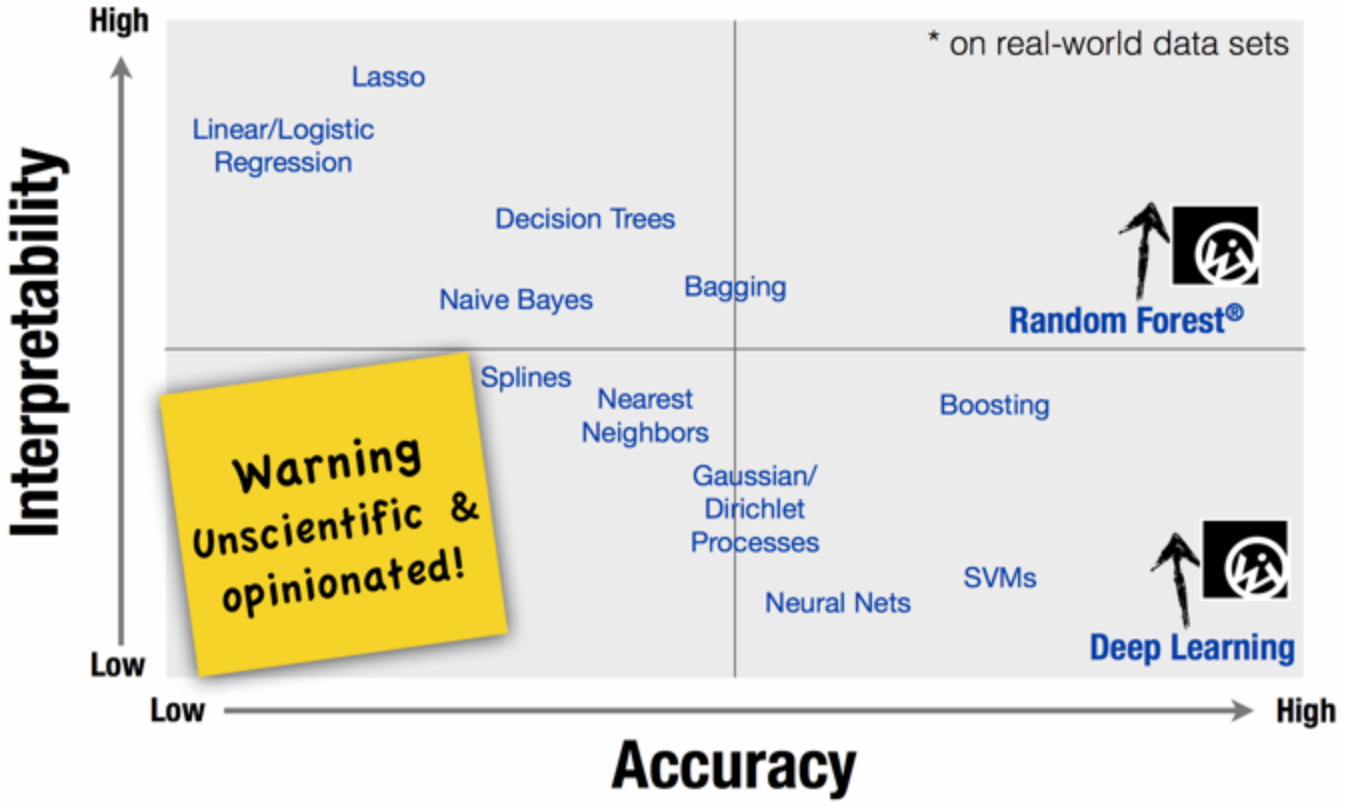
Practical advice is:

- Define a performance metric to evaluate your model
- Ask yourself: What performance score is desired, what hardware is required, what is the project deadline
- Start with the simplest model
- If you don't meet your expected goal, try more complex models

scikit-learn algorithm cheat-sheet



ML Algorithmic Trade-Off



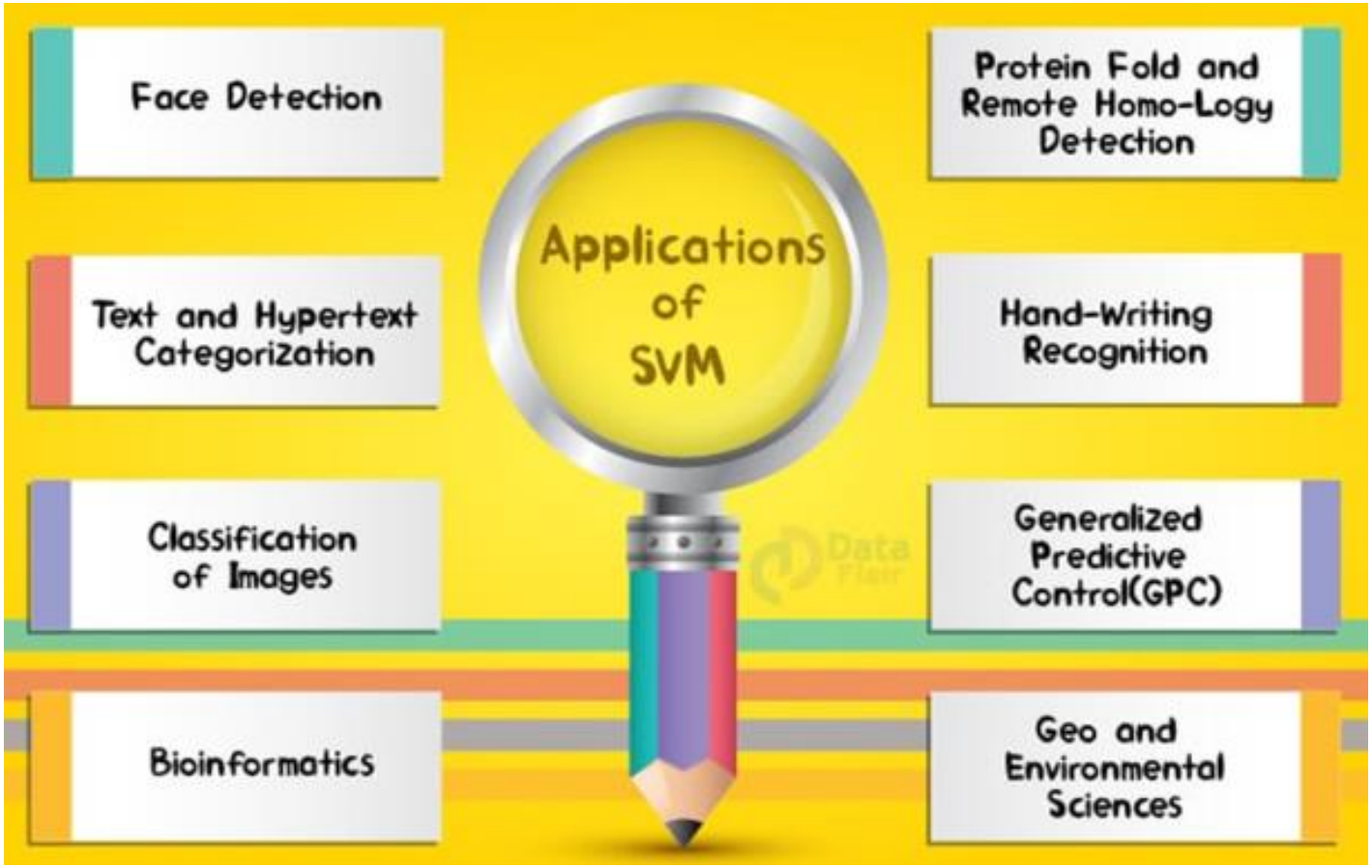
Machine Learning Model Selection

Machine Learning | Applicability

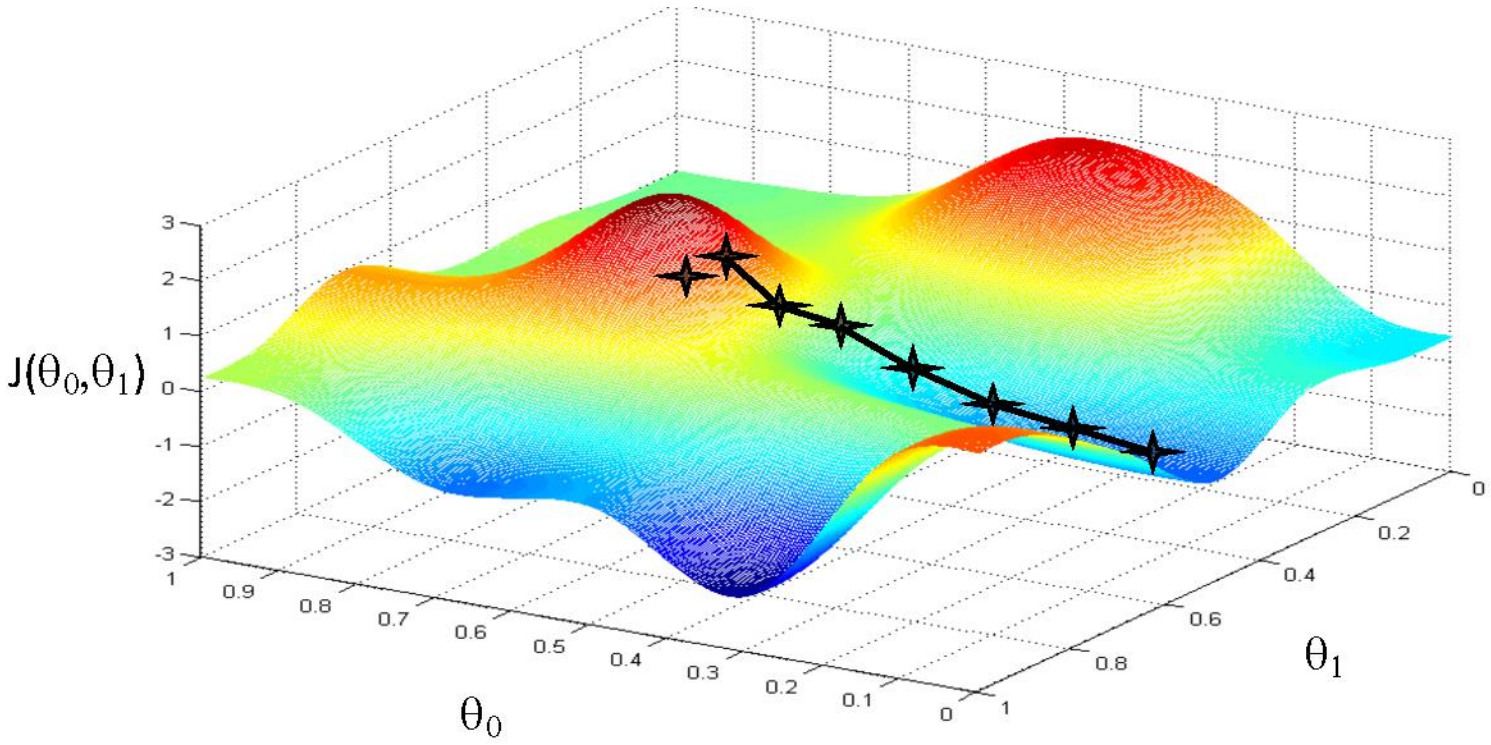
	Application	Model Type
IMAGENET	Object Localization and Image Classification	Convolutional Neural Networks (CNN), Support Vector Machines
NETFLIX RECOMMENDED TV	Collaborative Filtering, Recommendation Engines, Inputting Missing Interactions	Restricted Boltzmann Machines (RBM), ALS
	Anomaly Detection	Clustering, Decision Trees
	Forecasting or prediction of time-series and sequences like speech and video	Recurrent Neural Networks (RNN), Long-short Term Memory (LSTM), Hidden Markov Models
	Click Through Rate (CTR) Prediction	Logistic Regression
	State-Action Learning, Decision Making	Deep Q Networks (Reinforcement Learning)

SVM Applications

SVM has been used successfully in many real-world problems

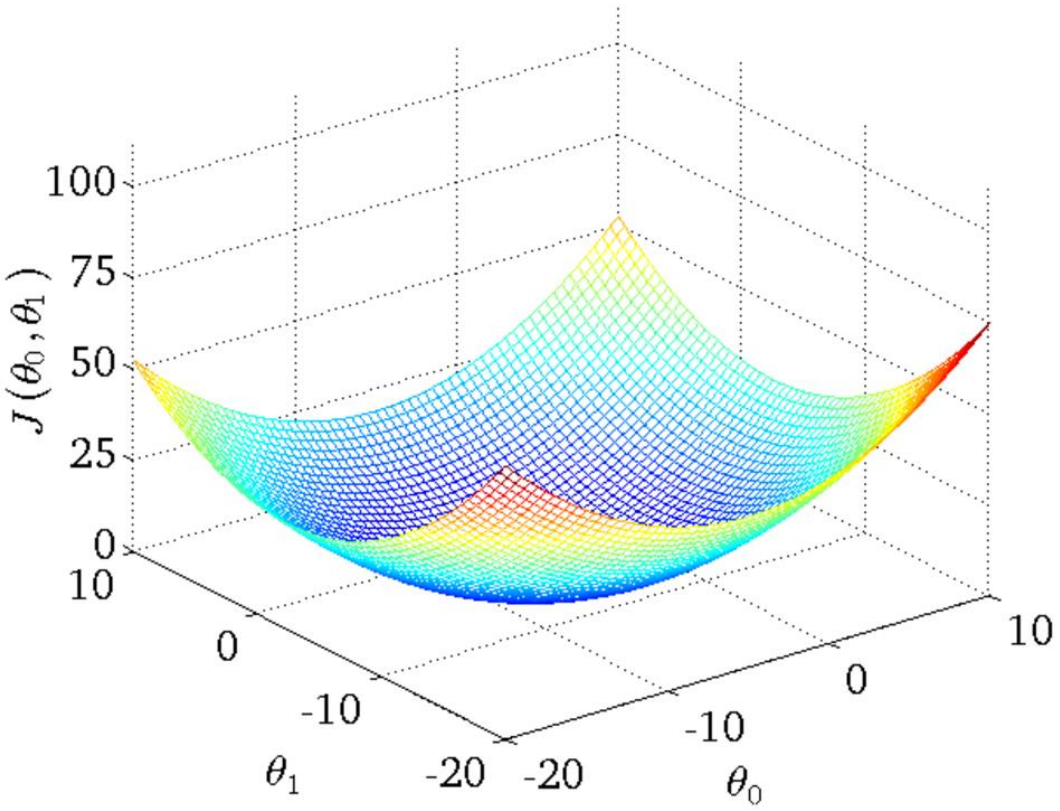


Minima and Maxima of a function Function



by Andrew Ng

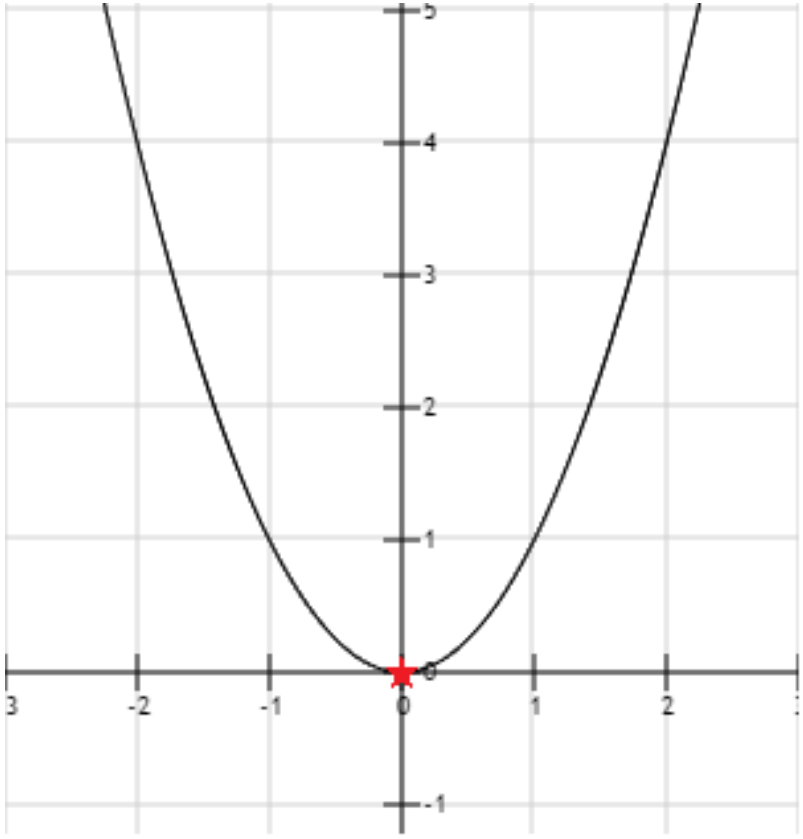
Convex Function



Andrew Ng

Unconstrained Optimization

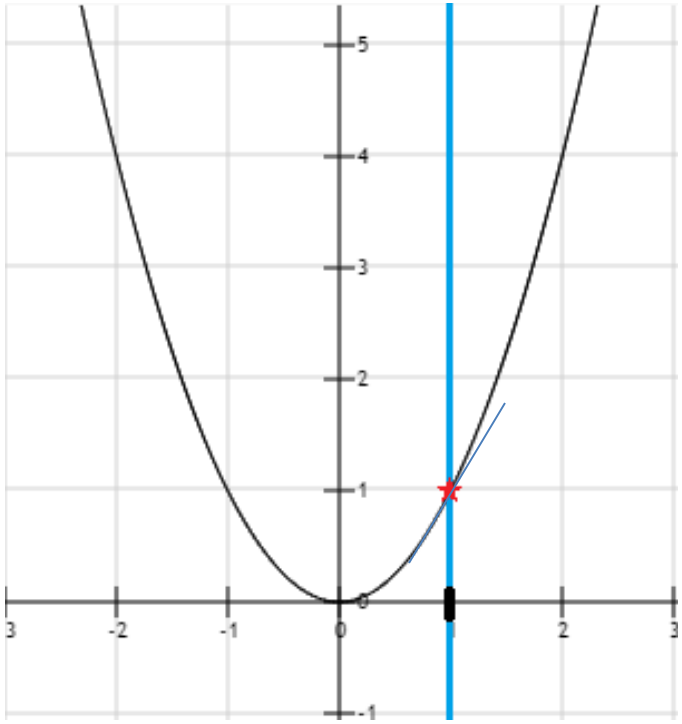
- Minimize x^2



Constrained Optimization -Equality Constraint

Minimize x^2

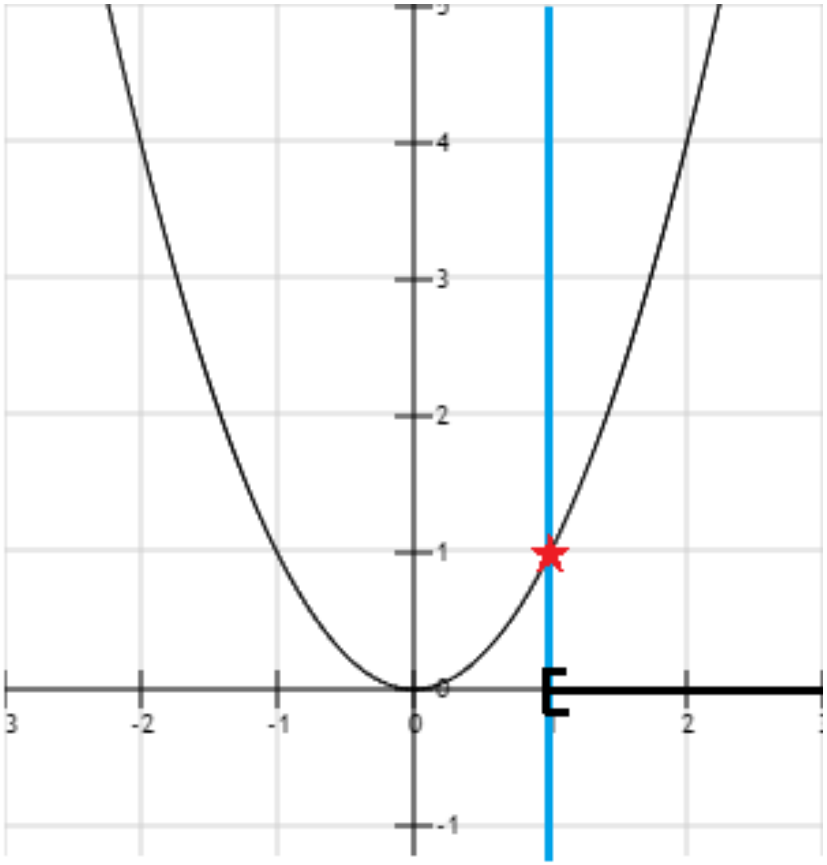
Subject to x = 1



Constrained Optimization -Inequality Constraint

Minimize x^2

Subject to x >= 1



Constrained optimization

- We can also have mix equality and inequality constraints together.
- Only restriction is that if we use contradictory constraints, we can end up with a problem which does not have a feasible set

Minimize x^2

Subject to

$x = 1$

$x < 0$

Impossible for x to be equal 1 and less than zero at the same

Constrained optimization

- A solution is an assignment of values to variables.
- A feasible solution is an assignment of values to variables such that all the constraints are satisfied.
- The objective function value of a solution is obtained by evaluating the objective function at the given solution.
- An optimal solution (assuming minimization) is one whose objective function value is less than or equal to that of all other feasible solutions.

Constrained Optimization Problem

- Optimization problem is typically written:

$$\text{Minimize } f(x)$$

$$\text{subject to}$$

$$g_i(x) = 0, \quad i=1,\ldots,p$$

$$h_i(x) <= 0, \quad i=1,\ldots,m$$

- f(x) is called the objective function
- By changing x (the optimization variable) we wish to find a value $x*$ for which f(x) is at its minimum.
- p functions of g_i define equality constraints and
- m functions h_i define inequality constraints.
- The value we find MUST respect these constraints!

Lagrange Multipliers

- **How do we find the solution to an optimization problem with constraints?**
- Constrained maximization (minimization) problem is rewritten as a Lagrange function whose optimal point is a saddle point, i.e. a global maximum (minimum)
- *Lagrange function use Lagrange multipliers as a strategy for finding the local maxima and minima of a function subject to equality constraints*
- Lagrange multipliers **only works with equality constraints**

Example of Lagrange Multiplier

Minimize $f(x,y) = x^2 + y^2$

Subject to $g(x, y) = x + y - 1 = 0$

Lagrange found that Minimum of f(x,y) under constraint g(x, y) is obtained **when their gradients point in the same direction**.

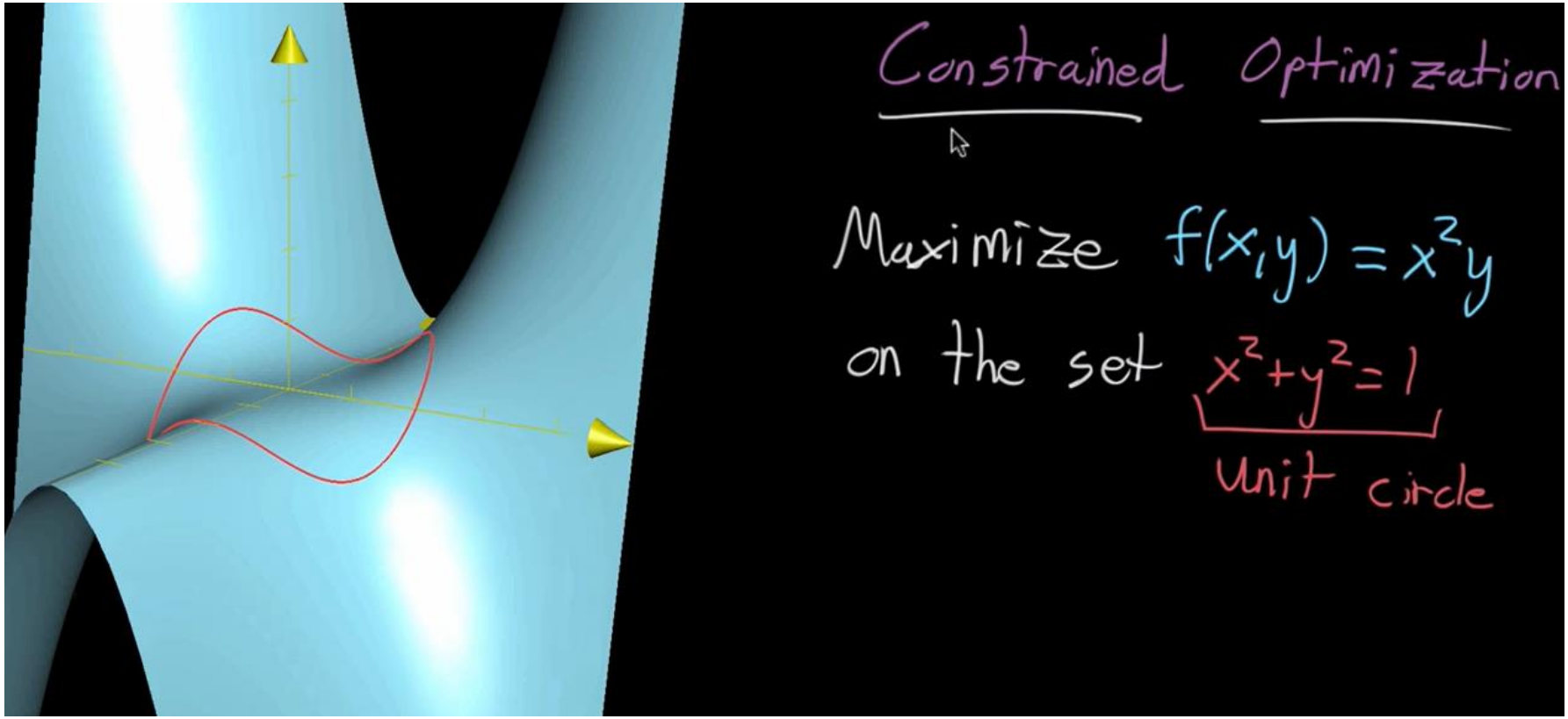
Mathematically,

$\nabla f(x,y) = \lambda \nabla g(x,y)$

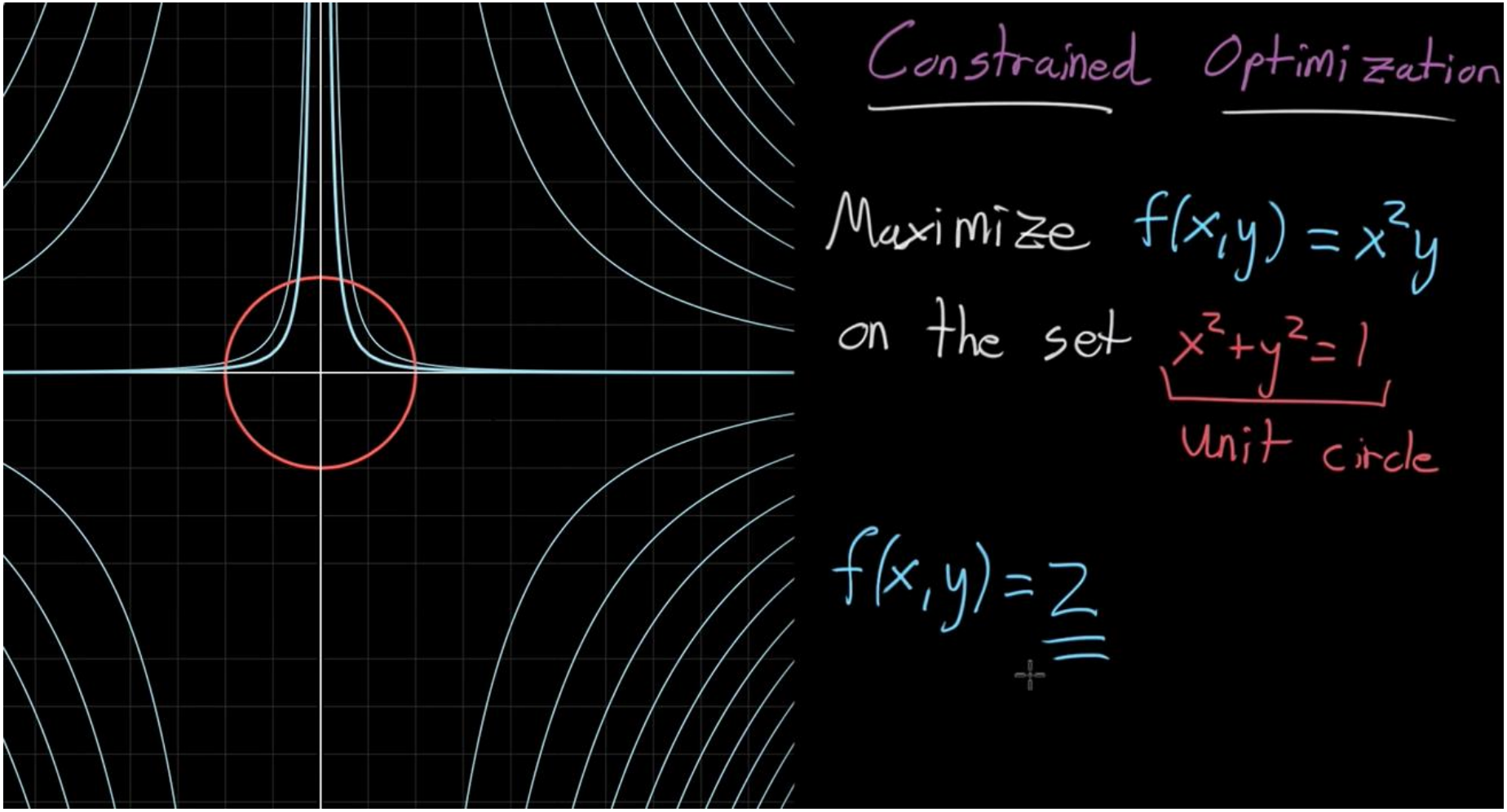
We introduce the Lagrangian function

$L(x,y,\lambda)=f(x,y)-\lambda g(x,y)$

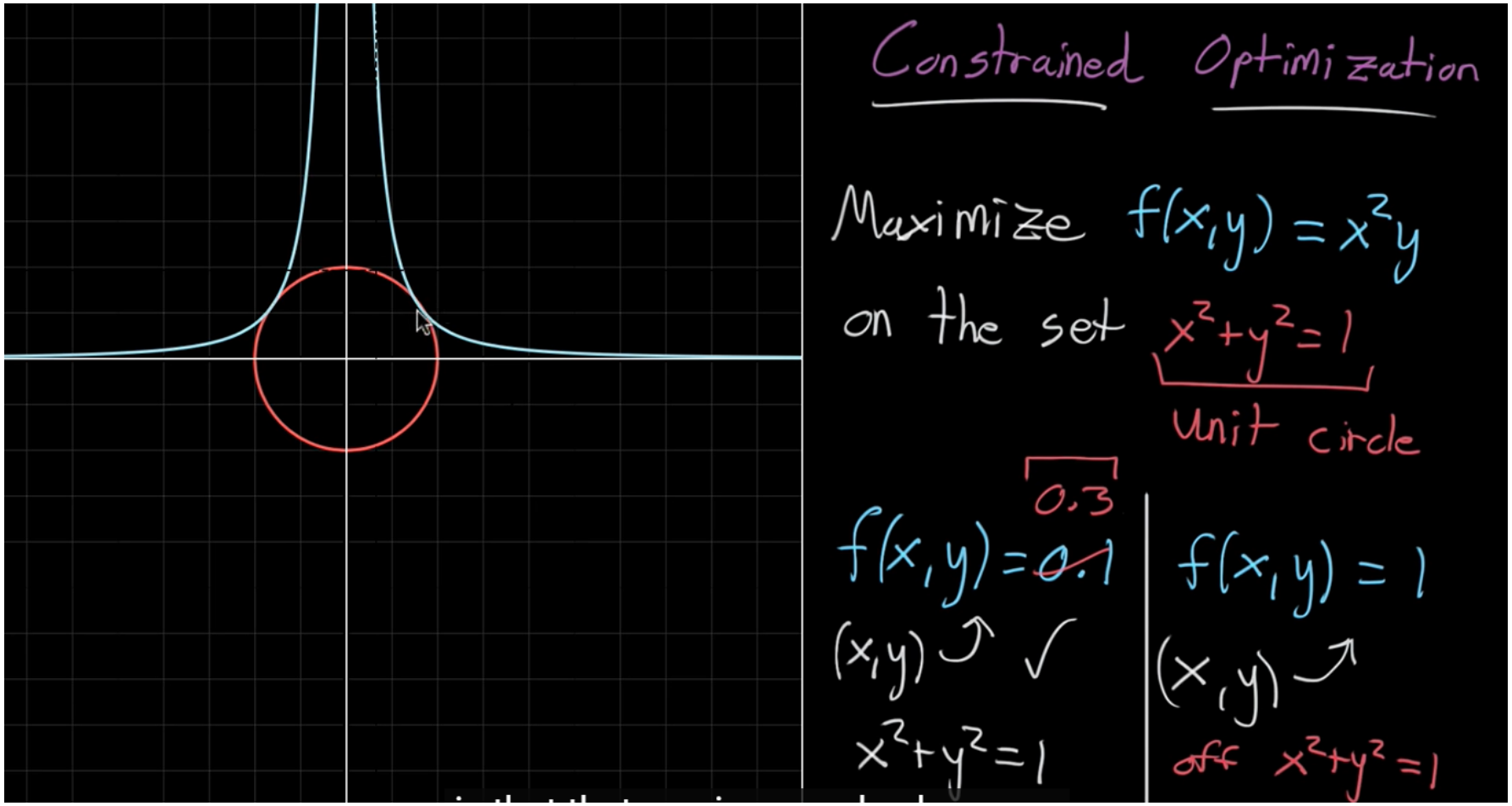
Constrained Optimization Problem



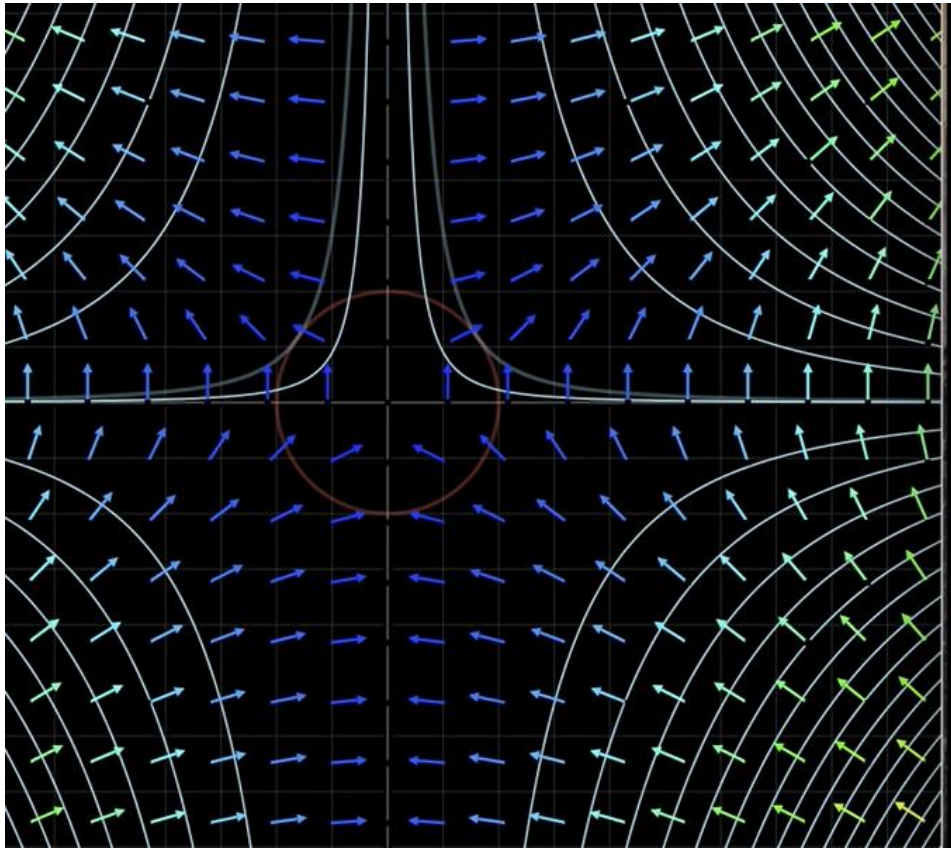
Contour Maps

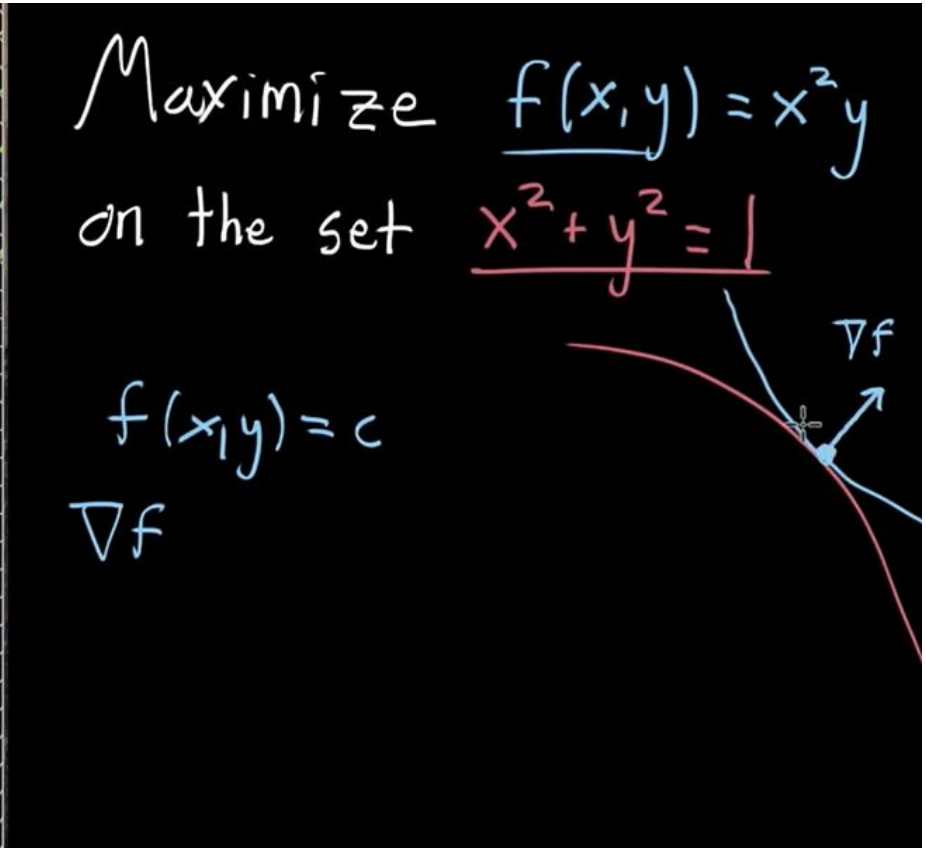


Contour Maps and Optimization

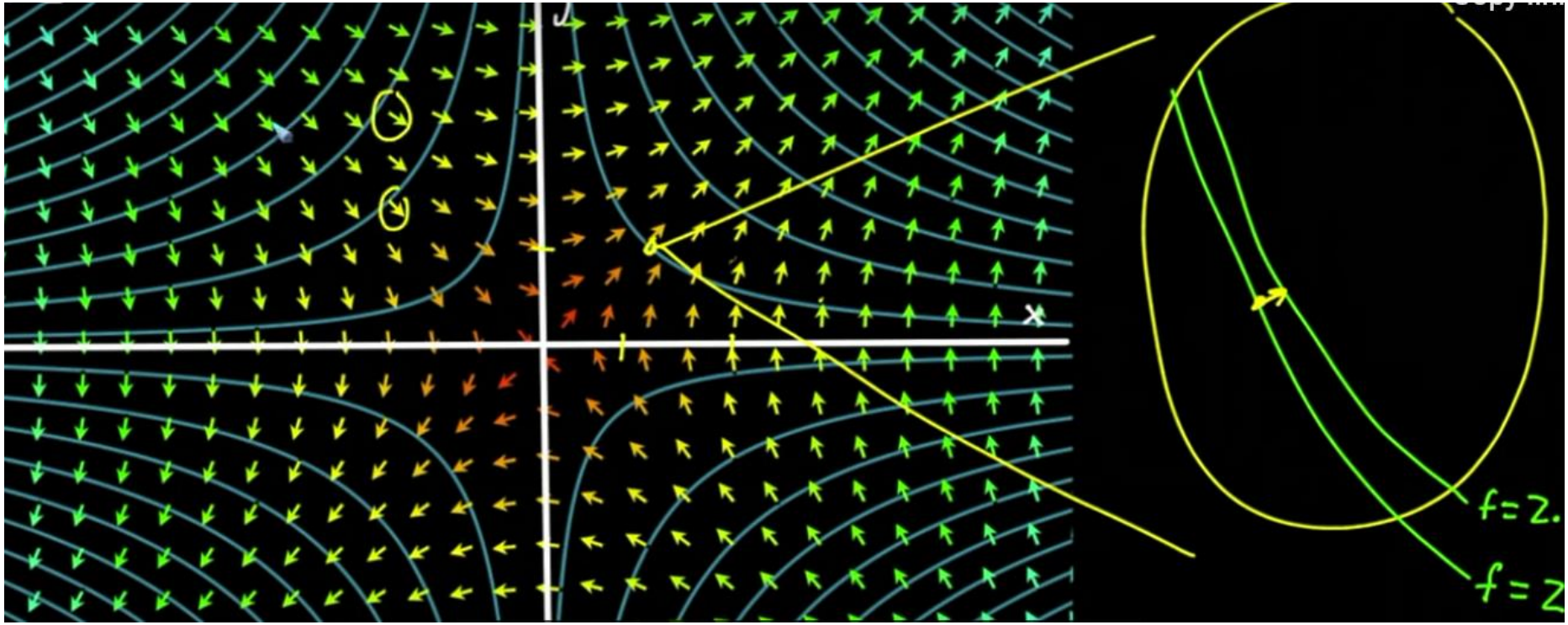


Contour maps and gradient

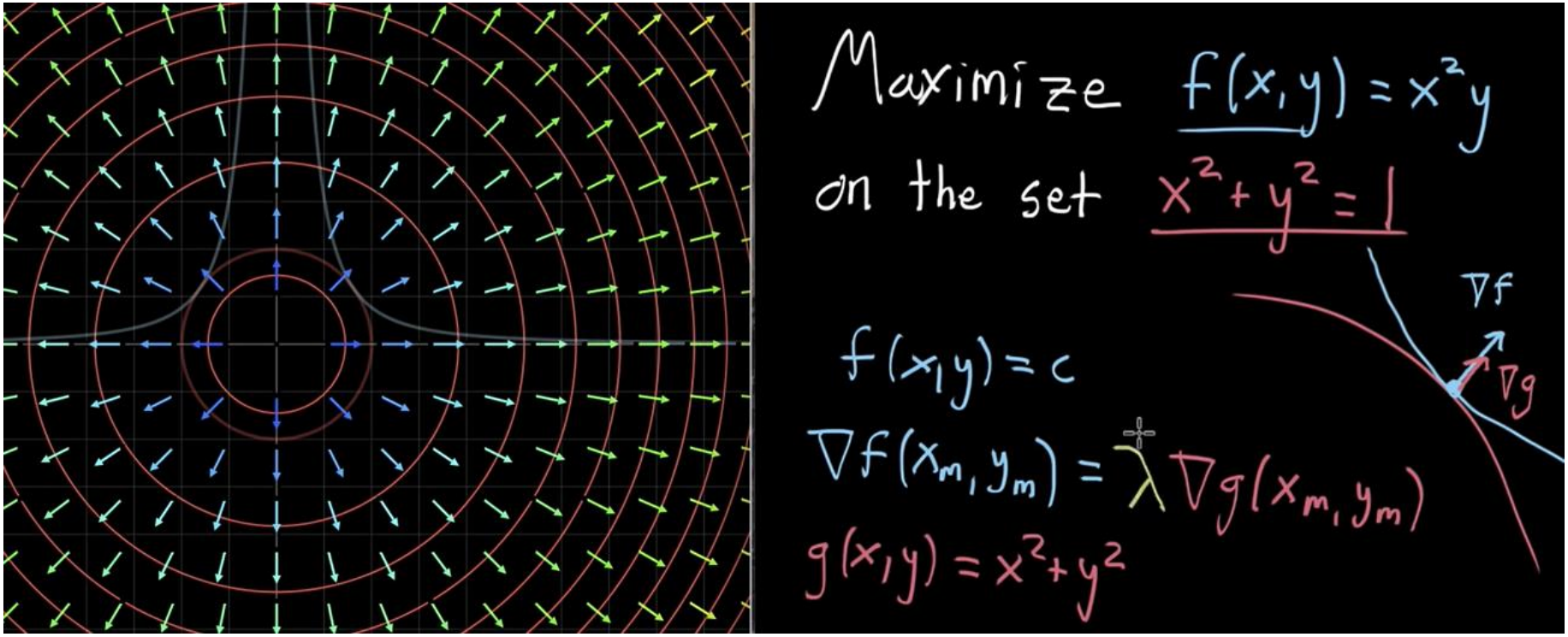




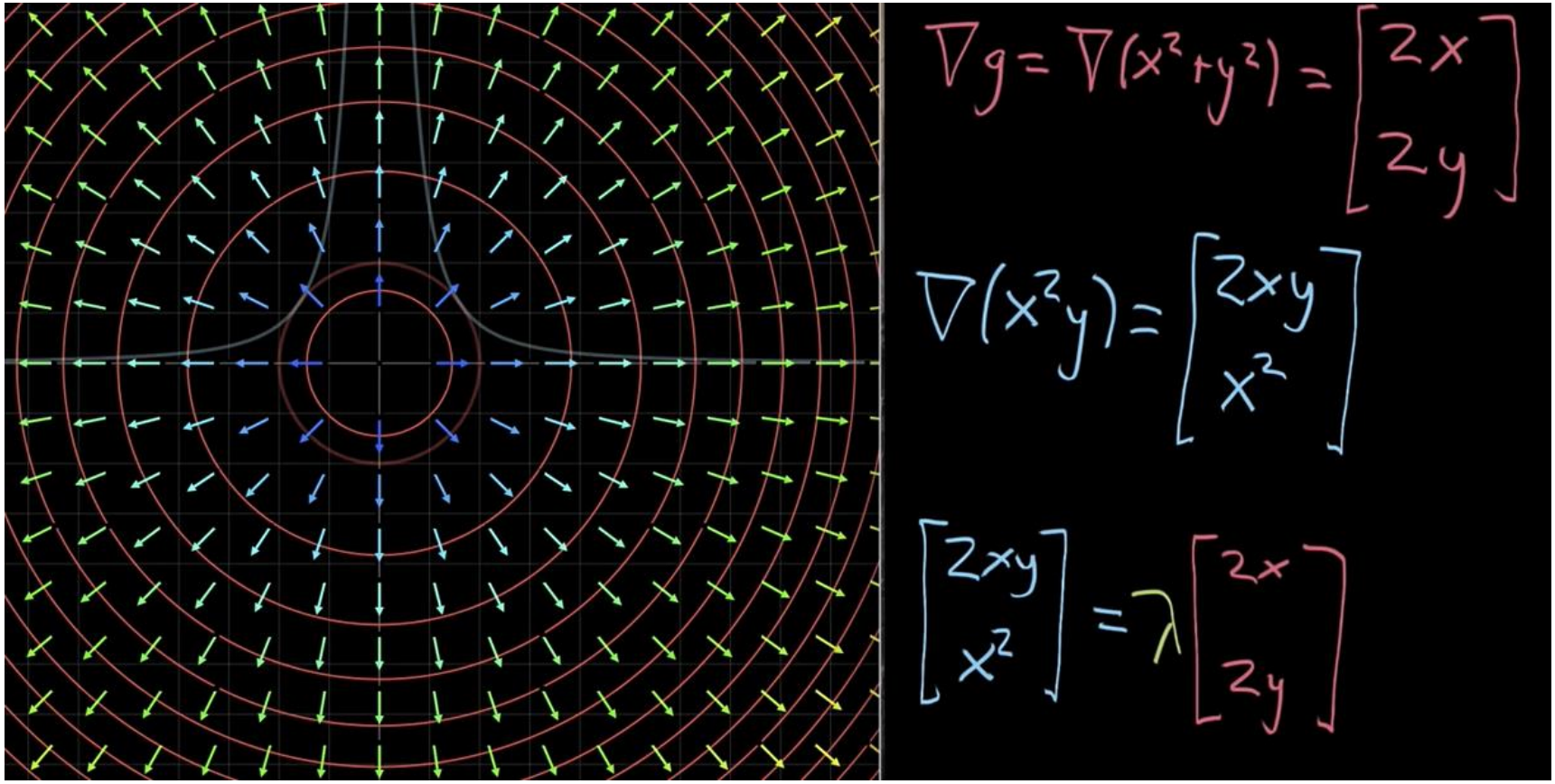
Contour maps and gradient



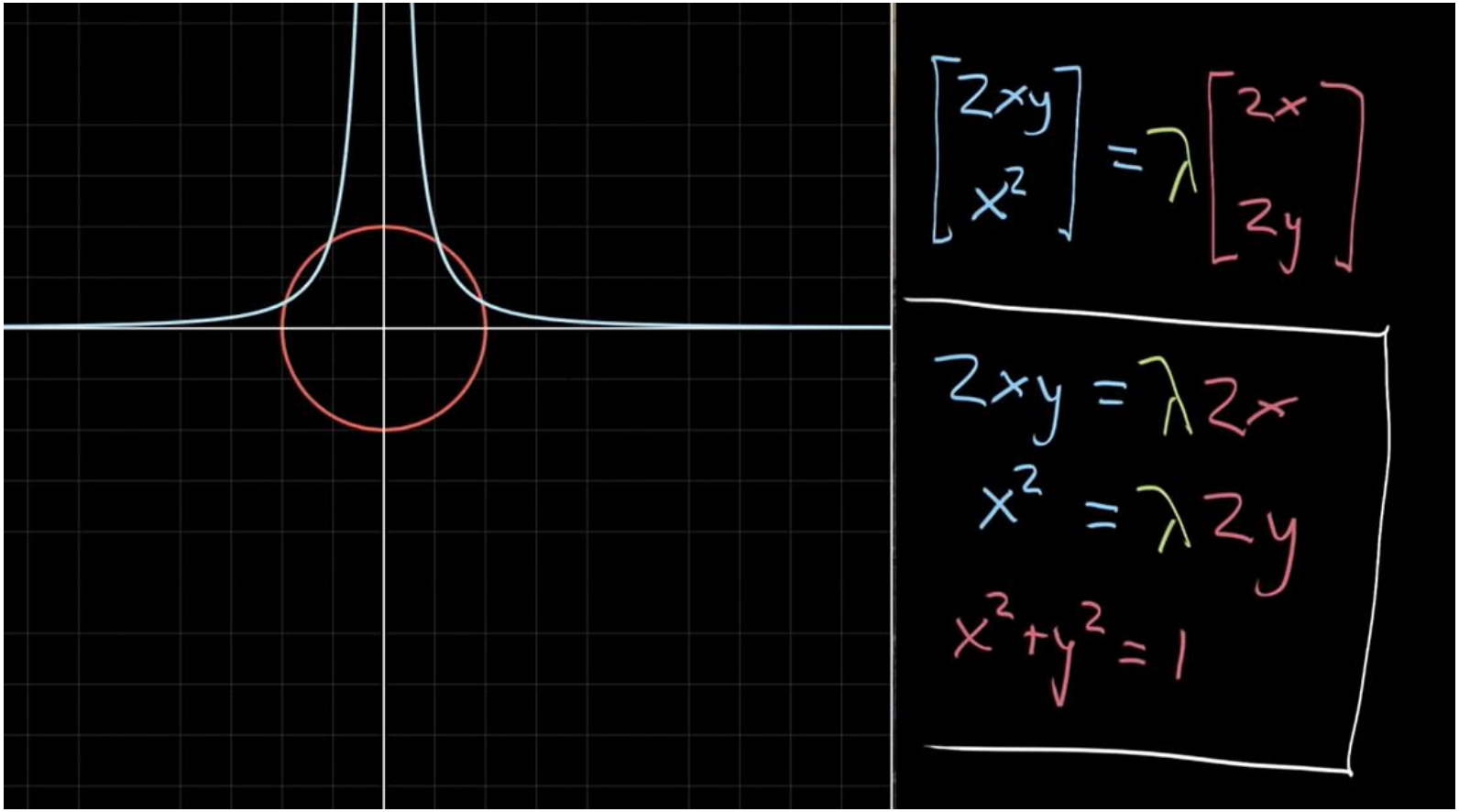
Lagrange Multiplier



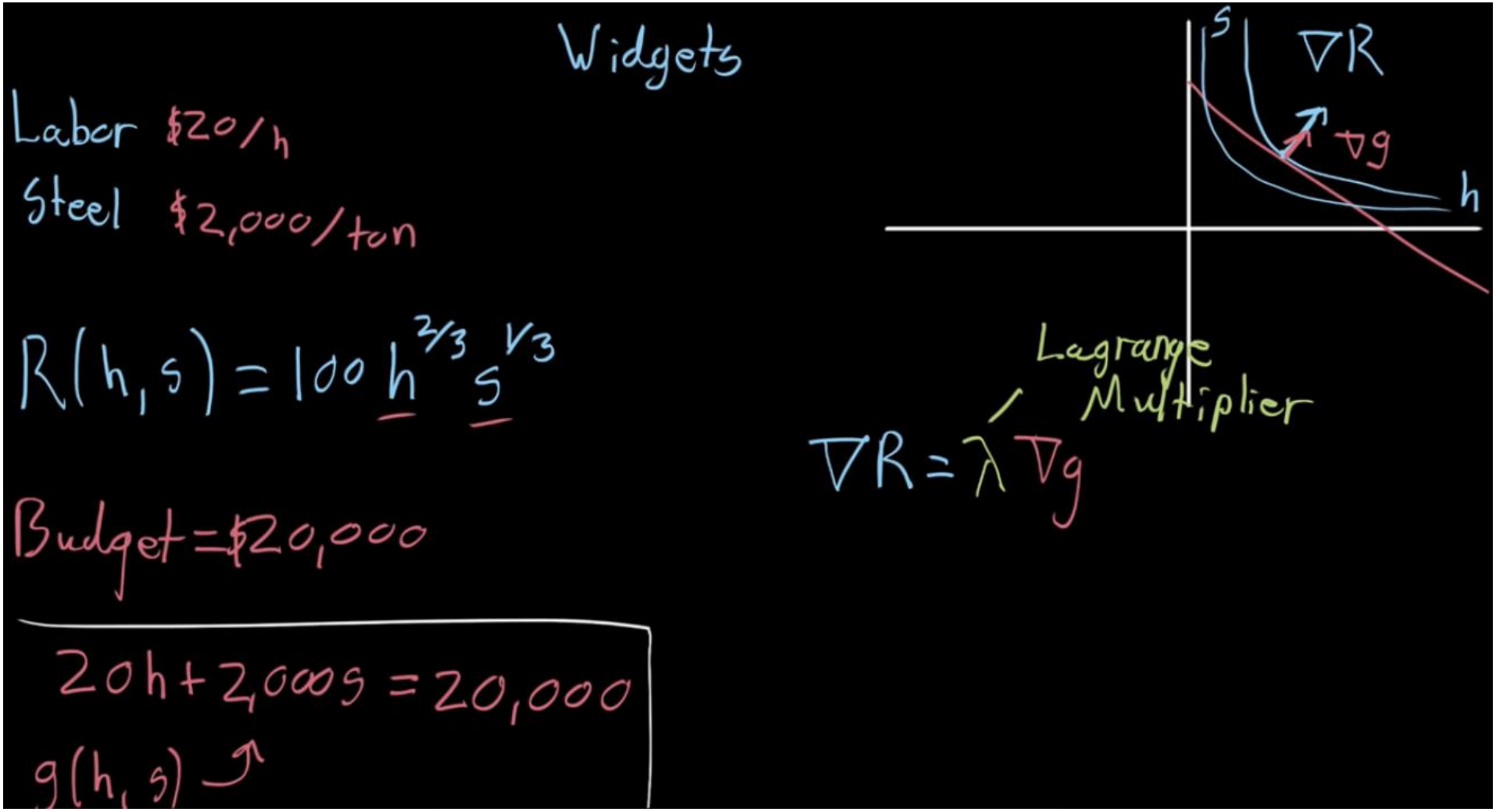
Lagrange Multiplier



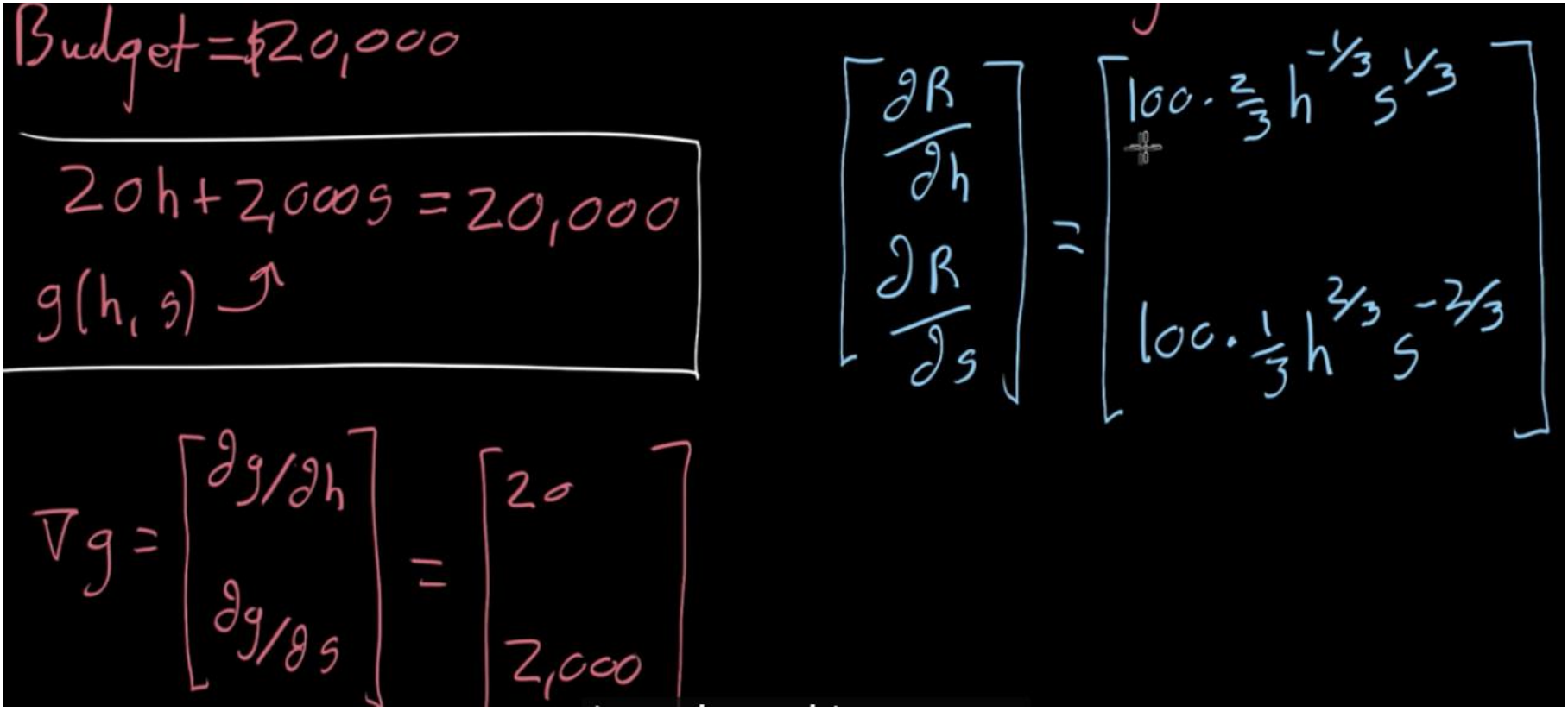
Lagrange Multiplier



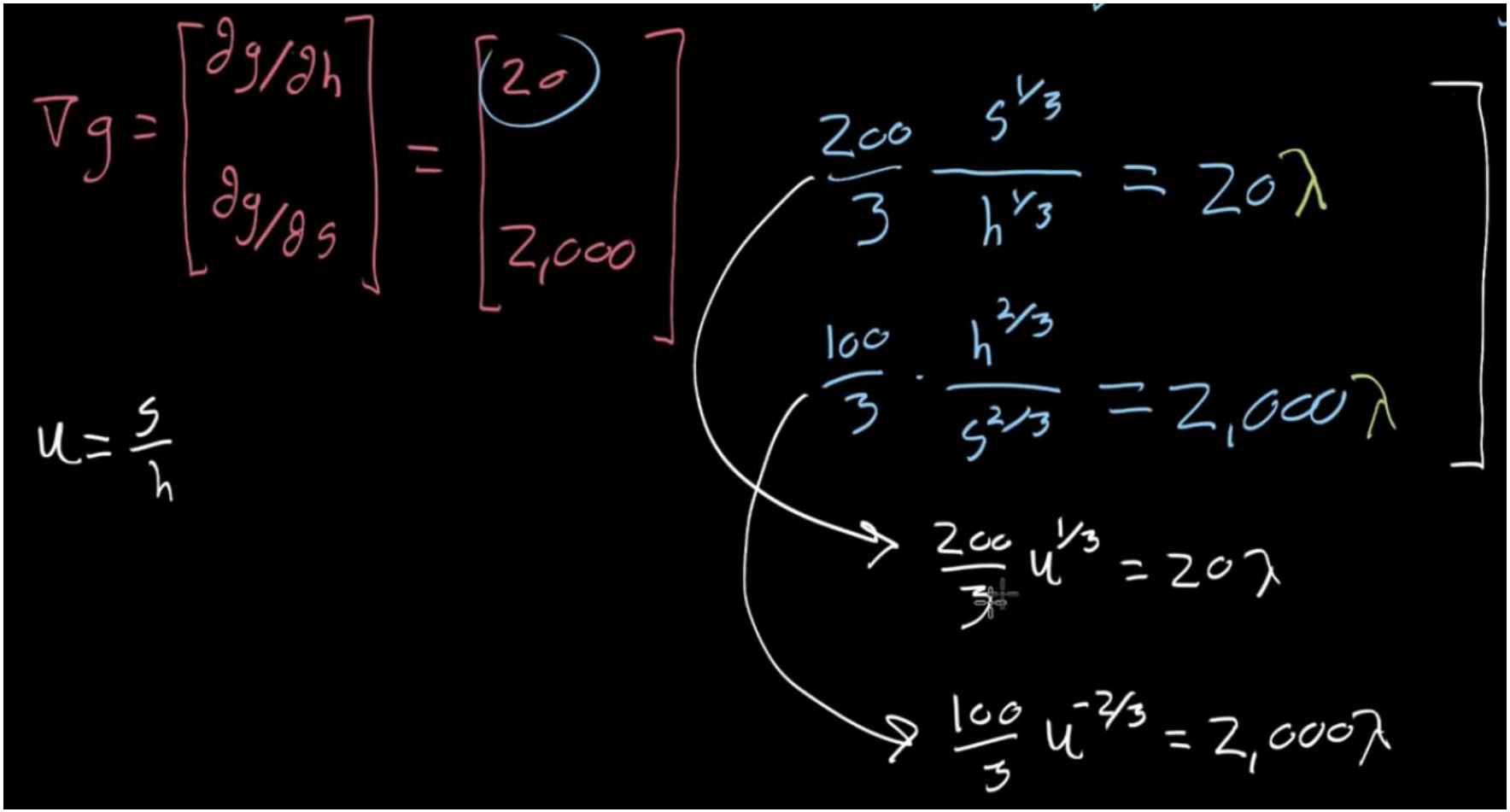
Real life example



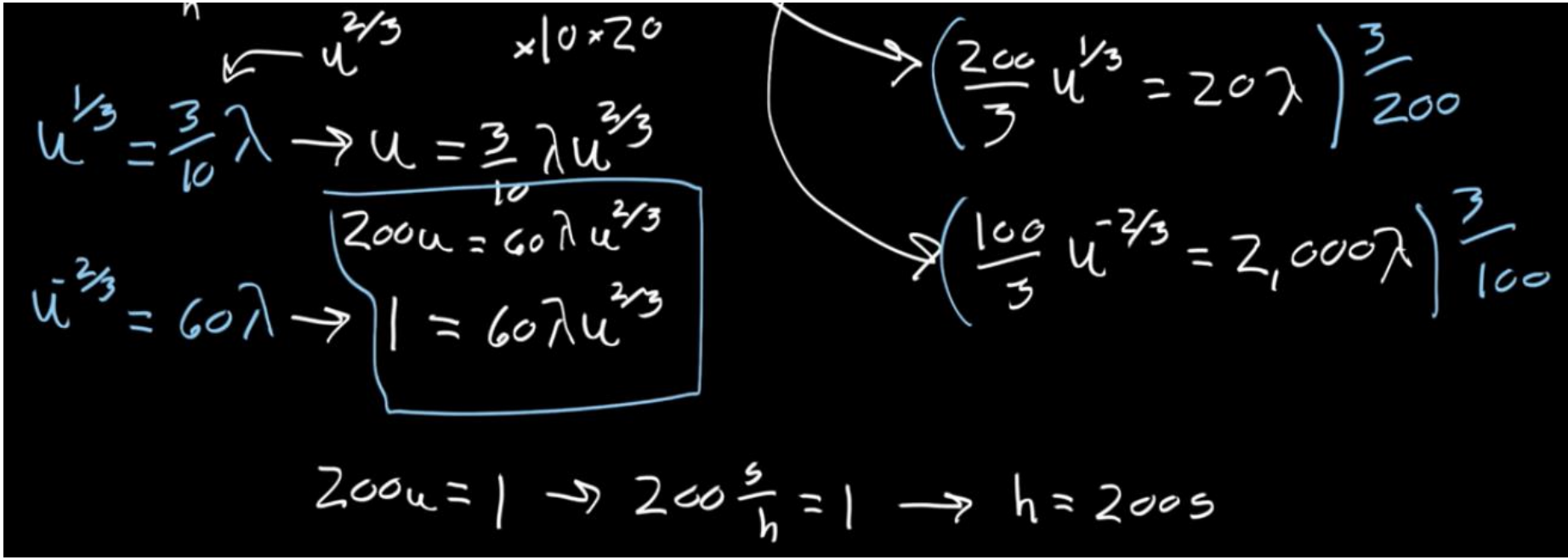
Gradient of function and constraint



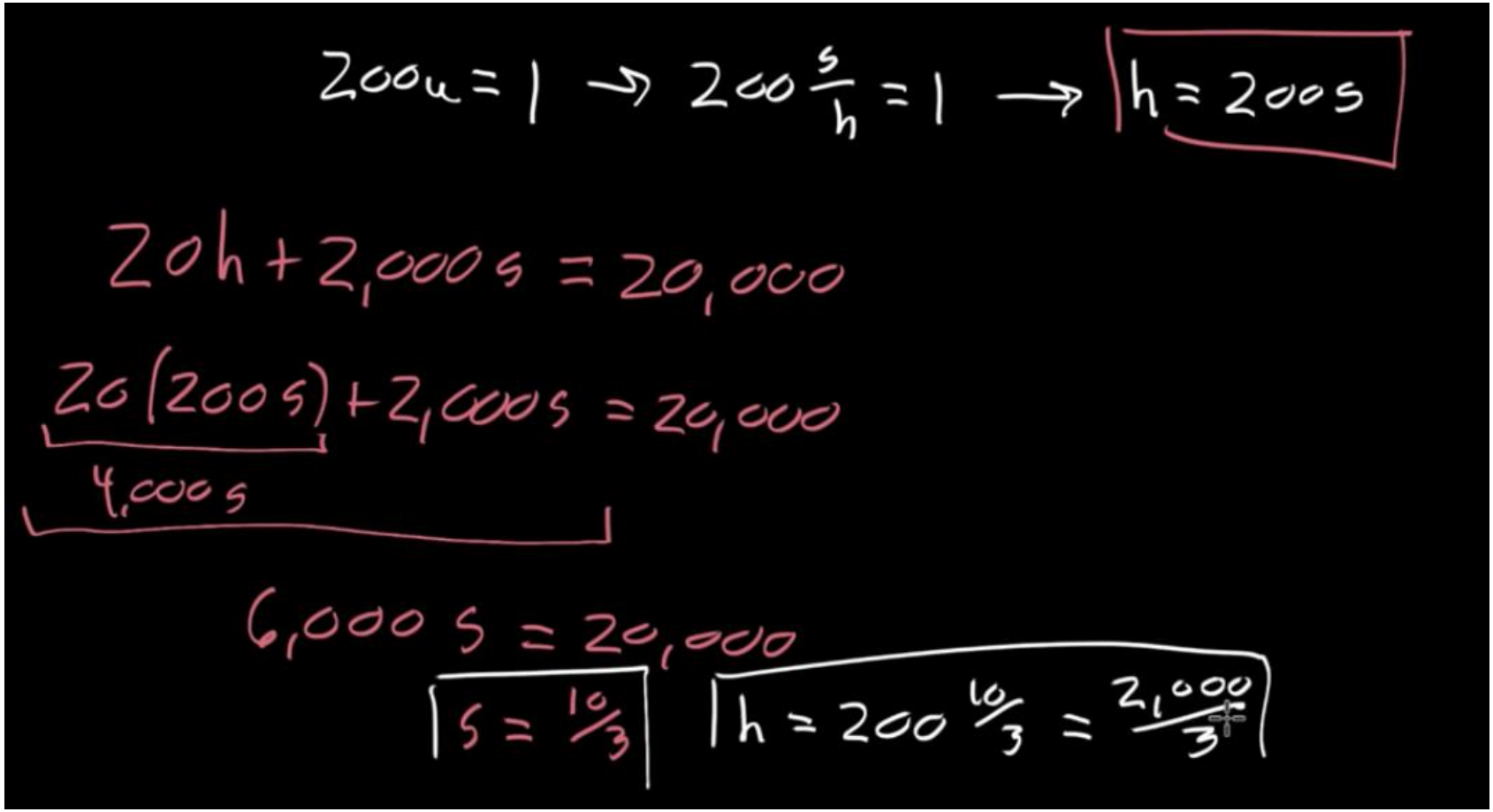
Solving for variable values



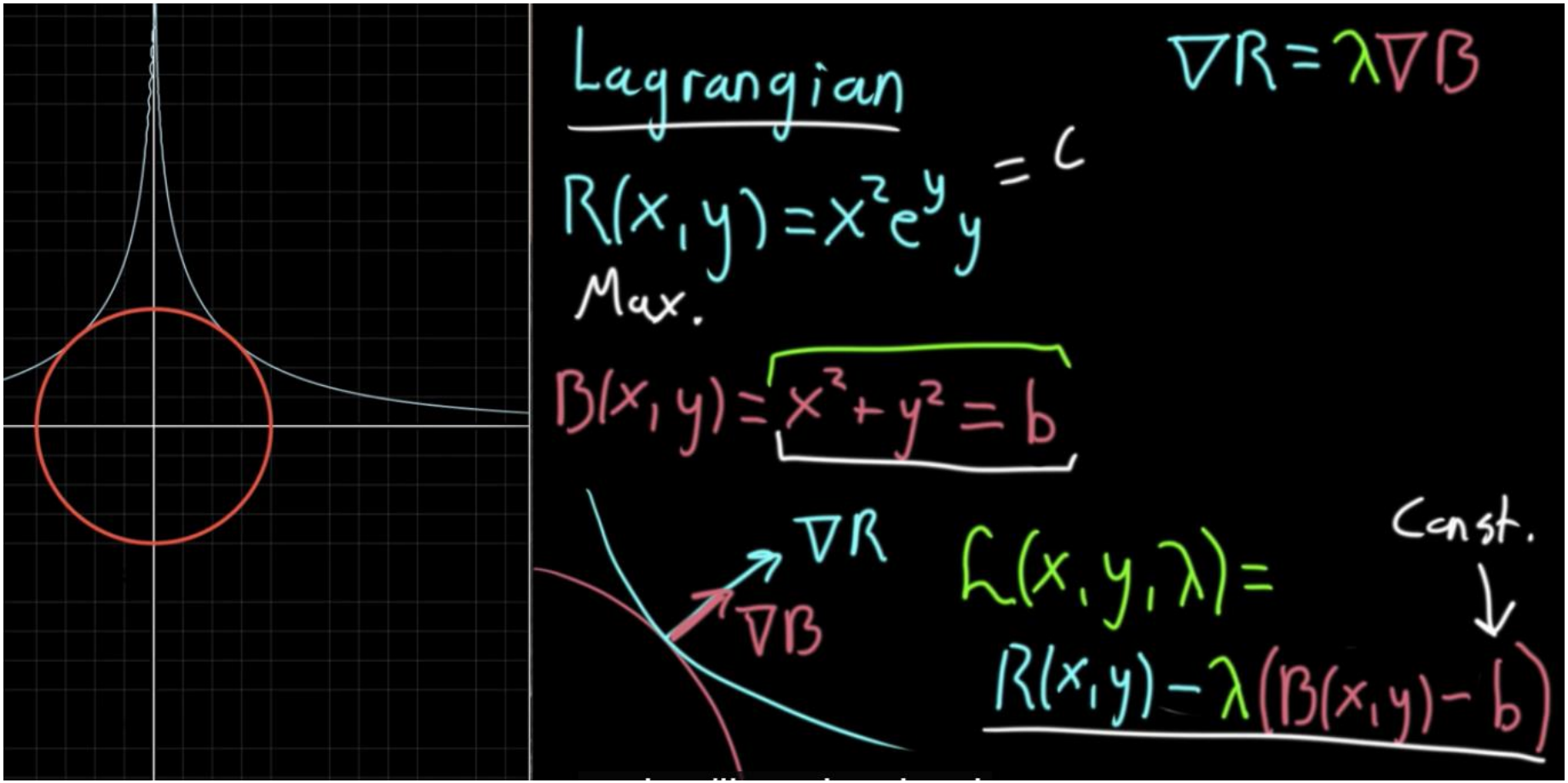
Solving for variable values



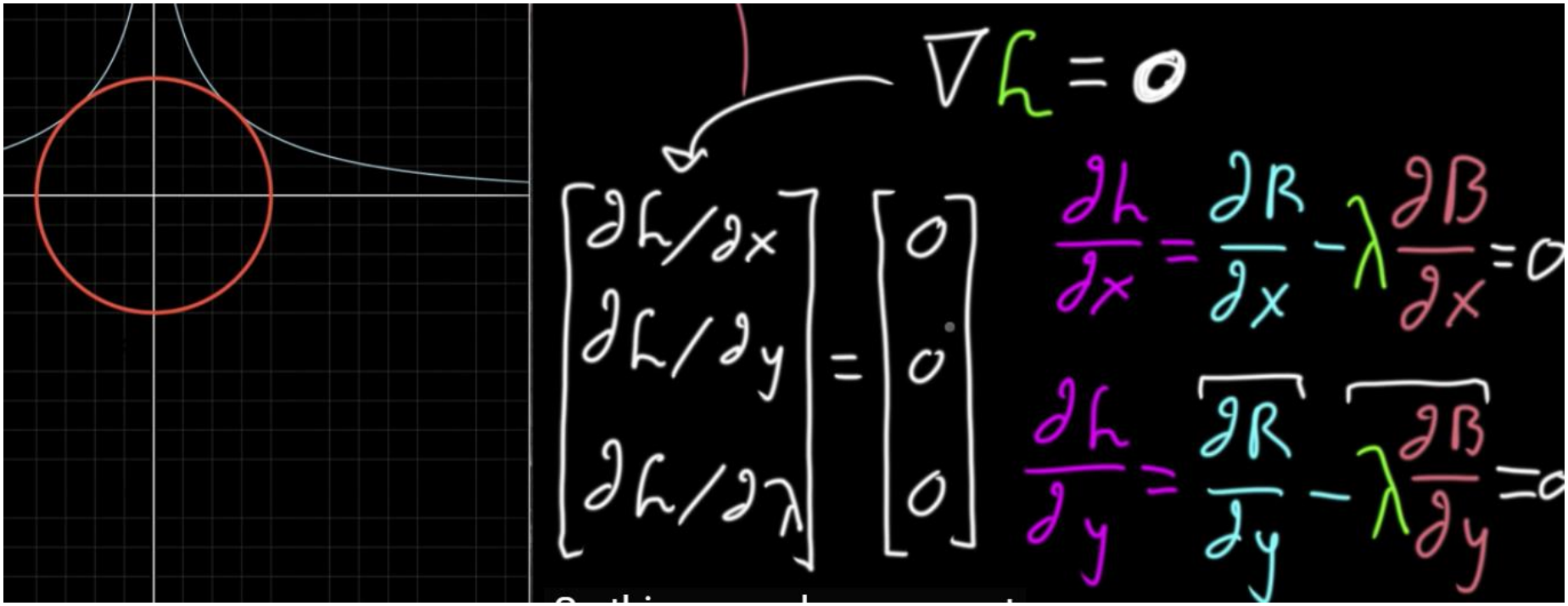
Solving for variable values



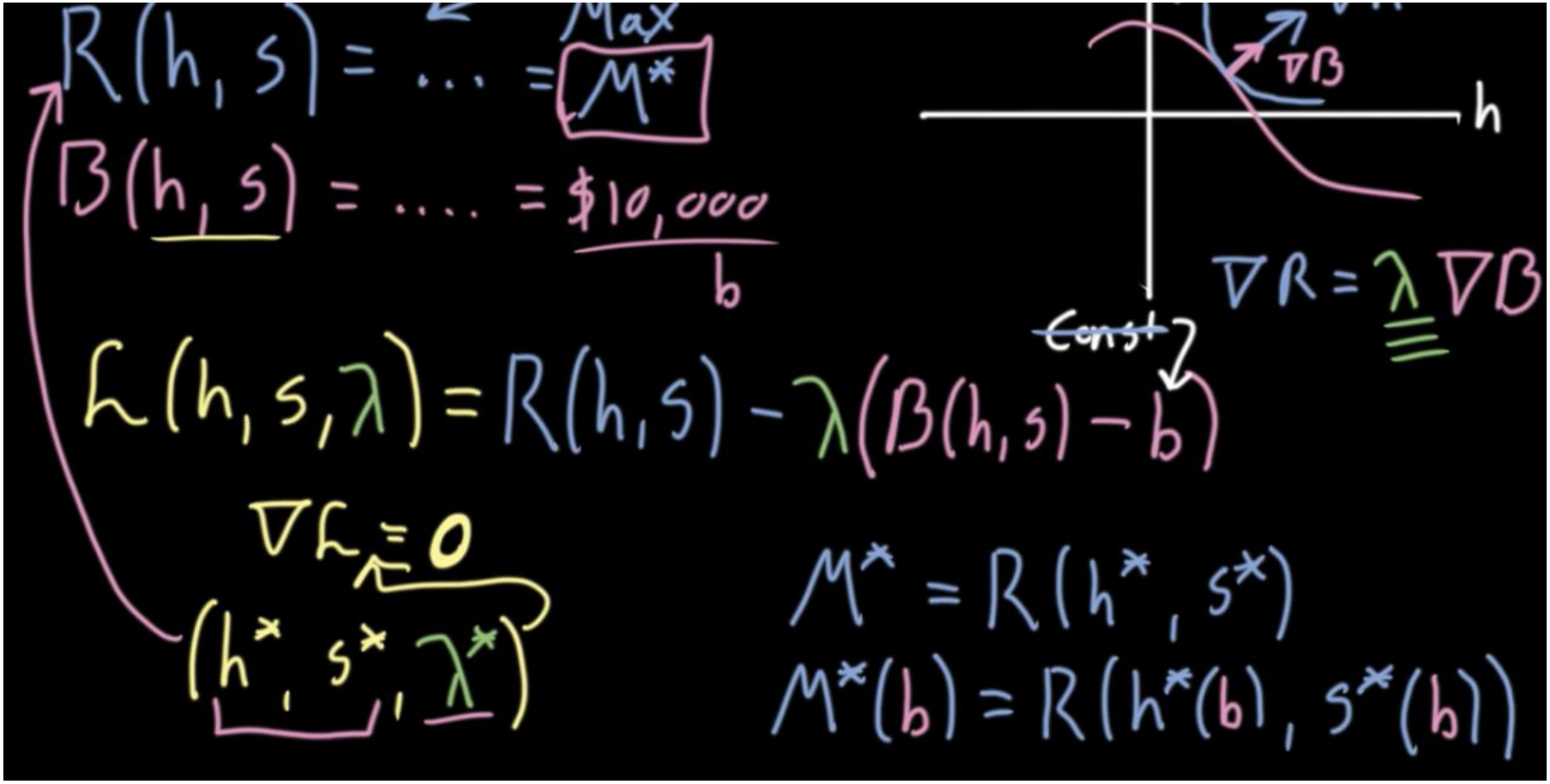
Budget and Revenue Example



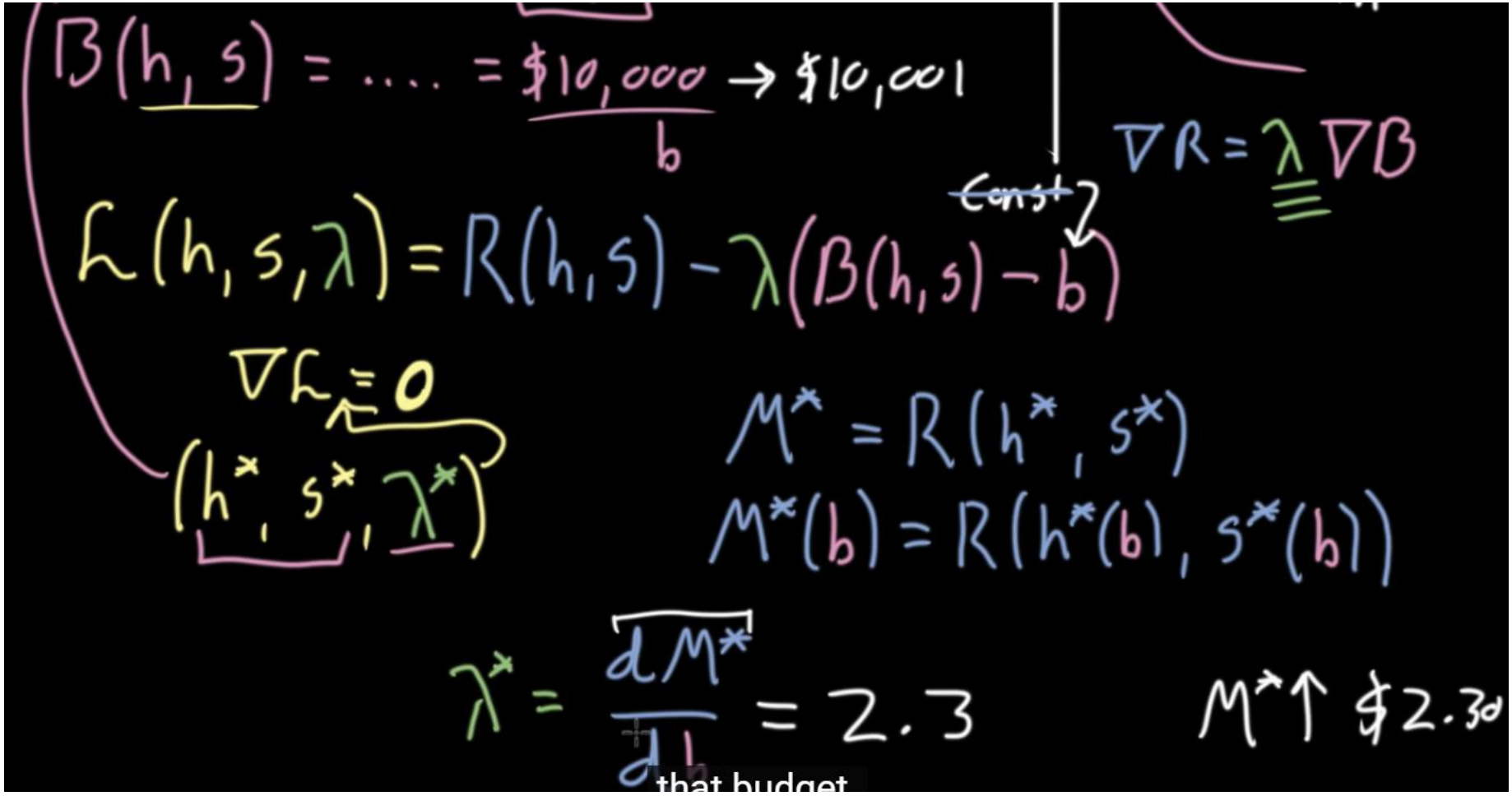
Budget and Revenue Example



Budget and Revenue Example



Budget and Revenue Example



Example:

$$\max_{x,y} (xy); \quad \textit{subject to } x + y = 6$$

- Introduce a Lagrange multiplier λ for constraint
- Construct the Lagrangian

$$L(x, y) = xy - \lambda(x + y - 6)$$

- Stationary points

$$\frac{\partial L(x,y)}{\partial \lambda} = x + y - 6 = 0$$

$$\left.\begin{aligned} \frac{\partial L(x,y)}{\partial x} &= y - \lambda = 0 \\ \frac{\partial L(x,y)}{\partial y} &= x - \lambda = 0 \end{aligned}\right\} \Rightarrow x = y = \lambda$$

$$\Rightarrow x = y = 3$$

x and y values remain same even if you take $+\lambda$ or $-\lambda$ for equality constraint

2 x = 6
x = y = 3
λ = 3

Linear programming(LP)

- **LP**, also called **linear optimization**
- Method to achieve the best outcome (such as maximum profit or lowest cost) in a mathematical model whose requirements are represented by linear relationships.
- It is a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints.

Duality Theory

- Every LP problem (called the 'Primal') has associated with another problem called the 'Dual'.
- The 'Dual' problem is an LP defined directly and systematically from the original (or Primal) LP model.
- The optimal solution of one problem yields the optimal solution to the other.
- Duality ease the calculations for the problems, whose number of variables is large.
- Conversion from primal to dual due to its mathematical elegance
- The dual is not only convex, but the duality gap is 0 even if the primal problem is not convex!

Duality or Duality principle

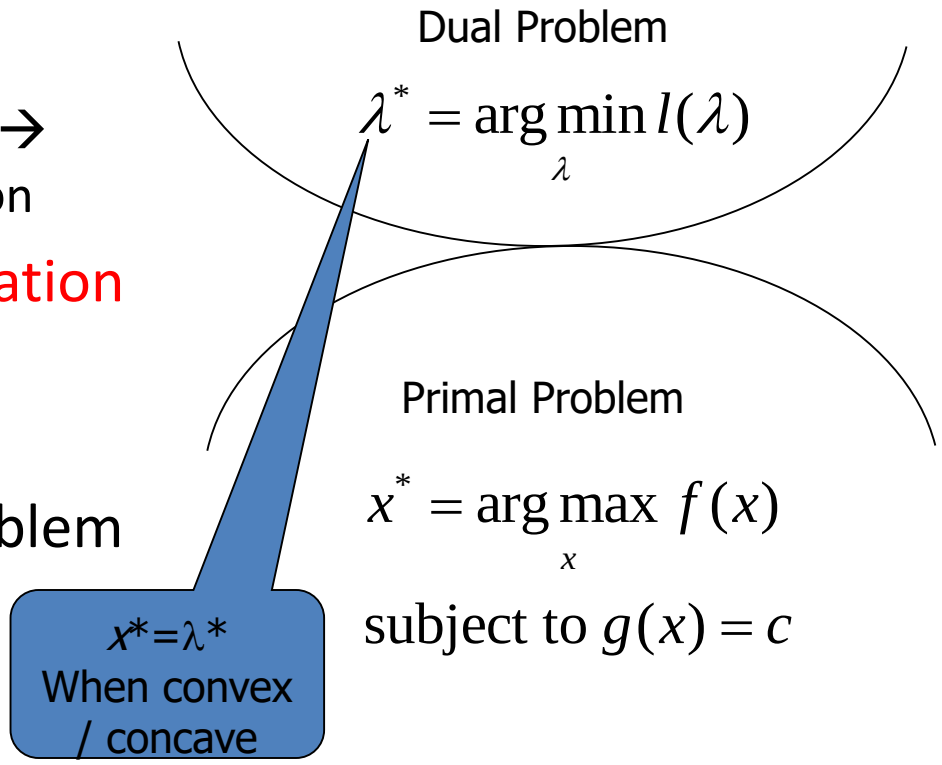
- ***Duality*** *means that optimization problems may be viewed from two perspectives,*
  - *Primal problem or the Dual problem*
  - *Solution to the dual problem provides a upper bound to the solution of the primal (maximization) problem.*

$$\text{Max } f(x)$$
$$\text{s.t. } g(x) \leq 0$$

- **Strong Duality Theorem:** If strong duality theorem holds, then the primal and dual optimal objective values are equal.
- Solving the dual problem is simpler than solving the primal problem.

Dual Problem

- Using dual problem
  - Constrained optimization → unconstrained optimization
- Need to change maximization to minimization
- strong duality when the original optimization problem is convex/concave



Strong Duality

Optimality Criterion theorem

If the primal and dual have feasible solution with the same value of the objective function, then both are optimal to the primal and dual respectively.

Maximize $10X_1 + 9X_2$
Subject to
$3X_1 + 3X_2 \leq 21$
$4X_1 + 3X_2 \leq 24$
$X_1, X_2 \geq 0$

(0,0) Z = 0
(1,1) Z = 19
(3, 1) Z = 39
(6, 0) Z = 60
(2, 5) Z = 65
(3, 4) Z = 66 ✓

Minimize $21Y_1 + 24Y_2$
Subject to
$3Y_1 + 4Y_2 \geq 10$
$3Y_1 + 3Y_2 \geq 9$
$Y_1, Y_2 \geq 0$

(10,10) W = 450
(6, 5) W = 246
(5, 4) W = 201
(3, 3) W = 135
(2, 2) W = 90
(2, 1) W = 66 ✓

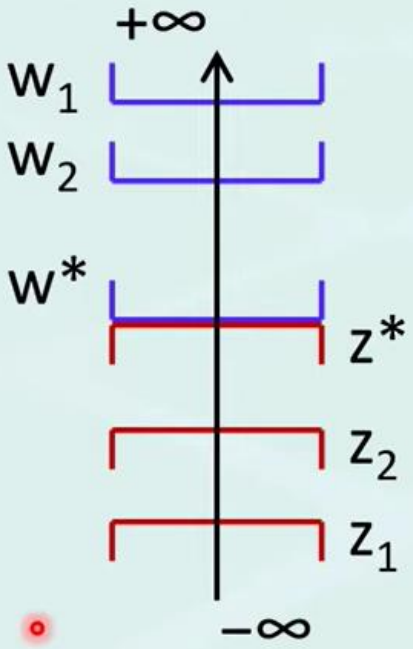
Lower Bound and Upper Bound

- If you pick a real number from the partially ordered set R and it is less than or equal to every element of a subset of R, then you can call this element a lower bound.

- Example:
  S={2,4,8,12}
  - 1 is less than or equal to 2, 4 ,8 and 12, 1 is a lower bound of S.
  - Same is true for -3 for instance.
  - 2 is also a lower bound of S.
- 2 is larger than any other lower bounds, it the greatest **lower bound.**

- The same logic apply with the relation "greater than or equal" and we have the concept of **upper-bound**.

Feasibility



Dual (Min w)

$z^*=w^*$

Primal (Max z)

- If the primal is unbounded, then the dual is infeasible
- If the dual is unbounded, then the primal is infeasible

Rules for converting Primal to Dual

- If the Primal is to maximize, the dual is to minimize.
- If the Primal is to minimize, the dual is to maximize.
- For every constraint in the primal, there is a dual variable.
- For every variable in the primal, there is a constraint in the dual.

Primal to Dual conversion table

Primal	Dual
Max	Min
Variables	Constraints
Constraints	Variables
RHS	Objective Function
Objective Function	RHS
A	A^T

Primal

Min. $Z = 10x_1+15x_2$

Subject to constraints:

$$5x_1 + 7x_2 \geq 80$$
$$6x_1 + 11x_2 \geq 100$$
$$x_1, x_2 \geq 0$$

Dual

Max. $Z' = 80y_1+100y_2$

Subject to constraints:

$$5y_1 + 6y_2 \leq 10$$
$$7y_1 + 11y_2 \leq 15$$
$$y_1, y_2 \geq 0$$

Example

Primal

Min. $Z' = 4y_1+12y_2+18y_3$

Subject to constraints:

$y_1 + 3y_3 \geq 3$

$2y_2 + 2y_3 \geq 5$

$y_1, y_2, y_3 \geq 0$

The Primal has:

3 variables and 2 constraints

The Dual has:

2 variables and 3 constraints.

Dual

Max. $Z = 3x_1+5x_2$

Subject to constraints:

$x_1 \leq 4$ y_1

$2x_2 \leq 12$ y_2

$3x_1+2x_2 \leq 18$ y_3

$x_1, x_2 \geq 0$

We define one dual variable for each primal constraint.

Example 1

Primal

$$\text{Min.. } Z = 10x_1 + 15x_2$$

Subject to constraints:

$$5x_1 + 7x_2 \geq 80$$

$$6x_1 + 11x_2 \leq 100$$

$$x_1, x_2 \geq 0$$

$$-6x_1 - 11x_2 \geq -100$$

Example Solution

innovate achieve lead

Dual

$$\text{Max.. } Z' = 80y_1 - 100y_2$$

Subject to constraints:

$$5y_1 - 6y_2 \leq 10$$
$$7y_1 - 11y_2 \leq 15$$
$$y_1, y_2 \geq 0$$

Non-linear Programming

- Process of solving an optimization problem where some of the constraints or the objective function are nonlinear. $x^3 - 2x + 1 = 0$
- An optimization problem is one of calculation of the extrema (maxima, minima or stationary points) of an objective function over a set of unknown real variables and conditional to the satisfaction of a system of equalities and inequalities

Quadratic Programming

- Problem of optimizing (minimizing or maximizing) a quadratic function (one or more variables in which the highest-degree term is of the second degree) of several variables subject to linear constraints on these variables.
- Quadratic programming is a particular type of nonlinear programming.
- Simplest form of non-linear programming

Karush–Kuhn–Tucker (KKT) theorem

- KKT approach to nonlinear programming (quadratic) generalizes the method of Lagrange multipliers, which allows only equality constraints.
- KKT allows inequality constraints

Karush-Kuhn-Tucker (KKT) conditions

- Start with

  min f(x) subject to

$$g_i(x) = 0 \text{ and } h_j(x) \geq 0 \text{ for all } i, j$$

- Make the Lagrangian function

$$\mathcal{L} = f(x) - \sum_i \lambda_i g_i(x) - \sum_j \mu_j h_j(x)$$

- Take gradient and set to 0 – but other conditions also.

KKT conditions – Equality and Inequality constraint

- Make the Lagrangian function for minimization:

$$\mathcal{L} = f(x) - \sum_i \lambda_i g_i(x) - \sum_j \mu_j h_j(x)$$

- Necessary conditions to have a minimum are

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$$
$$g_i(x^*) = 0 \text{ for all } i$$
$$h_j(x^*) \geq 0 \text{ for all } j$$
$$\mu_j \geq 0 \text{ for all } j$$
$$\mu_j^* h_j(x^*) = 0 \text{ for all } j$$

Good References for optimization

Lagrange Multiplier

- https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/lagrange-multipliers-and-constrained-optimization/v/constrained-optimization-introduction
- https://www.youtube.com/watch?v=liFWi2zR0MA&list=PLq-Gm0yRYwTipntZ17qTnGYAkyOPuhNEf&index=2
- Lagrange multipliers with visualizations and code | by Rohit Pandey | Towards Data Science
- http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf

Primal and dual

- https://www.youtube.com/watch?v=sUeX_JPlOAc&list=PLWoXNEI-KK1mCv_EL4OdF_-6FXryaZ11N&index=17
- https://www.quora.com/What-is-the-intuitive-explanation-for-the-duality-in-optimization-Why-are-the-primal-problem-and-the-dual-problem-equivalent

KKT

https://www.youtube.com/watch?v=Nbnd8KxRHGU

http://www.math.ubc.ca/~israel/m340/kkt2.pdf

Good References for SVM

Difference between SVM and NN

https://www.baeldung.com/cs/svm-vs-neural-network

Machine learning Github repository

- https://github.com/jermwatt/machine_learning_refined

Interpretability

- https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/

SVM Applications

- https://techvidvan.com/tutorials/svm-applications/
- https://data-flair.training/blogs/applications-of-svm/

Thank You

Dr. Chetana Gavankar has over 24 years of Teaching, Research and Industry experience. She has published papers in peer reviewed international conferences and journals. She is also reviewer for multiple conferences and journals. She has worked on different projects with multiple industries and received awards for her research work. Her areas of research interests include Natural Language Processing, Information Retrieval, Web Mining and Semantic Web, Ontology, Big Data Analytics, Machine learning, Deep learning and Artificial Intelligence.