Introduction to Mathematics for Political Science: Optimization

September 6, 2019

Optimization is at the core of many formal and quantitative problems in political science. How many yard signs should a mayoral candidate buy in order to *maximize* her probability of winning a local election? Who should an autocrat include in his ruling coalition in order to *maximize* rents, while ensuring he remains firmly in power? Game theoretic and decision theoretic models assume that political actors are goal oriented and make choices in order to realize those goals, subject to strategic and resource constraints. Choices are *optimal* given these constraints. Quantitative models of politics seek to fit models to data – choosing parameters in order to *minimize* the distance between the model and the data or to *maximize* the likelihood of data. We've now built up enough technical background to begin tackling these problems more rigorously. We'll tackle optimization problems of *constrained* and *unconstrained* varieties today.

Introduction

We'll focus on maximization problems in the exposition in order to fix ideas. The arguments presented here generalize easily to minimization problems, which we will encounter in quantitative applications.[1] Maxmization problems generically take the following form

$$\max_{\boldsymbol{x} \in G(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{1}$$

We call $f(\boldsymbol{x}, \theta)$ the *objective function* and $G(\boldsymbol{\theta})$ the *feasible set*. We can read the problem as "choose $\boldsymbol{x}$ in order to maximize the function f over the feasible set G," where both the function and the feasible set depend on some vector of parameters $\boldsymbol{\theta}$. The set of *maximizers* is written

$$\boldsymbol{x}^\star \in \arg\max_{\boldsymbol{x} \in G(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$$

A first order concern is whether or not it is possible to find such an $\boldsymbol{x}^\star$.[2] These problems become particularly acute when we work with general objective functions with sparse assumptions about their form. We'll sidestep these problems for the moment by working with the smooth functions we covered yesterday.

A *global optimum* is an $\boldsymbol{x}^\star$ that is "better" than any other $\boldsymbol{x}$ in the constraint set $G(\boldsymbol{\theta})$, $f(\boldsymbol{x}^\star; \boldsymbol{\theta}) \geq f(\boldsymbol{x}; \boldsymbol{\theta})$ for all $\boldsymbol{x} \in G(\boldsymbol{\theta})$. A *local*

[1] Equation 1 can be converted into a minimization problem by maximizing $-f(\boldsymbol{x}, \boldsymbol{\theta})$,

$$\min_{\boldsymbol{x} \in G(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{\boldsymbol{x} \in G(\boldsymbol{\theta})} -f(\boldsymbol{x}, \boldsymbol{\theta})$$

[2] Can you think of a function and domain for which no $\boldsymbol{x}^\star$ exists?

optimum is a $\boldsymbol{x}^\star$ that is better than all $\boldsymbol{x}$ in some neighborhood $S \subset G(\boldsymbol{\theta})$. Formally, $f(\boldsymbol{x}^\star; \boldsymbol{\theta}) \geq f(\boldsymbol{x}; \boldsymbol{\theta})$ for all $\boldsymbol{x} \in S$.

Example: Consider a spatial model of policy choice in which some decision maker with ideal point $\tilde{x}$ chooses some policy $x \in \mathbb{R}$ in order to maximize

$$u(x) = -(\tilde{x} - x)^2$$

What is the optimal policy choice?

Unconstrained Optimization

We start with unconstrained maximization problems, where the decision maker can choose any $\boldsymbol{x} \in X$.[3]

[3] Equivalently, $G(\boldsymbol{\theta}) = X$.

A local optimum is a point $\boldsymbol{x} \in G(\boldsymbol{\theta})$ that cannot be improved upon through small changes in $\boldsymbol{x}$. We want to find an $\boldsymbol{x}^\star \in S$ such that $f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S$. When f is smooth, we can approximate these $\boldsymbol{x}$ with linear functions,[4]

[4] It helps to think of the point being approximated as a $\boldsymbol{x} = \boldsymbol{x}^\star + \boldsymbol{dx}$

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^\star) + Df[\boldsymbol{x}^\star](\boldsymbol{x} - \boldsymbol{x}^\star)$$

If $x^\star$ is a maximum, then

$$\begin{aligned} f(\boldsymbol{x}^\star) &\geq f(\boldsymbol{x}^\star) + Df[\boldsymbol{x}^\star](\boldsymbol{x} - \boldsymbol{x}^\star) \\ 0 &\geq Df[\boldsymbol{x}^\star](\boldsymbol{x} - \boldsymbol{x}^\star) \end{aligned}$$

Linear approximations of $\boldsymbol{x}$ in the neighborhood of a local optimum slope downward. This holds in all directions, so if $\boldsymbol{x}^\star$ is an interior point,

$$Df[\boldsymbol{x}^\star](\boldsymbol{dx}) \leq 0 \qquad Df[\boldsymbol{x}^\star](-\boldsymbol{dx}) \leq 0$$

implying

$$Df[\boldsymbol{x}^\star](\boldsymbol{dx}) \leq 0 \qquad -Df[\boldsymbol{x}^\star](\boldsymbol{dx}) \leq 0$$

which can only be true if $Df[\boldsymbol{x}^\star](\boldsymbol{dx}) = 0$ for all steps $\boldsymbol{dx}$

The derivative of a functional evaluated at an interior local optimum must be the zero vector,

$$\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$$

Proposition (First Order Conditions): If $\boldsymbol{x}^\star$ is an interior local maximum of a functional f in X, then there exists an open neighborhood S of X such that

$$\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$$

Notice that the first order conditions are a necessary, but insufficient condition for finding a local maximum. It tells us that every local

maximum satisfies the first order conditions – *not* that the satisfying the first order conditions identifies a local maximum. The first order conditions identify stationary points, which may be maxima, minima, or saddle points, depending on the functions local concavity or convexity.

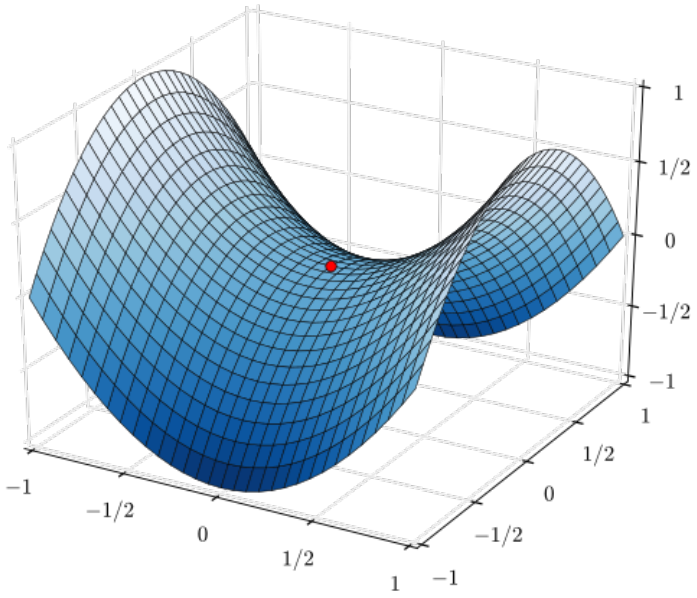

Figure 1: By Nicoguaro - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=20

The spatial model has a unique, interior $x^\star$ because the utility function is *concave*. This feature should be familiar from your study of univariate calculus. Here, we extend this intuition to a multivariate environment and study in greater depth the relationship between concavity, smoothness, and optimization.

Recall the definition of a concave function.[5] A function f is concave on a convex set S if

$$f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \geq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and $\alpha \in [0,1]$.

[5] See lecture notes on "Monotone, Linear, and Convex Functions"

Proposition (Tangent Hyperplanes): If a function f is differentiable and concave on a open, convex set S, then

$$f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_2) + Df[\boldsymbol{x}_2](\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$.

The proposition claims that linear approximations of concave functions always lie above their target. This is easy to visualize in the

one-dimensional case. Every line tangent to a concave function f lies above the function.

Proof: If f is concave,

$$\begin{aligned} f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) &\geq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) \\ f(\alpha(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \boldsymbol{x}_2) &\geq \alpha\left(f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\right) + f(\boldsymbol{x}_2) \\ \frac{f(\alpha(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \boldsymbol{x}_2) - f(\boldsymbol{x}_2)}{\alpha} &\geq f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \\ \lim_{\alpha \to 0} \frac{f(\alpha(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \boldsymbol{x}_2) - f(\boldsymbol{x}_2)}{\alpha} &\geq f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \\ Df[\boldsymbol{x}_2](\boldsymbol{x}_1 - \boldsymbol{x}_2) + f(\boldsymbol{x}_2) &\geq f(\boldsymbol{x}_1) \end{aligned}$$

which holds for arbitrary $\boldsymbol{x}_1$, allowing us to conclude

$$f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_2) + Df[\boldsymbol{x}_2](\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

as desired.

If we're standing at the peak of a mountain, all paths lead down – there is no direction to walk which will increase our elevation. If the mountain is *smooth*, and the peak is interior to the neighborhood of interest S then the peak itself must be flat! Why is this the case?

Rolle's Theorem: Let $f \in C[a,b]$ with f differentiable on (a,b). If $f(a) = f(b)$ then there exists some $x \in (a,b)$ such that $f'(x) = 0$.[6]

[6] We leave the proof of this fact as an exercise.

Proposition (Second Order Conditions): If $\boldsymbol{x}^\star$ is a stationary point of f with $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$ and f is strictly locally concave at $\boldsymbol{x}^\star$ ($H_f(\boldsymbol{x}^\star)$ is negative definite) then $\boldsymbol{x}^\star$ is a strict local maximum of f.

Corollary: If f is strictly concave and $\boldsymbol{x}^\star$ is interior to X, then $\boldsymbol{x}^\star$ is a global maximum of f iff $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$.

Proof: Applying the proposition on tangent hyperplanes, we have

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}^\star) + \underbrace{\nabla f(\boldsymbol{x}^\star)^T}_{\mathbf{0}}(\boldsymbol{x} - \boldsymbol{x}^\star) = f(\boldsymbol{x}^\star)$$

for all $\boldsymbol{x}$ in the interior of X.

Combining the first order conditions with the strict concavity of f ensures that we have located global maxima of the function or local maxima on concave neighborhoods of X.

Example: Let

$$f(x,y) = x^2 - 6xy + 2y^2 + 10x + 2y - 5$$

Find all critical points of the function and classify them as local maxima, local minima, or saddle points.

Example: A linear model relates *independent variables* $\boldsymbol{x}_i$ to a *dependent variable* y_i through the following equation

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

The distance between the predictions of the model and the data is given by

$$\epsilon_i = \boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{\beta}$$

The *least squares estimator* seeks to choose the model's parameters $\boldsymbol{\beta}$ in order to minimize the *sum of squared errors* across all observations $i \in \{1, ..., N\}$

$$\min_{\beta} \sum_i \epsilon_i^2 = \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right)^2 \tag{2}$$

Assume for the moment that $\sum_i \epsilon_i^2$ is a convex function.[7] Then the first order conditions must characterize the optimal choice of $\boldsymbol{\beta}$, which we'll denote with $\hat{\boldsymbol{\beta}}$. Remember, there are multiple independent variables for each observations, each of which we'll denote with x_{ij}, where x_{ij} is individual[8] i's value for variable j. Each β in the vector of coefficients will be denoted with β_j Then Equation 2 becomes

[7] We leave the proof of this as an exercise.

[8] or country, state, etc.

$$\min_{\beta} \sum_i \epsilon_i^2 = \sum_i \left(y_i - \sum_j x_{ij} \beta_j \right)^2$$

and the first order conditions are

$$\frac{\partial \epsilon_i^2}{\partial \beta_k} = 2 \sum_i \left(y_i - \sum_j x_{ij} \beta_j \right) (-x_{ik}) = 0$$

$$\sum_i y_i x_{ik} = \sum_i x_{ik} \sum_j \beta_j x_{ij}$$

Which must hold for all $j \in \{1, ..., J\}$. Stacking these first order conditions and writing sums as dot products gives

$$\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NJ} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_J \end{bmatrix}$$

which gives[9]

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

[9] Verify that $\hat{\boldsymbol{\beta}}$ is a $J \times 1$ vector

Constrained Optimization

Suppose a candidate for Senate of the United States must decide how many dollars to allocate to TV advertising in each TV market in his state. Let $x_i \in \mathbb{R}_+$ denote the amount that the candidate chooses to allocate to each district. The campaign manager has given the candidate a fixed budget B to spend on TV advertising across the state and demanded that he spend each and every dollar. An increasing and concave function $v_i(x_i)$ determines how much voter support the candidate receives as a function of his TV advertisment allocation. The candidate seeks to maximize the amount of voter support. How should he allocate spending across markets?

This is a *constrained optimization problem*, where the feasible set of allocations is defined by the *constraint condition*. In the case of our Senate candidate, he must spend B dollars on TV advertising, or

$$\sum_i x_i = B$$

This equation defines the $G(\boldsymbol{\theta})$ constraint set defined in our general optimization problem 1.[10] We now have

[10] $\max_{\boldsymbol{x} \in G(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \sum_i v_i(x_i) \\ \text{subject to} \quad & \sum_i x_i = B \end{aligned} \tag{3}$$

How can we solve this problem? The first order condition approach outlined above will clearly not work, because $v_i(x_i)$ are increasing, implying there are no stationary points. Now, we look for stationarity along the constraint set – among feasible allocations, which maximizes $\sum_i v_i(x_i)$?

As we did above, let's work axiomatically to characterize what an optimal solution to 3 looks like. Suppose $\boldsymbol{x}^\star$ is the optimal allocation across markets and satisfies the budget constraint

$$\sum_i x_i^\star = B$$

If this is true, then we can't rearrange the elements of $x^\star$ in a manner that satifies the budget constraint and increases total vote share. Consider a vector of proposed changes $\boldsymbol{dx}$. Satisfying the budget constraint requires

$$\sum_i dx_i = 0 \tag{4}$$

If $\boldsymbol{x}^\star$ is optimal, then we also know

$$\sum_i v_i(x_i^\star) \geq \underbrace{\sum_i v_i(x_i^\star + dx_i)}_{\text{I}}$$

We can approximate (I) with a derivative, giving

$$\sum_i v_i(x_i^\star + dx_i) \approx \sum_i v_i(x_i^\star) + Dv_i[x_i^\star]dx_i$$

Combining these gives

$$\sum_i v_i(x_i^\star) \geq \sum_i v_i(x_i^\star + dx_i) \approx \sum_i v_i(x_i^\star) + Dv_i[x_i^\star]dx_i$$

implying

$$0 \geq \sum_i Dv_i[x_i^\star]dx_i \tag{5}$$

Consider the case in which there are only two TV markets in the state. Then we have

$$0 \geq Dv_1[x_1^\star]dx_1 + Dv_2[x_2^\star]dx_2$$

and the constraint implies

$$dx_1 + dx_2 = 0 \implies dx_2 = -dx_1$$

Then,

$$\begin{aligned} 0 \geq Dv_1[x_1^\star]dx_1 - Dv_2[x_2^\star]dx_1 \\ Dv_2[x_2^\star]dx_1 \geq Dv_1[x_1^\star]dx_1 \\ Dv_2[x_2^\star] \geq Dv_1[x_1^\star] \end{aligned}$$

But we also have $dx_1 = -dx_2$, giving

$$\begin{aligned} 0 \geq -Dv_1[x_1^\star]dx_2 + Dv_2[x_2^\star]dx_2 \\ Dv_1[x_1^\star]dx_2 \geq Dv_2[x_2^\star]dx_2 \\ Dv_1[x_1^\star] \geq Dv_2[x_2^\star] \end{aligned}$$

We conclude

$$Dv_1[x_1^\star] = Dv_2[x_2^\star]$$

This condition states that the *marginal vote share* gained by spending in market 1 must equal the marginal vote share gained by spending in market 2 at the optimal allocation. The optimal allocation is therefore stationary with respect to the constraint set. We can see the geometry implied by this condition below.

The intuition applies more generally. Return to the case of many TV markets over which to allocate spending. If we consider an arbitrary set of $\boldsymbol{dx}$, then Inequality 5 implies

$$Dv_i[x_i^\star] = Dv_j[x_j^\star] \qquad \text{for all } i, j$$

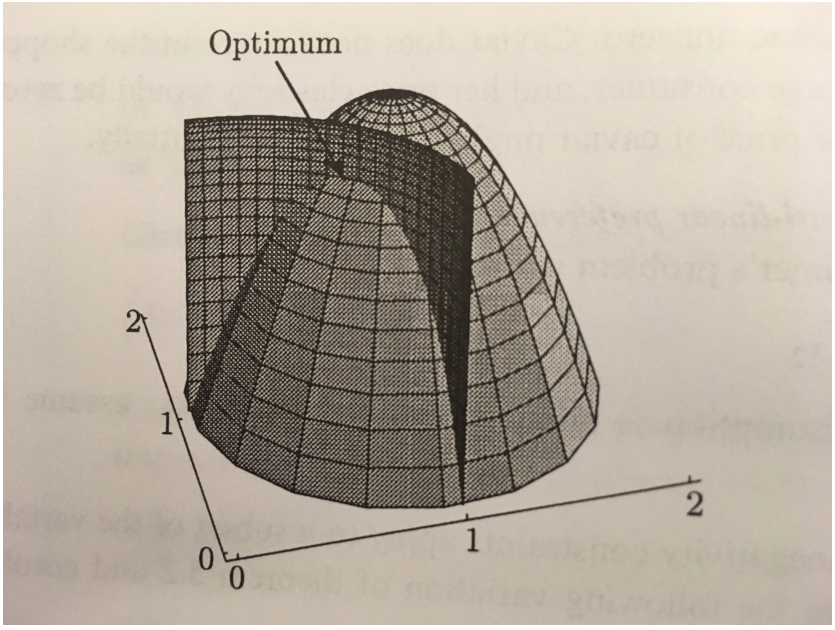

Figure 2: Carter pg. 526

The Method of Lagrange

In practice, we usually solve these sorts of problems with the *method of Lagrange*, which converts constrained optimization problems into unconstrained problems whose solution implicitly satisfies our constraint. The idea is to bake the constraints we've derived into the first order condtions of an unconstrained optimization problem. First, we'll state the theorem and then show how it works in the context of our TV advertisement problem.

Theorem (Lagrange): If $\boldsymbol{x}^\star$ is a local optimum[11] of

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g(\boldsymbol{x}) = 0 \end{aligned} \tag{6}$$

Then there exists a unique set of multipliers $\boldsymbol{\lambda}$ such that

$$\nabla f(\boldsymbol{x}^\star) = \boldsymbol{\lambda}^T \nabla g(\boldsymbol{x}^\star)$$

When we have a problem that satisfies these conditions, we can write the constrained problem as an unconstrained problem with a penalty, or

$$\max_{\boldsymbol{x}} \quad \mathcal{L} = f(\boldsymbol{x}) - \boldsymbol{\lambda}^T g(\boldsymbol{x})$$

whose first order conditions yield the multiplier FOC in the Theorem. We call the maximand in this setting the *Legrangian* and denote it

[11] Notice that the method only guarantees that it will find local optima. If we want the method to identify global optima, then we need the objective function to be concave along the constraint set. It is easy to see that the objective function pictured in Figure X satisfies this condition. But we will not go into more analytic depth here.

with $\mathcal{L}$. Going back to our TV advertisement example, we can write the constraint as

$$B - \sum_i x_i = 0$$

which gives the Legrangian

$$\mathcal{L} = \sum_i v_i(x_i) - \lambda \left(B - \sum_i x_i \right)$$

This yields the following set order conditions

$$\begin{aligned} v_1'(x_1) &= \lambda \\ \cdots &= \cdots \\ v_n'(x_n) &= \lambda \end{aligned}$$

Notice that this replicates the condition we derived using the derivative approximation approach, with

$$\lambda = Dv_i[x_i^\star] = Dv_j[x_j^\star] \qquad \text{for all } i, j$$

The marginal utility of TV ad spending must be equalized across markets. We combine this set of conditions with the constraint condition $B = \sum_i x_i$ to get a set of $N + 1$ equations that can be solved for λ and $\boldsymbol{x}$.

Example: Utility Maximization[12]

$$\begin{aligned} &\max_{x,y} x^\alpha y^\beta \\ &\text{subject to } x + y = B \end{aligned} \tag{7}$$

[12] Assume α and β are such that $x^\alpha y^\beta$ is concave... you found a sufficient condition for this property on your last problem set.

Example: Ridge Regression It is sometimes helpful to think about the Lagrange multiplier as a "penalty" enacted on solutions that violate the constraint. Consider again the linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

Under some conditions, we might worry that the line of best fit given by the $\hat{\boldsymbol{\beta}}$ vector might mistake noise for signal, and find a relationship between variables that is actually due to random noise. We call this phenomenon *overfitting*. The method of *ridge regression* provides a simple fix to the problem of overfitting.

By penalizing complexity in the $\hat{\boldsymbol{\beta}}$ vector, we balance a tradeoff between model fit and model parsimony. The ridge estimator $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ solves the following constrained optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right)^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_2 \leq \lambda \end{aligned} \tag{8}$$

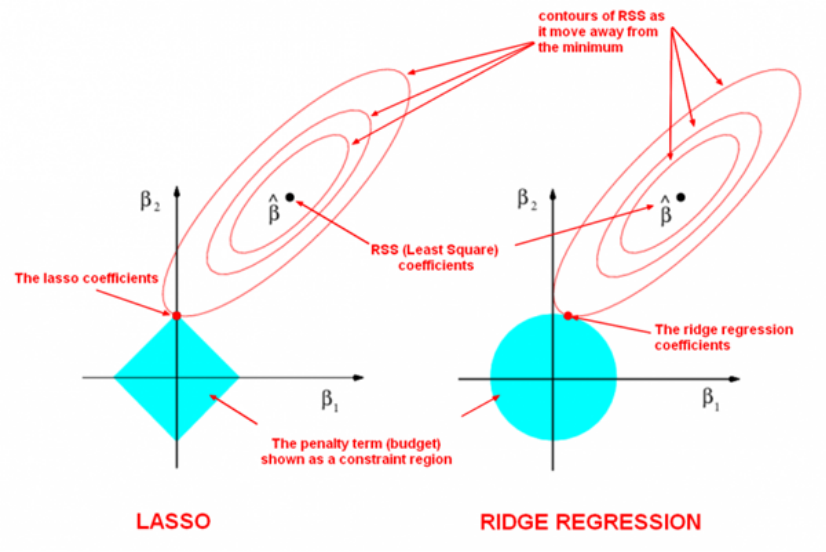

Figure 3: https://www.quora.com/How-would-you-describe-the-difference-between-linear-regression-lasso-regression-and-ridge-regression

$\|\boldsymbol{\beta}\|_2$ is a measure of the complexity of the $\boldsymbol{\beta}$ vector.[13] The constraint demands that this vector not become too complex. Alternatively, we can formulate the problem using the method of Lagrange

$$\mathcal{L} = \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 - \lambda \|\boldsymbol{\beta}\|_2$$

We see that $\lambda\|\boldsymbol{\beta}\|_2$ acts as a *penalty* on the objective function, decreasing $\mathcal{L}$ as $\boldsymbol{\beta}$ becomes more complex. This λ is called a *tuning parameter* and is usually strategically chosen to maximize out-of-sample prediction quality of a model.

[13] See lecture notes on Normed Linear Spaces.

References

1. Carter, Michael. *Foundations of Mathematical Economics*. Chapter 5.