Introduction to Mathematics for Political Science: Smooth Functions

4 September 2018

Linear functions are rarely sufficient to describe political phenomena of interest. Smooth functions are often better suited for this purpose. Smoothness is particularly important feature in optimization problems. You've encountered smooth functions and optimization in a 1-dimensional setting. Today, we generalize these intuitions to multi-dimensional spaces, preparing ourselves to approximate and optimize functions of many variables.

Smoothness in One Dimension (Univariate Calculus)

Consider a *policy production function*, $p : y \rightarrow \mathbb{R}_+$, which maps a legislator's staff size ($y \in \mathbb{R}_+$) to a number of bills produced during a session of congress.[1] Assume this function is continuous – small changes in staff size produce small changes in policy output. You learned what it means for this function to be *smooth* in your self-study of calculus. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ was said to be *differentiable* at x if

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

exists.[2] A function is said to be smooth if it is everywhere differentiable. Geometrically, a function is differentiable at x if we can draw a *line approximating* f that is tangent to $f(x)$ at x. The derivative at x is the slope of this line at x and is given by

$$\alpha = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Rearranging gives

$$\begin{aligned}
\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} - \alpha &= 0 \\
\lim_{h \rightarrow 0} \frac{f(x+h) - f(x) - \alpha h}{h} &= 0 \\
\lim_{h \rightarrow 0} \frac{f(x+h) - f(x) - g(x)}{h} &= 0
\end{aligned}$$

where $g(x) = \alpha h$ is a linear function approximating f in the neighborhood of x.[3] The difference $f(x+h) - f(x) - g(x)$ is the *approximation error* that $g(x)$ seeks to minimize in the neighborhood of x

$$\epsilon(x, h) = f(x+h) - f(x) - g(x)$$

[1] We might parameterize this function as follows

$$p(y) = y^{\beta}$$

for $\beta > 0$

[2] Give an example of a continuous but not smooth function. How do these concepts differ from one another?

[3] Note that the function $f(x) + g(x)$ is not linear but *affine*.

For the derivative to exist, we must have

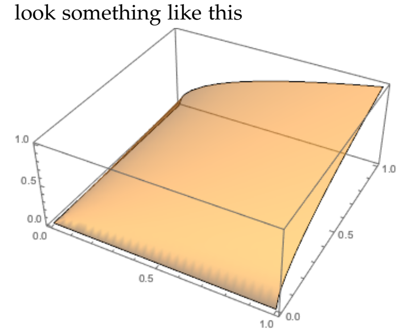
$$\lim_{h \to 0} \epsilon(x, h) = 0$$

Policy productivity is obviously more complicated than we postulated when writing $p(y)$. Consider the slightly more complex $p : \{x \times y\} \to \mathbb{R}_+$, where x is the legislator's policy expertise $x \in \mathbb{R}_+$ and y is again the number of staff members.[4] What does it mean for this function to be smooth? What about a generic $f : \mathbb{R}^n \to \mathbb{R}^n$? We want to generalize our one-dimensional notion of smoothness to these arbitrary spaces. We say a function is *smooth* if it *can be well approximated locally by a linear function*.

[4] We might parameterize this function with

$$p(x, y) = x^\alpha y^\beta$$

Such a function would look something like this



Smooth Functions and Differentiability

Formally, let $f : X \to Y$ be a function between normed linear spaces. We want to find a linear function $g(\boldsymbol{x})$ that well approximates f in the neighborhood of a particular element $\boldsymbol{x}_0 \in X$ as we move away from $\boldsymbol{x}_0$ in the direction of some $\boldsymbol{x}$.

$$f(\boldsymbol{x}_0 + x) \approx f(\boldsymbol{x}_0) + g(\boldsymbol{x})$$

The approximation error in this setting is given by

$$\epsilon(\boldsymbol{x}) = f(\boldsymbol{x}_0 + \boldsymbol{x}) - (f(\boldsymbol{x}_0) + g(\boldsymbol{x}))$$

We can scale the approximation error by its distance from $\boldsymbol{x}_0$, giving us a familiar quotient

$$\eta(\boldsymbol{x}) = \frac{\epsilon(\boldsymbol{x})}{\|\boldsymbol{x}\|} = \frac{f(\boldsymbol{x}_0 + \boldsymbol{x}) - (f(\boldsymbol{x}_0) + g(\boldsymbol{x}))}{\|\boldsymbol{x}\|}$$

We want $\eta(\boldsymbol{x})$ to get small as $\boldsymbol{x}$ gets small. This presents a natural definition for differentiability.

Definition: A function $f : X \to Y$ is *differentiable* at $\boldsymbol{x}_0 \in X$ if there exists a linear function $g : X \to Y$ such that for all $\boldsymbol{x} \in X$,

$$f(\boldsymbol{x}_0 + \boldsymbol{x}) = f(\boldsymbol{x}_0) + g(\boldsymbol{x}) + \eta(\boldsymbol{x})\|\boldsymbol{x}\|$$

and $\boldsymbol{x} \to \boldsymbol{0}_X \implies \eta(\boldsymbol{x}) \to \boldsymbol{0}_Y$.

Note: We'll call this quantity the derivative of f at $\boldsymbol{x}_0$ and denote it with $Df[\boldsymbol{x}_0]$ or $f'[\boldsymbol{x}_0]$.

Definition: A function $f : X \to Y$ is differentiable if it is differentiable at all $\boldsymbol{x}_0 \in X$. The derivative defines a function $Df : X \to \{X \times Y\}$.

Note: If Df is a continuous function then we say f is *continuously differentiable* or C^1. If Df itself is differentiable and its derivative is a continuous function then we say f is *twice continuously differentiable* or C^2, and so on.

Proposition: If a function is differentiable at $\boldsymbol{x}_0$ then it is continuous at $\boldsymbol{x}_0$.

Proof: Take a sequence $\boldsymbol{x}^m \to \boldsymbol{x}_0$. For a function f to be continuous at $\boldsymbol{x}_0$, we must have $\boldsymbol{x}^m \to \boldsymbol{x}_0 \implies f(\boldsymbol{x}^m) \to f(\boldsymbol{x}_0)$.[5] Let $\boldsymbol{x}^m = \boldsymbol{x}_0 + \boldsymbol{x}^n$ with $\boldsymbol{x}^n n \to \mathbf{0}_X \implies \boldsymbol{x}^m \to \boldsymbol{x}_0$. If f is differentiable, then, $\boldsymbol{x}^n \to \mathbf{0}_X \implies \eta(\boldsymbol{x}^n) \to \mathbf{0}_Y$ and

[5] See Definition 3 in the notes on continuous functions.

$$\begin{aligned} f(\boldsymbol{x}_0 + \boldsymbol{x}^n) - (f(\boldsymbol{x}_0) + g(\boldsymbol{x}^n)) &\to \mathbf{0}_Y \\ f(x^m) - (f(\boldsymbol{x}_0) + g(\boldsymbol{x}^n)) &\to \mathbf{0}_Y \\ f(x^m) &\to f(\boldsymbol{x}_0) + g(\boldsymbol{x}^n) \\ f(x^m) &\to f(\boldsymbol{x}_0) \end{aligned}$$

Since $\boldsymbol{x}^n \to \mathbf{0}_X$ and g is a linear function. This demonstrates $\boldsymbol{x}^m \to \boldsymbol{x}_0 \implies f(x^m) \to f(\boldsymbol{x}_0)$ as desired.

Note the analogue here with linear regression. Estimation of a regression coefficient entails searching for a linear function that minimizes the distance itself between data ($\boldsymbol{x}$). Formally, outcomes y are modeled as a function of the linear approximation, an error term, and an intercept

$$\begin{aligned} y &= f(\boldsymbol{x}_0) + g(\boldsymbol{x}) + \epsilon \\ &= \alpha + \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon \end{aligned}$$

Derivatives help us construct affine approximations of smooth functions. The approximate change in a function f in the neighborhood of $\boldsymbol{x}_0$ is

$$df = f(\boldsymbol{x}_0 + \boldsymbol{dx}) - f(\boldsymbol{x}_0) \approx Df[\boldsymbol{x}_0](\boldsymbol{dx}) f(\boldsymbol{x}_0 + \boldsymbol{dx}) \quad \approx f(\boldsymbol{x}_0) + Df[\boldsymbol{x}_0](\boldsymbol{dx})$$

Partial Derivatives, the Gradient Vector, and the Hessian Matrix

Considering the set of *functionals* allows us to consider how the value of a function $f : \mathbb{R}^n \to \mathbb{R}$ changes as a function of a single variable, holding all other variables constant. Let

$$h_i(t) = f(x_1^0, ..., t, ..., x_n^0)$$

denote the value of the ith partial of f at $\boldsymbol{x}^0$.

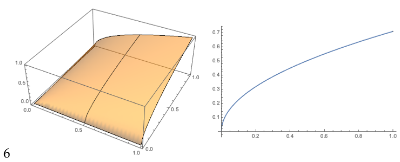
Definition: The *ith partial derivative* of a differentiable function f at $\boldsymbol{x}^0$ is

$$\lim_{t \to 0} \frac{h(x_i^0 + t) - h(x_i^0)}{t}$$

We denote these partial derivatives with

$$\frac{\partial f[\boldsymbol{x}^0]}{\partial \boldsymbol{x}_i}$$

Geometrically, we're asking for the slope of the function when sliced at the ith cross section of f through $\boldsymbol{x}^0$.[6] Why should we care about partial derivatives?



6

Example: Consider the policy production function again, $p(y, x)$. Suppose the candidates policy expertise is fixed. What is the marginal effect of adding staff members on the policy output? This is precisely what the partial derivative gives us.

Iteratively taking partial derivatives of a function f would tell us how the function is changing in every direction. We could collect these partial derivatives into a vector, called the *gradient*, which summarizes all of these changes.[7]

[7] The ∇ you see in the notes can be written backslash "nabla" in LaTeX

Definition: The *gradient* of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at $\boldsymbol{x}^0$ is

$$\nabla f(\boldsymbol{x}^0) = \left(\frac{\partial f[\boldsymbol{x}^0]}{\partial x_1}, ..., \frac{\partial f[\boldsymbol{x}^0]}{\partial x_n} \right)$$

and

$$Df[\boldsymbol{x}^0](\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\partial f[\boldsymbol{x}^0]}{\partial x_i} x_i^0$$

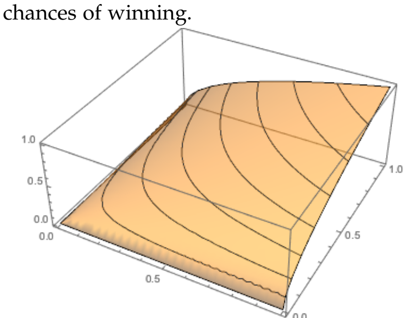
We see that for functions on $\mathbb{R}^n$, we can decompose the derivative into a sum of partial derivatives.[8] The gradient is a vector, and it has interesting geometric properties. In particular, it points in the direction of greatest change.[9] To get a better understanding of this geometry, it's useful to define contours.

[8] Note that the gradient can be expressed as the inner product $\nabla f(\boldsymbol{x}^0)^T \boldsymbol{x}$

[9] If you were a campaign strategist advising a candidate facing the election function $p(y, x)$, knowing the gradient would be useful for improving your chances of winning.

Definition: The contour of a functional f through $c = f(\boldsymbol{x}^0)$ is[10]

$$f^{-1}(c) = \{\boldsymbol{x} \in X | f(\boldsymbol{x} = c\}$$

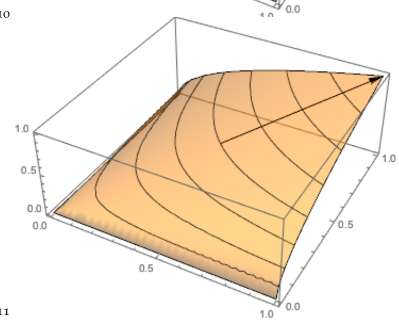


10

Think about our election function again. The concept of the contour is simple in this setting. $p^{-1}(c)$ gives us the set of $\{x \times y\}$ that give the candidate a c chance of victory. There are likely to be an infinity of potential $x \times y$ combinations of campaign spending and policy that could produce this result. Contours are sometimes useful for visualizing high-dimensional functions.

Remark: The gradient is orthogonal to the contour[11]

$$\nabla f(\boldsymbol{x}^0)^T f^{-1}(c) = 0$$



11

If a function is twice differentiable, we can compute its *second-order partial derivatives* by differentiating the partial derivatives with respect to another variable. We write these

$$\frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_i \partial x_j}$$

When $i \neq j$, we call this the *cross partial* of f at $\boldsymbol{x}^0$. If $i = j$ we write

$$\frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_i^2}$$

Example: The cross partial of the election function would be written

$$\frac{\partial p(y,x)}{\partial y \partial x}$$

and describes how the effectiveness of campaign spending changes with respect to the candidate's policy position (or vice versa).

Notice that there are many cross partials we can take as n gets large.[12] The gradient vector can be differentiated with respect to each of its elements. We store these derivatives in the *Hessian* matrix.

[12] n^2 to be precise.

Definition: The *Hessian* of a function f at $\boldsymbol{x}^0$ is a matrix of second-order derivatives, with

$$H_f(\boldsymbol{x}^0) = \begin{bmatrix} \frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_1^2} & \cdots & \frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f[\boldsymbol{x}^0]}{\partial x_n^2} \end{bmatrix}$$

We can describe the concavity or convexity of a high-dimensional function using its Hessian.

Proposition: A twice-differentiable function f is strictly locally convex at $\boldsymbol{x}$ iff $H_f(\boldsymbol{x})$ is positive definite. The function is strictly locally concave at $\boldsymbol{x}$ iff $H_f(\boldsymbol{x})$ is negative definite.

The Jacobian

We often represent political systems as systems of equations. In a multi-candidate election, for example, we might have a family of policy production $p_i(\boldsymbol{y}, \boldsymbol{x})$ that map *every* legislator's expertise and staff size into policy output.[13] The *Jacobian* is a matrix that stores a linear approximation of this system.

[13] Notice that the arguments of the function p are now vectors.

More formally, consider a system of m functionals, with each $f_m : \mathbb{R}^n \rightarrow \mathbb{R}$.[14]

[14] Note that the system $\boldsymbol{f} : \mathbb{R}^n \mathbb{R}^m$, because each of the m functionals returns a single real number.

Definition: The *Jacobian* of a system of functionals $\boldsymbol{f}$ is a matrix of partial derivatives, where each element ij is the ith partial of the jth function evaluated at $\boldsymbol{x}^0$

$$J_f(\boldsymbol{x}^0) = \begin{bmatrix} \frac{\partial f_1[\boldsymbol{x}^0]}{\partial \boldsymbol{x}_1} & \cdots & \frac{\partial f[\boldsymbol{x}^0]}{\partial \boldsymbol{x}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m[\boldsymbol{x}^0]}{\partial \boldsymbol{x}_1} & \cdots & \frac{\partial f_m[\boldsymbol{x}^0]}{\partial \boldsymbol{x}_n} \end{bmatrix}$$

Notice that the Jacobian can also be written as a vector of gradients stacked on top of one another

$$J_f(\boldsymbol{x}^0) = \begin{pmatrix} \nabla f_1(\boldsymbol{x}^0) \\ \vdots \\ \nabla f_m(\boldsymbol{x}^0) \end{pmatrix}$$

References

1. Carter, Michael. *Foundations of Mathematical Economics*. Chapter 4.