Optimization

Class 7: Scaling Factors

1 Nearly Stationary

Newton's optimization method (and many other methods) for optimizing f tend stop progressing when the current guess becomes *nearly stationary* by having gradient ∇f nearly the zero-vector.

$$\begin{bmatrix} x_{\text{new}} \\ y_{\text{new}} \end{bmatrix} = \begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} - \nabla^2 f(x_{\text{old}}, y_{\text{old}})^{-1} \nabla f(x_{\text{old}}, y_{\text{old}}) \quad \text{NEWTON UPDATE}$$

$$\Downarrow$$

$$\begin{bmatrix} x_{\text{new}} \\ y_{\text{new}} \end{bmatrix} \approx \begin{bmatrix} x_{\text{old}} \\ y_{\text{old}} \end{bmatrix} \quad \Longleftrightarrow \quad \nabla^2 f(x_{\text{old}}, y_{\text{old}})^{-1} \nabla f(x_{\text{old}}, y_{\text{old}}) \approx \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

One possibility here is that the inverse of the second-derivative matrix is nearly zeroing out the matrix-vector multiplication. A more satisfying possibility (since we're seeking stationary points) is that the gradient vector $\nabla f(x_{\text{old}}, y_{\text{old}})$ itself is nearly the zero-vector. We can always track the components of the gradient vector to see if this is the case:

x	y	fx	fy
50.	1.	1	1
54.789	14.121	0.0584058	0.353042
54.2671	12.6756	0.0227975	0.141721
56.0322	10.5546	0.00851942	0.0558006
60.6215	7.66995	0.00314236	0.0213464
66.2276	4.45039	0.0012233	0.00761365
70.1934	1.86954	0.000412774	0.00209785
71.6872	0.714	0.0000615384	0.000270094
71.8931	0.536726	1.32855×10^{-6}	5.55492×10^{-6}
71.8973	0.533041	5.74076×10^{-10}	2.38742×10^{-9}
71.8973	0.53304	1.00614×10^{-16}	4.51028×10^{-16}
71.8973	0.53304	-1.38778×10^{-17}	0.

Q: Does this mean we are near a stationary point?

A: No.

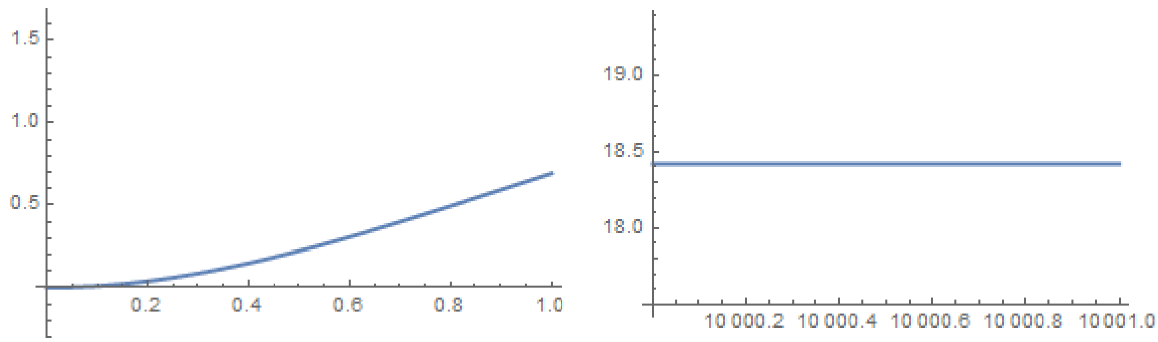

Figure 1: The stationary point (at $\bar{x} = 0$) and a nearly stationary point (at $x = 10,000.6$).

See Nearly stationary not near stationary point in Class7.nb.

This example shows that nearly stationary points can be a long way from a true stationary point, and that the function value at a nearly stationary point can be very far from optimal. We need a test to give us confidence that our nearly stationary current guess is not a long way from a true stationary point.

2 Reverse continuity

- Continuity of F at $\bar{x}$: inputs $x \to \bar{x} \Rightarrow$ outputs $F(x) \to F(\bar{x})$

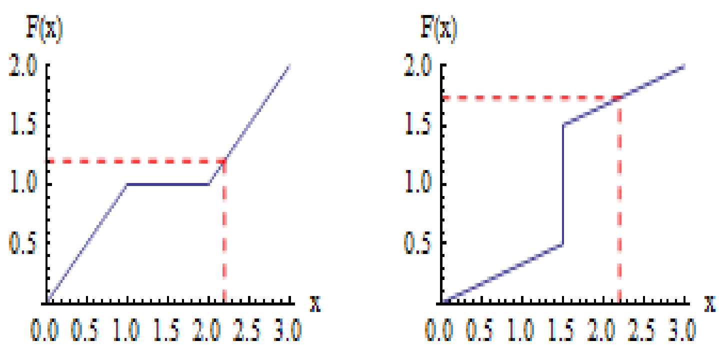

Figure 2: A function that is continuous everywhere, and one that is not continuous at $\bar{x} = 1.5$

- Reverse continuity of F at $\bar{x}$: outputs $F(x) \to F(\bar{x}) \Rightarrow$ inputs $x \to \bar{x}$

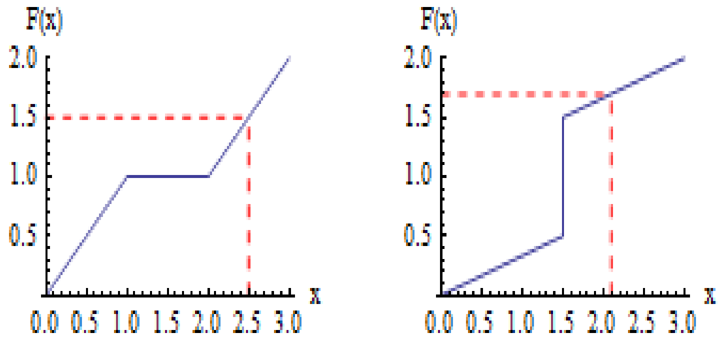

Figure 3: A function that is not reverse continuous at any $\bar{x} \in [1, 2]$, and one that is reverse continuous everywhere

See Continuity and Reverse Continuity in Class7.nb.

We want reverse continuity of the objective gradient $F = \nabla f$ near a stationary point $\bar{x}$; then we can deduce something we don't know (that the nearly stationary points x approach the stationary point $\bar{x}$) from something we can test (that the output $\nabla f(x)$ approaches zero).

The objective $f(x) = \log(x^2 + 1)$ was used to generate the images at the beginning of class showing a nearly stationary point at $x = 10,000.6$ with the true stationary point at $\bar{x} = 0$. Here is a graph of $f'(x) = \frac{2x}{x^2+1}$:

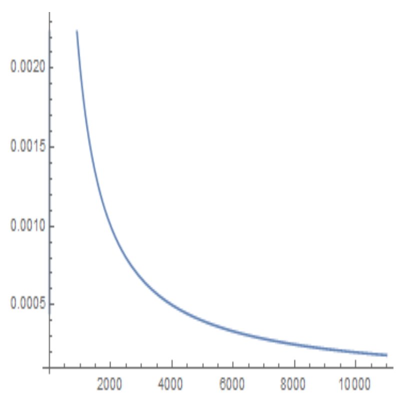

Figure 4: Reverse continuity of $f'(x) = \frac{2x}{x^2+1}$

This derivative function appears to be reverse continuous at the nearly stationary point $x = 10,000.6$....but look at the scale on the vertical axis. This graph is actually relatively close to horizontal, and so is practically like the image above (which is not reverse continuous).

See A look at $f'(x)$ for the first example above in Class7.nb.

3 Scaling factor

We need to know the scale of closeness of the inputs of a reverse continuous function (and not just that they are close). For a single-variable function $f(x)$, we want

$$|x - \bar{x}| \leq \frac{|f'(x) - f'(\bar{x})|}{L} \tag{1}$$

for a fixed *scaling factor* $L > 0$.

Group Problem 3.7.

(a) Explain why this relationship guarantees that f' is reverse continuous at the point $\bar{x}$.

(b) Suppose $L = 1$. If $f'(x)$ agrees with zero up to 6 decimal places, how close is x guaranteed to be to a stationary point $\bar{x}$?

(c) What if instead $L = 1000$?

(d) In general, is a relatively large or a relatively small L more desirable?

Group Problem 3.7. (solutions)

(a) Any f' satisfying $L|x - \bar{x}| \leq \frac{|f'(x)-f'(\bar{x})|}{L}$ is reverse continuous at the point $\bar{x}$ since the input distances $|x - \bar{x}|$ decrease by the factor $1/L$ of the output distances $|f'(x) - f'(\bar{x})|$ as the latter decrease.

(b) If the derivative satisfies $|x - \bar{x}| \leq |f'(x) - f'(\bar{x})|$, and $f'(x)$ agrees with zero up to 6 decimal places, we know that x is within 10^{-6} of the stationary point $\bar{x}$ (since $f'(\bar{x}) = 0$ by the definition of being a stationary point).

(c) If instead $|x - \bar{x}| \leq \frac{|f'(x)-f'(\bar{x})|}{1000}$, we know that x is within $10^{-6}/1000 = 10^{-9}$ of the stationary point $\bar{x}$.

(d) Larger L are more desirable. In particular, any $L > 1$ ensures that inputs are closer than outputs.

3.1 Graphical implication

Cross-multiplying the scaling factor relationship (1) gives:

$$L \leq \frac{|f'(x) - f'(\bar{x})|}{|x - \bar{x}|}.$$

The term on the right is the magnitude of the slope of the secant line through the base point $(\bar{x}, f'(\bar{x}))$ on the graph of f' and any other point $(x, f'(x))$ on the graph of f'.

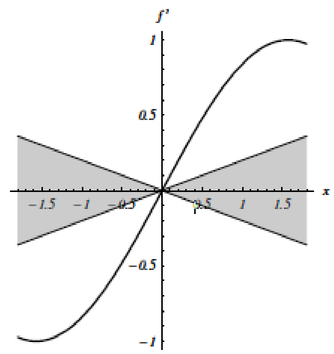

Figure 5: Graphical implication of scaling factor with $\bar{x} = 0$

As long as the graph of f' avoids the shaded bowtie, any secant line through $\big(0, f'(0)\big)$ will also avoid these shaded regions. It follows that the magnitude of the slope of any such secant line will be at least as great as the value of L used to construct the bowtie. Recall that larger scaling factors are better, which means fatter bowties!

Group Exercises 3.2.2.

(a) The bowtie above is centered on the graph of f' at the point $\big(\bar{x}, f'(\bar{x})\big) = (0, 0)$. Indicate the points $\bar{x}$ on the x-axis for which the same bowtie, but now centered on the graph of f' at the new point $\big(\bar{x}, f'(\bar{x})\big)$, is too fat to correspond to a scaling factor for f' there.

(b) Indicate the additional points $\bar{x}$ on the x-axis for which a miniature version of the same bowtie works.

(c) There are exactly two points where no matter how skinny or miniaturized the bowtie, there will be overlap with the graph of f'. Which are these and what value do they produce in the second-derivative f''?

Group Exercises 3.2.2. (solutions)

(a) Any center point $(\bar{x}, f'(\bar{x}))$ for which the graph somewhere intersects the bowtie indicates a place where this bowtie is too fat. This includes all the pairs $(\bar{x}, f'(\bar{x}))$ for which $|\bar{x}| \geq .75$.

(b) As we zoom in arbitrarily close to the graph of f', the graph begins to resemble its tangent line. So, any points along the graph of f' where the tangent line has slope between L and $-L$ will generate an intersection, no matter how much we miniaturize the bowtie. This corresponds to x-values near $\bar{x} = -1.5$ and $\bar{x} = 1.5$; so any other points work.

(c) No matter what scaling factor $L > 0$ is used, bowties constructed at the peak and valley of the graph of f' will not work since the graph is flat there. The corresponding x-values are $\bar{x} = -1.5$ and $\bar{x} = 1.5$, and the second derivative at both is zero.

NOTE: Miniaturizing corresponds to local optimizers; so apparently scaling factors $L > 0$ will work locally as long as $L < |f''(\bar{x})|$.

See Animated bow-tie in Class7.nb.

3.2 Multi-variable scaling factors

$$L \left\|\mathbf{x} - \bar{\mathbf{x}}\right\| \leq \left\|\nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}})\right\|.$$

where the *magnitude* of a vector is the square-root of the sum of the components squared:

$$\|\mathbf{x}\| = \left\| \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \right\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Scaling Factor Test

$$L := \min\{|\epsilon| \text{ over all eigenvalues } \epsilon \text{ of } \nabla^2 f(\bar{\mathbf{x}})\} > 0$$

$$\Downarrow$$

∇f is locally reverse continuous at $\bar{\mathbf{x}}$ with scaling factor L:

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \frac{\|\nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}})\|}{L} \quad \text{for } \mathbf{x} \text{ near } \bar{\mathbf{x}}$$

The minimal-magnitude eigenvalue of the second-derivative matrix at $\bar{\mathbf{x}}$ is the multi-variable analog of the term $|f''(\bar{x})|$ in the single-variable case.