

ARE211, Fall 2009

CALCULUS3: TUE, OCT 27, 2009

PRINTED: AUGUST 27, 2009

(LEC# 18)

CONTENTS

4. Univariate and Multivariate Differentiation (cont)	1
4.4. Multivariate Calculus: functions from \mathbb{R}^n to \mathbb{R}^m	2
4.5. Four graphical examples.	3
4.6. Taylor's Theorem	13

4. UNIVARIATE AND MULTIVARIATE DIFFERENTIATION (CONT)

Key Points:

- (1) for $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, understanding the differential as a *linear* map from \mathbb{R}^n to \mathbb{R}^m ,
- (2) understanding applications of the differential from \mathbb{R}^n to \mathbb{R}^m ,
- (3) Global Taylor theorem: if f is $(K + 1)$ times continuously differentiable, then for any $0 \leq k \leq K$, any $\bar{\mathbf{x}}$, and any $\mathbf{dx} \in \mathbb{R}^n$, there exists $\lambda \in [0, 1]$ such that

$$f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) = T^k(f, \bar{\mathbf{x}}, \mathbf{dx}) + \frac{Tf_{k+1}(\bar{\mathbf{x}} + \lambda \mathbf{dx}, \mathbf{dx})}{(k+1)!} \quad (\text{eq (1) on p. 15}).$$

- (4) Local Taylor theorem: if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $K + 1$ times continuously differentiable, then for any $\mathbf{x} \in \mathbb{R}^n$, any $0 \leq k \leq K$ and any $\mathbf{v} \in \mathbb{R}^n$ such that $Tf_k(\mathbf{x}, \mathbf{v}) \neq 0$, there exists $M \in \mathbb{N}$ such that for $m > M$, $|T^k(f, \mathbf{x}, \mathbf{v}/m)|$ strictly exceeds the absolute value of the remainder term (see p. 16).

4.4. Multivariate Calculus: functions from \mathbb{R}^n to \mathbb{R}^m

We'll now generalize what we did last time to a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. In general, if you have a function from \mathbb{R}^n to \mathbb{R}^m , what is the notion of slope (or gradient or derivative)? Not suprisingly, it is a $m \times n$ *matrix*. The *matrix* which is the derivative of a function from \mathbb{R}^n to \mathbb{R}^m is called the *Jacobian matrix* for that function.

Note well: I tend to talk about the Jacobian of a function, when what I mean is the Jacobian matrix. But this is potentially confusing. The Jacobian matrix has a determinant, which is called the Jacobian determinant. There are (respectable) books that use the unqualified word Jacobian to refer to the determinant, not the matrix. De Groot is one of these. So need to be aware of which is which.

Example: A particularly important function from \mathbb{R}^n to \mathbb{R}^n is the gradient of $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Specifically, think of the gradient as being n functions from $\mathbb{R}^n \rightarrow \mathbb{R}$, i.e., each of the partial derivatives

of f , stacked on top of each other: $\nabla f = \begin{bmatrix} f_1(\cdot) \\ \vdots \\ f_n(\cdot) \end{bmatrix}$. The derivative of the gradient function is

the matrix constructed by stacking the gradients of each of these partial derivatives *viewed as row*

vectors on top of each other, i.e., $\begin{bmatrix} \nabla f_1(\cdot) \\ \vdots \\ \nabla f_n(\cdot) \end{bmatrix}$. This derivative of the derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$, which

in generic language would be called the Jacobian of the gradient of f , is more concisely known as the *Hessian* of f .

More generally, to visualize the derivative and differential associated with an *arbitrary* function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, it is helpful to think of f , once again, as a vertical stack of m functions $f^i : \mathbb{R}^n \rightarrow \mathbb{R}$, all stacked on top of each other. (Notationally, the only difference between this and the previous paragraph is that now we use superscripts rather than subscripts to distinguish the functions from each other.) It is now natural to think of the derivative of f as a vertical stack of all the derivatives

(gradients) of the f^i 's. That is, $f'(\cdot) \equiv \mathbf{J}f(\cdot) = \begin{bmatrix} \nabla f^1(\cdot) \\ \nabla f^2(\cdot) \\ \vdots \\ \nabla f^m(\cdot) \end{bmatrix}$, where each $\nabla f^i(\cdot)$ is a row vector consisting of the partial derivatives of $f^i(\cdot)$.

Next think of the *differential* of ∇f at \mathbf{x} , i.e., the linear function $\mathbf{d}f^{\mathbf{x}}(\cdot) = \mathbf{J}f(\mathbf{x})(\cdot)$ as a vertical stack consisting of the differentials of the f^i 's at \mathbf{x} , i.e.,

$$\mathbf{d}f^{\mathbf{x}}(\mathbf{d}\mathbf{x}) = \mathbf{J}f(\mathbf{x})(\mathbf{d}\mathbf{x}) = \mathbf{J}f(\mathbf{x}) \cdot \mathbf{d}\mathbf{x} = \begin{bmatrix} \nabla f^1(\mathbf{x}) \cdot \mathbf{d}\mathbf{x} \\ \nabla f^2(\mathbf{x}) \cdot \mathbf{d}\mathbf{x} \\ \vdots \\ \nabla f^m(\mathbf{x}) \cdot \mathbf{d}\mathbf{x} \end{bmatrix}.$$

4.5. Four graphical examples.

We can now apply all the graphical intuitions we've developed from the last lecture about the differential of a real-valued function, to the general case: instead of considering one 3-D picture like Figure 1 in the previous lecture, you just visualize a stack of m such pictures.

The following example is intended to illustrate this idea.

We start out with a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Its gradient, then, maps \mathbb{R}^2 to \mathbb{R}^2 . The function we are interested in is graphed in Fig. 1. Note that the function *decreases* with both arguments so that the gradient is a strictly negative vector. We are interested in how the gradient changes in response to a small change \mathbf{dx} in the domain.

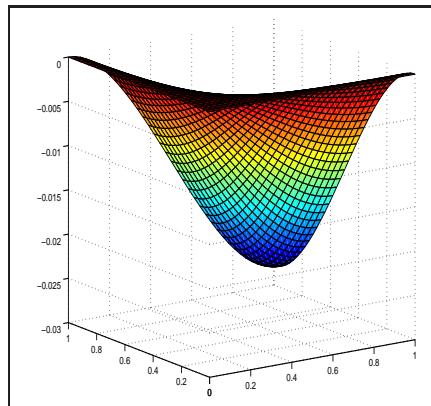


FIGURE 1. Graph of f

To get some intuition, it's helpful to return to the 3-D diagrams that we were looking at in the last lecture, as we do in Fig. 2 below.

It is

$$f(\mathbf{x}) = (x_1^2/2 - x_1^3/3)(x_2^3/3 - x_2^2/2)$$

whose gradient is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} (x_1 - x_1^2)(x_2^3/3 - x_2^2/2) \\ (x_2^2 - x_2)(x_1^2/2 - x_1^3/3) \end{bmatrix}$$

so that $d \nabla f^{\mathbf{x}}(\mathbf{dx}) = \mathbf{H}f(\mathbf{x}) \cdot \mathbf{dx} = \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_2(\mathbf{x}) \end{bmatrix} \mathbf{dx}$, where

$$\begin{aligned} \nabla f_1(\mathbf{x}) &= \begin{bmatrix} (1 - 2x_1)(x_2^3/3 - x_2^2/2) & (x_1 - x_1^2)(x_2^2 - x_2) \end{bmatrix} & \text{and} \\ \nabla f_2(\mathbf{x}) &= \begin{bmatrix} (x_2^2 - x_2)(x_1 - x_1^2) & (2x_2 - 1)(x_1^2/2 - x_1^3/3) \end{bmatrix} \end{aligned}$$

We'll evaluate the gradient of this function at the point $\mathbf{x} = [0.667, 0.667]$, and consider a shift in the domain of $\mathbf{dx} = [-0.1944, 0.2222]$, which takes us to the point $\mathbf{x} + \mathbf{dx} = [0.4722, 0.8889]$.

Plugging in the numbers, we obtain

$$\nabla f(\mathbf{x}) = \begin{bmatrix} -0.0274 \\ -0.0274 \end{bmatrix}; \quad \nabla f(\mathbf{x} + \mathbf{dx}) = \begin{bmatrix} -0.0401 \\ -0.0075 \end{bmatrix} \quad \text{so that} \quad \nabla f(\mathbf{x} + \mathbf{dx}) - \nabla f(\mathbf{x}) = \begin{bmatrix} -0.0127 \\ 0.0199 \end{bmatrix}$$

i.e., the first partial becomes *more* negative while the second becomes *less* so. Evaluating the differential of ∇f at \mathbf{x} at the magnitude of the change we obtain

$$d \nabla f^{\mathbf{x}}(\mathbf{dx}) = \mathbf{H}f(\mathbf{x}) \cdot \mathbf{dx} = \begin{bmatrix} 0.0412 & -0.0494 \\ -0.0494 & 0.0412 \end{bmatrix} \begin{bmatrix} -0.1944 \\ 0.2222 \end{bmatrix} = \begin{bmatrix} -0.0190 \\ 0.0187 \end{bmatrix}$$

graphical analog of these computations, we'll now do exactly what we were doing for a function mapping \mathbb{R}^2 to \mathbb{R} , except that we are going to look at two 3-D graphs simultaneously. **It's much easier to understand Fig. 2 if can view it in color, so if you don't have access to a color printer, you might want to look at it on a color screen.** Here's a guide to the colors:

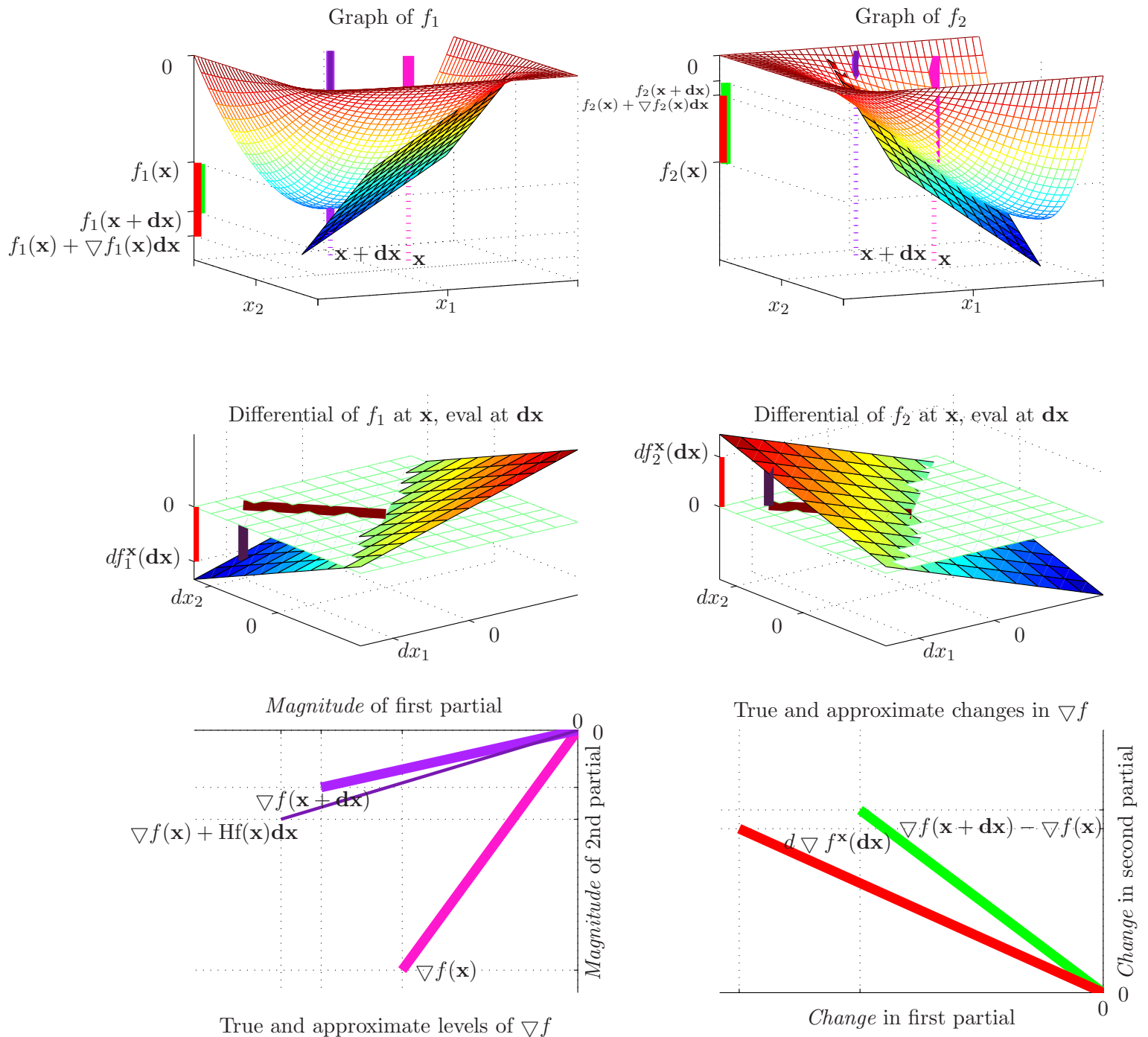


FIGURE 2. The differential approximation to a change in gradient

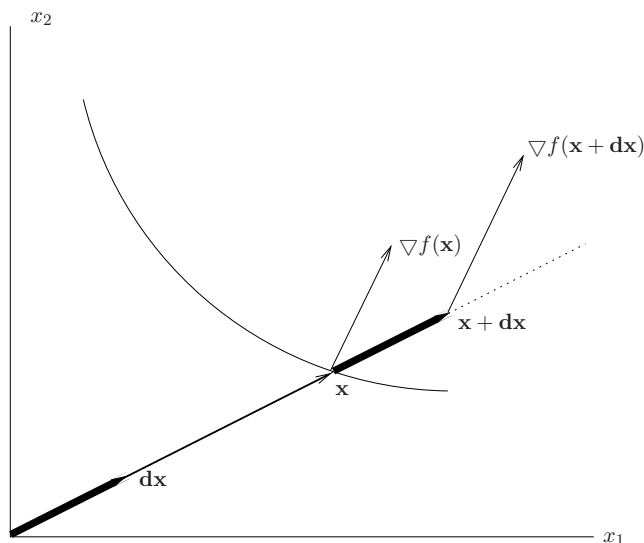
- The *level* of $\nabla f(\mathbf{x})$ is indicated by pink lines;
- The *level* of $\nabla f(\mathbf{x} + \mathbf{dx})$ is indicated by purple lines
- The *true change* in $\nabla f(\cdot)$ is indicated by green lines;
- The *evaluation of the differential* is indicated by red lines

Observe in Fig. 2 that because of the shape of $f_2(\cdot)$, the first order linear approximation to $f_2(\mathbf{x} + \mathbf{dx})$ is almost perfect, while the first order linear approximation to $f_1(\mathbf{x} + \mathbf{dx})$ is much less so. This is reflected in the bottom right panel, where there is a big gap between $(f_1(\mathbf{x} + \mathbf{dx}) - f_1(\mathbf{x}))$ and $df_1^{\mathbf{x}}(\mathbf{dx})$ and a negligible one between $(f_2(\mathbf{x} + \mathbf{dx}) - f_2(\mathbf{x}))$ and $df_2^{\mathbf{x}}(\mathbf{dx})$.

We now consider three more examples, using the differential of the gradient of f to explore how the gradient vector changes as we change \mathbf{x} . Since the gradient of f at \mathbf{x} is always perpendicular to the level set of f corresponding to $f(\mathbf{x})$, what we learn about these changes indirectly tells us about things like the curvature of the level set of f at \mathbf{x} . Here are a couple of examples, applied to the function $f(\mathbf{x}) = x_1x_2$.

Second example: The function $f(\mathbf{x}) = x_1x_2$, depicted in Fig. 3, is an example of a *homothetic* function, i.e., a function with the property that the *slopes* of its level sets are constant along rays through the origin. More precisely, if $\mathbf{y} = \alpha\mathbf{x}$, for some scalar $\alpha \in \mathbb{R}_+$, then the slope of the level set of f through \mathbf{y} is equal to the slope of the level set of f through \mathbf{x} . Since gradient vectors are perpendicular to level sets, this implies that the gradients of f at both \mathbf{x} and \mathbf{y} must point in the same direction. Let's check that this is true for this function.

$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{bmatrix} x_2 & x_1 \end{bmatrix} \\ \text{Hf}(\mathbf{x}) &= J \nabla f(\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

FIGURE 3. f is homothetic

so the differential of ∇f at \mathbf{x} is

$$d \nabla f(\mathbf{x})(d\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix}$$

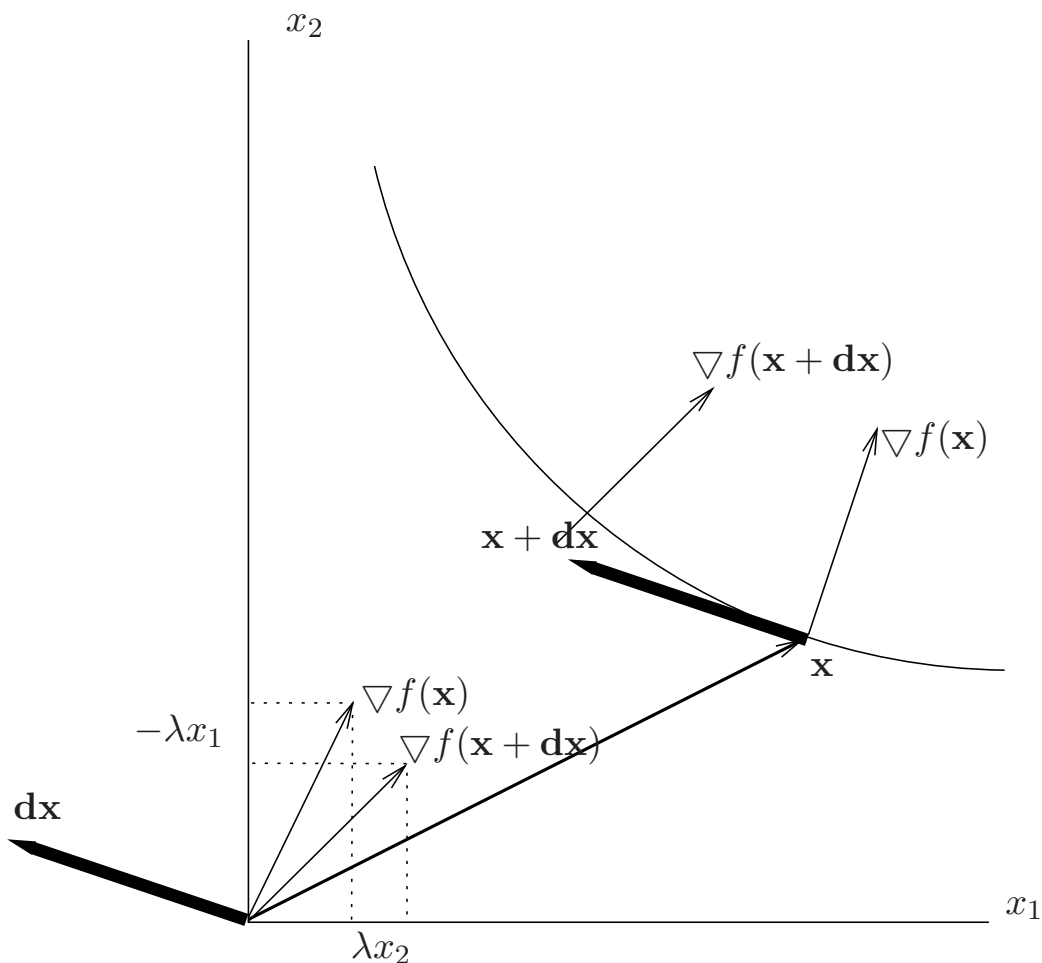
In this case $J \nabla f(\mathbf{x})$ is a constant, so that the higher order terms in the Taylor approx are all zero, so that the first approximation must be exactly correct. Now consider a move $d\mathbf{x}$ along the ray through the origin passing through \mathbf{x} , i.e., choose $d\mathbf{x} = \lambda\mathbf{x}$, for some scalar $\lambda > 0$. In this case, we have

$$d \nabla f(\mathbf{x})(d\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix} = \begin{bmatrix} \lambda x_2 \\ \lambda x_1 \end{bmatrix}$$

so that, taking a first order approximation to $\nabla f(\mathbf{x} + \lambda\mathbf{x})$:

$$\nabla f(\mathbf{x} + \lambda\mathbf{x}) \approx \nabla f(\mathbf{x}) + d \nabla f(\mathbf{x})(d\mathbf{x}) = \begin{bmatrix} (1 + \lambda)x_2 \\ (1 + \lambda)x_1 \end{bmatrix}$$

But in this case, we can replace the approximation symbol with an equality. That is, the gradient of f at $(1 + \lambda)\mathbf{x}$ is a scalar multiple of the gradient of f at \mathbf{x} , confirming homotheticity. (Note additionally that the gradient gets *longer* as you move out along a ray through the origin, indicating that f exhibits increasing returns to scale.)

FIGURE 4. f exhibits diminishing MRS

Third example (see Fig. 4): We'll now show that f exhibits diminishing marginal rate of substitution. Recall that the marginal rate of substitution of x_2 for x_1 is the ratio $\left| \frac{f_1(x)}{f_2(x)} \right|$. In Fig. 4, this is the length of the horizontal component of the gradient vector divided by the length of the vertical component. i.e., "run over rise." Diminishing MRS means that the gradient vector becomes flatter (steeper) as move to the northwest (south east) along a level set. We consider a northwesterly movement of x , and verify that the gradient vector becomes flatter. Fix \mathbf{x} and consider a north-west movement in the domain, orthogonal to the gradient of f . Since, $\nabla f(\mathbf{x}) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$ a north-west movement orthogonal to this vector would be $\mathbf{dx} = (-\lambda x_1, \lambda x_2)$. (Observe that $\nabla f(\mathbf{x}) \cdot \mathbf{dx} = -\lambda x_1 x_2 + \lambda x_1 x_2 = 0$, so that indeed \mathbf{dx} and $\nabla f(\mathbf{x})$ are orthogonal to each other).

Now

$$d \nabla f(\mathbf{x})(d\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\lambda x_1 \\ \lambda x_2 \end{bmatrix} = \begin{bmatrix} \lambda x_2 \\ -\lambda x_1 \end{bmatrix}$$

so that

$$\nabla f(\mathbf{x} + \lambda \mathbf{x}) = \nabla f(\mathbf{x}) + d \nabla f(\mathbf{x})(d\mathbf{x}) = \begin{bmatrix} (1 + \lambda)x_2 \\ (1 - \lambda)x_1 \end{bmatrix}$$

i.e., the partial with respect to x_1 gets bigger while the partial with respect to x_2 gets smaller, i.e., the gradient gets *flatter*.

Fourth example: Consider the demand system: $\boldsymbol{\xi}(\mathbf{p}) = \begin{pmatrix} \xi^1(\mathbf{p}) \\ \vdots \\ \xi^n(\mathbf{p}) \end{pmatrix}$. The Jacobian of this function

is written as $J\boldsymbol{\xi}(\cdot)$. Note that I'm using superscripts rather than subscripts, to distinguish between the components of an arbitrary vector-valued function (here the system of demand equations) and the specific vector valued function which is the gradient, i.e., vector of partial derivatives. Start out at $\bar{\mathbf{p}}$. Want to know the effect of a change in the price vector from $\bar{\mathbf{p}}$ to \mathbf{p} :

$$\begin{aligned} & \boldsymbol{\xi}(\mathbf{p}) - \boldsymbol{\xi}(\bar{\mathbf{p}}) \\ & \approx d\boldsymbol{\xi} \\ & = J\boldsymbol{\xi}(\bar{\mathbf{p}})(\mathbf{p} - \bar{\mathbf{p}}) \end{aligned}$$

Explain that $J\boldsymbol{\xi}(\cdot)$ is the matrix constructed by stacking on top of each other the gradients of each of the demand functions. i.e.,

$$J\boldsymbol{\xi}(\bar{\mathbf{p}}) = \begin{bmatrix} \nabla \xi^1(\bar{\mathbf{p}})' \\ \vdots \\ \nabla \xi^n(\bar{\mathbf{p}})' \end{bmatrix}$$

To do a specific example, we are going to set $n = m = 2$. Start out with a given vector $\bar{\mathbf{p}}$, then move it to \mathbf{p} . We are interested in approximating the *difference* between the values of the nonlinear function $\boldsymbol{\xi}$, evaluated at these two vectors, i.e., $\boldsymbol{\xi}(\mathbf{p}) - \boldsymbol{\xi}(\bar{\mathbf{p}}) = (dp_1, dp_2)$. We have

$$\begin{aligned} d\boldsymbol{\xi} &= \begin{pmatrix} d\xi^1 \\ d\xi^2 \end{pmatrix} \\ &= \begin{bmatrix} \nabla \xi^1(\bar{\mathbf{p}})' \\ \nabla \xi^2(\bar{\mathbf{p}})' \end{bmatrix} \begin{pmatrix} dp_1 \\ dp_2 \end{pmatrix} \\ &= \begin{bmatrix} \xi_1^1(\bar{\mathbf{p}}) & \xi_2^1(\bar{\mathbf{p}}) \\ \xi_1^2(\bar{\mathbf{p}}) & \xi_2^2(\bar{\mathbf{p}}) \end{bmatrix} \begin{pmatrix} dp_1 \\ dp_2 \end{pmatrix} \\ &= \begin{pmatrix} \xi_1^1(\bar{\mathbf{p}})dp_1 + \xi_2^1(\bar{\mathbf{p}})dp_2 \\ \xi_1^2(\bar{\mathbf{p}})dp_1 + \xi_2^2(\bar{\mathbf{p}})dp_2 \end{pmatrix} \end{aligned}$$

Emphasize again that what's going on in all of these examples is that we are approximating the true effect of a change in some variable by the value of the differential, evaluated at the change, in this case a vector.

Do a concrete example with real numbers.

$$\boldsymbol{\xi}(\mathbf{p}) = \begin{pmatrix} y/2p_1 \\ y/2p_2 \end{pmatrix}$$

$$J\boldsymbol{\xi}(\cdot) = \begin{bmatrix} -y/2p_1^2 & 0 \\ 0 & -y/2p_2^2 \end{bmatrix}$$

Set $y = 8000$; $\bar{p}_1 = \bar{p}_2 = 4$; $p_1 = p_2 = 4.1$, so that $\boldsymbol{\xi}(\bar{\mathbf{p}}) = (1000, 1000)$; $\boldsymbol{\xi}(\mathbf{p}) = (975.6, 975.6)$;

Thus $\mathbf{p} - \bar{\mathbf{p}} = (0.1, 0.1)$ while $\boldsymbol{\xi}(\mathbf{p}) - \boldsymbol{\xi}(\bar{\mathbf{p}}) = (-24.4, -24.4)$.

Calculate the approximation:

$$\begin{aligned}
 d\xi(\cdot) &= \begin{bmatrix} -y/2p_1^2 & 0 \\ 0 & -y/2p_2^2 \end{bmatrix} \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix} \\
 &= \begin{bmatrix} -8000/32 & 0 \\ 0 & -8000/32 \end{bmatrix} \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix} \\
 &= \begin{bmatrix} -250 & 0 \\ 0 & -250 \end{bmatrix} \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix} \\
 &= (-25, -25)
 \end{aligned}$$

So the approximation is within about 2.5% of the right answer.

Graphically, what is going on here is very similar to what we did in the linear algebra section. That is, we are going to look at the image of \mathbf{dp} under the linear function defined by the Jacobian matrix. Fig. 5 shows the change in price in \mathbf{dp} space, the pair of gradient vectors, the image of \mathbf{dp} under the linear function defined by the Jacobian matrix, and finally the original demand vector together with the approximate location of the new demand vector.

- top left picture is a circle of dp 's. The horizontal axis is the first component of \mathbf{dp} , the vertical axis is the second.
- bottom left picture has the columns of the Jacobian: emphasize that the vectors are each gradient vectors for each demand function. The label on the horizontal axis is: derivative w.r.t. first price.
- bottom right is what happens to the dp 's under the function defined by the Jacobian matrix.
- Top right is in quantity space, and show that where the quantity ends up is roughly obtained by adding the image of dp to the starting vector.

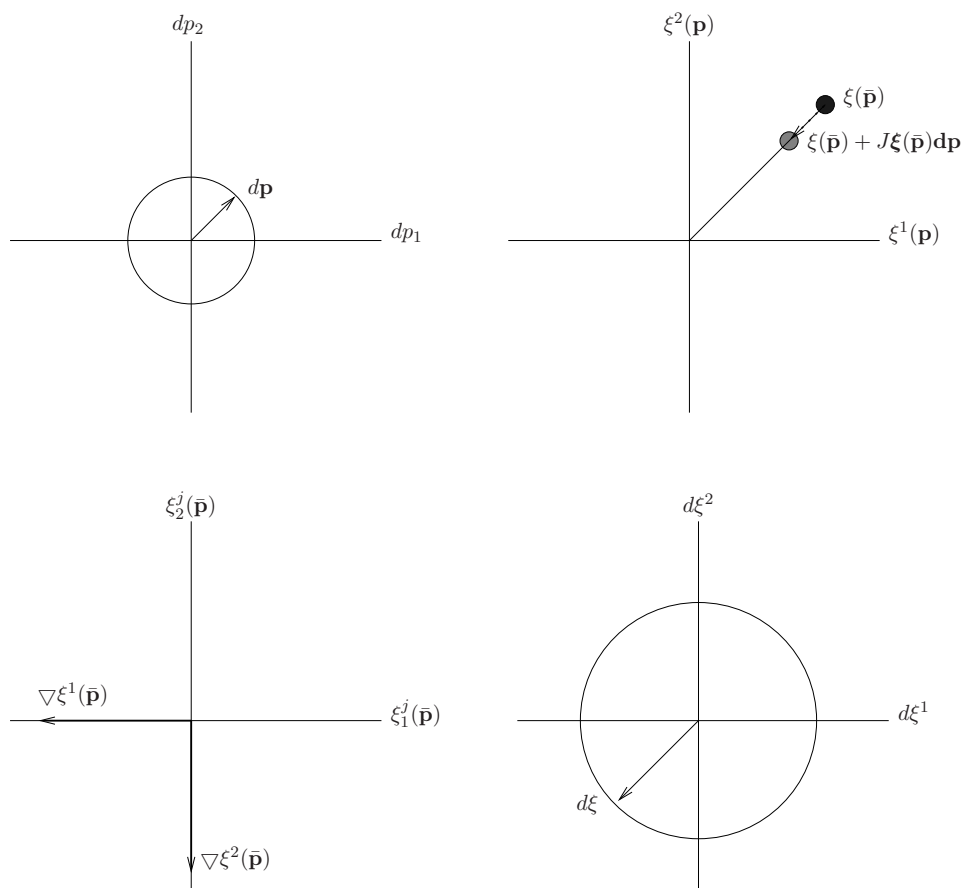


FIGURE 5. Demand as a function of price

4.6. Taylor's Theorem

Approximating the change in a nonlinear function by evaluating the differential is only a good approximation if the change is small. As we noted last time, we can improve our approximation by adding in extra terms; instead of doing a *linear* or *first-order* approximation, can do a *quadratic* or *second-order* approximation. Consider Fig. 6. The function $f(\cdot)$ is quite well approximated by the affine function $A(\cdot)$, but it is better approximated by the quadratic function $Q(\cdot)$. And would be

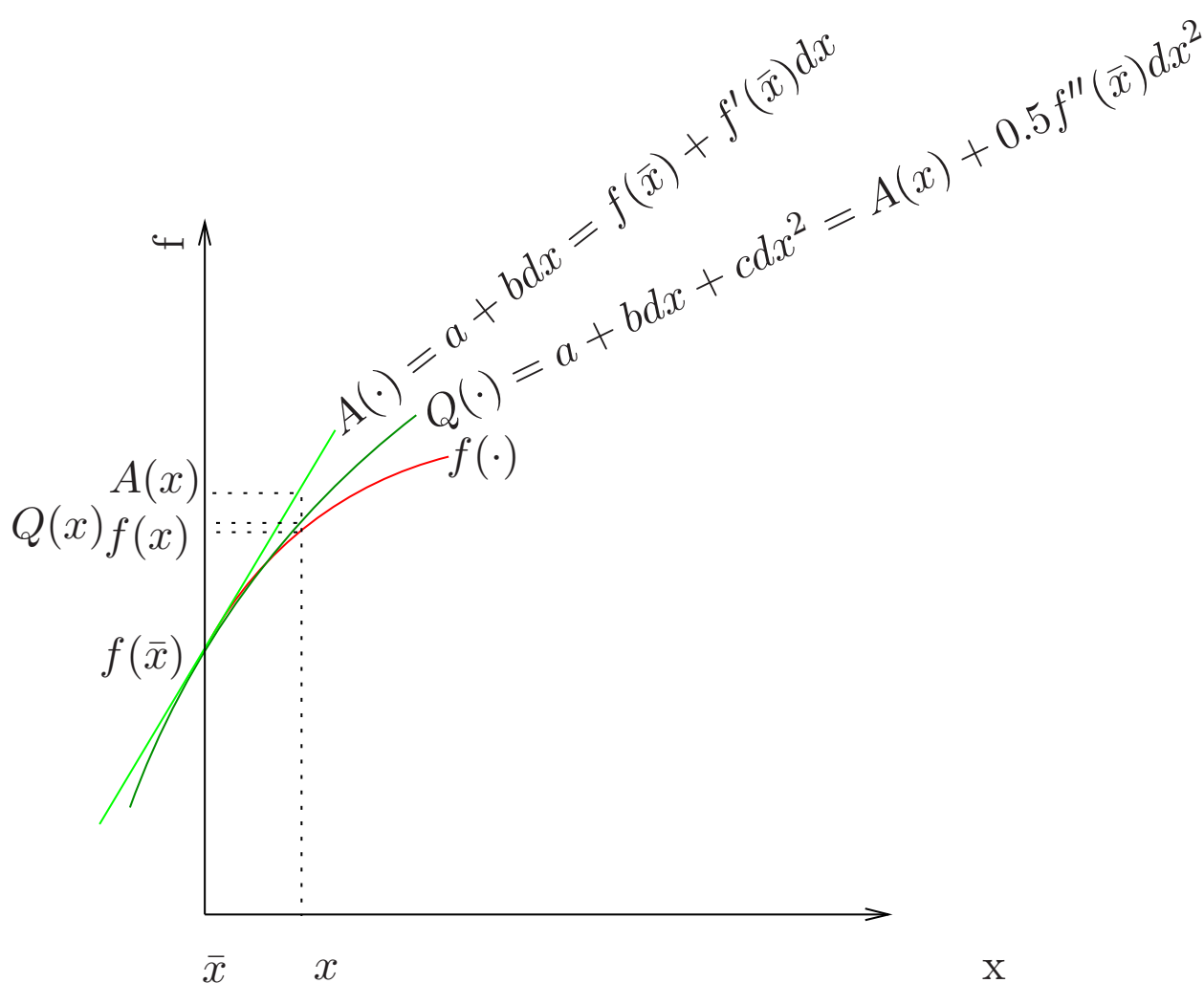


FIGURE 6. Approximating a function with linear and quadratic functions

better still approximated by a cubic function, etc. Indeed, if f were a k 'th order polynomial, then f would be *perfectly* approximated by a Taylor approximation that included k terms. Similarly, if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and is twice continuously differentiable, then

$$f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) \approx \nabla f(\bar{\mathbf{x}}) \cdot \mathbf{dx}$$

$$f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) \approx \nabla f(\bar{\mathbf{x}})\mathbf{dx} + \frac{1}{2}\mathbf{dx}'\mathbf{H}f(\bar{\mathbf{x}})\mathbf{dx}$$

even better

$$f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) = \nabla f(\bar{\mathbf{x}})\mathbf{dx} + \frac{1}{2}\mathbf{dx}'\mathbf{H}f(\bar{\mathbf{x}})\mathbf{dx} + \text{a remainder term}$$

We'll refer to the last line as a second order Taylor expansion of f about $\bar{\mathbf{x}}$ in the direction \mathbf{dx} . To write down a higher order expansion, we need hyper-matrix notation, which is a royal pain. I'm going to

cheat and, for each κ , simply define $\text{Tf}_\kappa(\bar{\mathbf{x}}, \mathbf{dx}) = \begin{cases} \nabla f(\bar{\mathbf{x}}) \cdot \mathbf{dx} & \text{if } \kappa = 1 \\ \mathbf{dx}' \cdot \text{Hf}(\bar{\mathbf{x}}) \cdot \mathbf{dx} & \text{if } \kappa = 2 \\ \text{the analogous hypermatrix term} & \text{if } \kappa > 2 \end{cases}$

Next, define the k 'th order Taylor expansion of f about $\bar{\mathbf{x}}$ in the direction \mathbf{dx} , to be the following weighted sum of the Tf_κ 's:

$$T^k(f, \bar{\mathbf{x}}, \mathbf{dx}) = \sum_{\kappa=1}^k \frac{\text{Tf}_\kappa(\bar{\mathbf{x}}, \mathbf{dx})}{\kappa!}$$

We now have the following “global” version of Taylor's theorem, known as the Taylor-Lagrange theorem.

Theorem (Taylor-Lagrange or Global Taylor): If f is $(K + 1)$ times continuously differentiable, then for any $0 \leq k \leq K$ and any $\bar{\mathbf{x}}, \mathbf{dx} \in \mathbb{R}^n$, there exists $\lambda \in [0, 1]$ such that

$$f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) = T^k(f, \bar{\mathbf{x}}, \mathbf{dx}) + \underbrace{\frac{\text{Tf}_{k+1}(\bar{\mathbf{x}} + \lambda \mathbf{dx}, \mathbf{dx})}{(k+1)!}}_{\text{the remainder term}} \quad (1)$$

Note that the remainder term differs from the other terms in the expansion because it is evaluated at some point on the line-segment between $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}} + \mathbf{dx}$. A priori, we have no idea of the value of λ . So how can this theorem be of any use to us? There are two cases in which it is very useful. Of these, the second is by far the more important.

- (1) if f is an $k + 1$ 'th order polynomial, then $\text{Tf}_{k+1}(\cdot, \mathbf{dx})$ is independent of its first argument, so that for the k 'th order Taylor expansion, the equality holds for *all* values of λ . E.g., let $k = 1$ and consider the quadratic $a + bx + cx^2$. In this case $\text{Tf}_{k+1}(\cdot, dx) = cdx^2/2$.
- (2) if the Hessian of f is a *definite* (or even semi-definite) matrix, then you always know the *sign* of $\text{Tf}_2(\cdot, dx)$, even if you don't know the value of λ . E.g., if f is a strictly concave, twice differentiable function, then $\text{Hf}(\cdot)$ is everywhere negative definite, i.e., $\text{Tf}_2(\cdot, \mathbf{dx}) = \mathbf{dx}' \text{Hf}(\cdot) \mathbf{dx}$ is *always* negative. Hence we know that $f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}})$ is *always* strictly less than the differential approximation $f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}}) \cdot \mathbf{dx}$.

Returning to functions where λ really does make a difference, you might think that if the $(k + 1)$ 'th derivative of f at $\bar{\mathbf{x}}$ were really huge, then the remainder term, which is determined by this term, would be really huge also, and thus mess up your approximation in the sense that the remainder term would be much larger in absolute value than the terms that have been written out explicitly. However, if an important caveat is satisfied, it turns out that *any* order of the Taylor expansion will be “good enough”—in the sense of determining the *sign* of the left hand side—provided that the length of \mathbf{dx} is small enough. The caveat is that for small enough \mathbf{dx} 's, the sum of the first k terms in the approximation (i.e., the k 'th order expansion, etc.) must be non-zero. For some k 's, in particular the important case of $k = 1$, whether or not this caveat is satisfied depends on the direction of \mathbf{dx} . Indeed, if the domain of f is \mathbb{R}^n , $n > 1$, it will *always* fail to be satisfied for *some* direction(s) \mathbf{dx} (since there always exists \mathbf{dx} such that $\nabla f(\mathbf{x}) \cdot \mathbf{dx} = 0$).

Theorem (Taylor-Young or Local Taylor):¹ Consider a $K + 1$ times continuously differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$, and any $0 \leq k \leq K$. Consider also a sequence (\mathbf{dx}^m) such that $\lim_m \|\mathbf{dx}^m\| \rightarrow 0$ but $\lim_m \frac{|T^k(f, \mathbf{x}, \mathbf{dx}^m)|}{\|\mathbf{dx}^m\|^k} > 0$. Then there exists $M \in \mathbb{N}$ such that for $m > M$, $|T^k(f, \mathbf{x}, \mathbf{dx}^m)|$ strictly exceeds the absolute value of the remainder term.

In applications of this theorem, we typically are interested in sequences of the form, $\mathbf{dx}^m = \mathbf{v}/m$, i.e., sequences in which the \mathbf{dx}^m 's all point in the same direction \mathbf{v} , but become increasingly short in length. For this special case, we have a simpler theorem:

A less general version of Taylor-Young's Theorem (Local Taylor): Consider a $K + 1$ times continuously differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$, and any $0 \leq k \leq K$. Fix $\mathbf{v} \in \mathbb{R}^n$ such that $Tf_k(\mathbf{x}, \mathbf{v}) \neq 0$. Then there exists $M \in \mathbb{N}$ such that for $m > M$, $|T^k(f, \mathbf{x}, \mathbf{v}/m)|$ strictly exceeds the absolute value of the remainder term.

The intuition for the theorem is clearest when f maps \mathbb{R} to \mathbb{R} and (a) the first thru $k - 1$ 'th order terms are zero; and (b) the k 'th order term, $\frac{f^{(k)} dx^k}{k!}$, is nonzero. By the Taylor Lagrange theorem,

¹ This version is slightly different from the one you see in most mathematics books. But this one is more useful for the kinds of applications we use.

we have in this case that for some $x' \in [x, x + dx]$,

$$\begin{aligned} f(x + dx) - f(x) &= \underbrace{\frac{f^{(k)}(\bar{x})dx^k}{k!}}_{k\text{'th order Taylor expansion}} + \underbrace{\frac{f^{(k+1)}(x')dx^{k+1}}{(k+1)!}}_{\text{Remainder term}} \\ &= \frac{dx^k}{k!} \left(f^{(k)}(\bar{x}) + dx \frac{f^{(k+1)}(x')}{(k+1)} \right) \end{aligned} \quad (2)$$

Consider $dx > 0$. If dx is sufficiently small, then the first term in parentheses is going to dominate the second term, and $(f(x + dx) - f(x))$ is going to have the same sign as $f^{(k)}(x)$.

Notice, however, that if condition (a) above is *not* satisfied, i.e., if there is some $0 < \kappa < k$ such that $f^{(\kappa)}(\bar{x}) \neq 0$, then display (2) is misleading, since

- (1) the M identified by the theorem above is *not*, in general, going to be the M such that for $dx = 1/M$, the term in parentheses in (2) is zero, so that for $dx < 1/M$, $|f^{(k)}(\bar{x})|$ dominates $|dx \frac{f^{(k+1)}(x')}{(k+1)}|$.
- (2) in general the M identified by the theorem will be larger
- (3) by the time the “real” M for the theorem is reached, the k 'th term in the expansion will in fact be dominated by the κ 'th term

To illustrate this point, suppose that $f(x) = 13x - 9x^2 + 2x^3$ and $k = 2$, and consider the k 'th order Taylor expansion around $x = 1$. In this case, since $f'''(\cdot)$ is independent of x , we have

$$\begin{aligned} f'(x) &= 13 - 18x + 6x^2 \\ f''(x) &= -18 + 12x \\ f'''(x) &= 12 \end{aligned}$$

so that, evaluating the Taylor expansion of f about 1, we get

$$T^2(f, 1, dx) = dx - 3dx^2$$

while since the remainder term is $2dx^3$

$$f(1 + dx) - f(1) = dx - 3dx^2 + 2dx^3$$

Now for $k = 2$.

- (1) the M that solves $|f^{(k)}(\bar{x})| = |f^{(2)}(1)| = 6 = \frac{f^{(k+1)}(x')}{(k+1)}/M = 4/M$ is of course $M = 2/3$ (not an integer, but the math works out).
- (2) but at $dx = 1/M = 3/2$, $dx - 3dx^2 = -21/4$; it is certainly *not* the case that for all $dx < 3/2$, $T^2(f, 1, dx)$ dominates the remainder in absolute value.
- (3) In particular, for $dx = 1/3$, $T^2(f, 1, dx) = 0$ while the remainder term is $2/9 > 0$
- (4) the threshold dx needs to be sufficiently small (approx 0.1577) that $1 - 3dx = 2dx^2$ before the condition of the theorem is satisfied, for all $dx' < dx$.

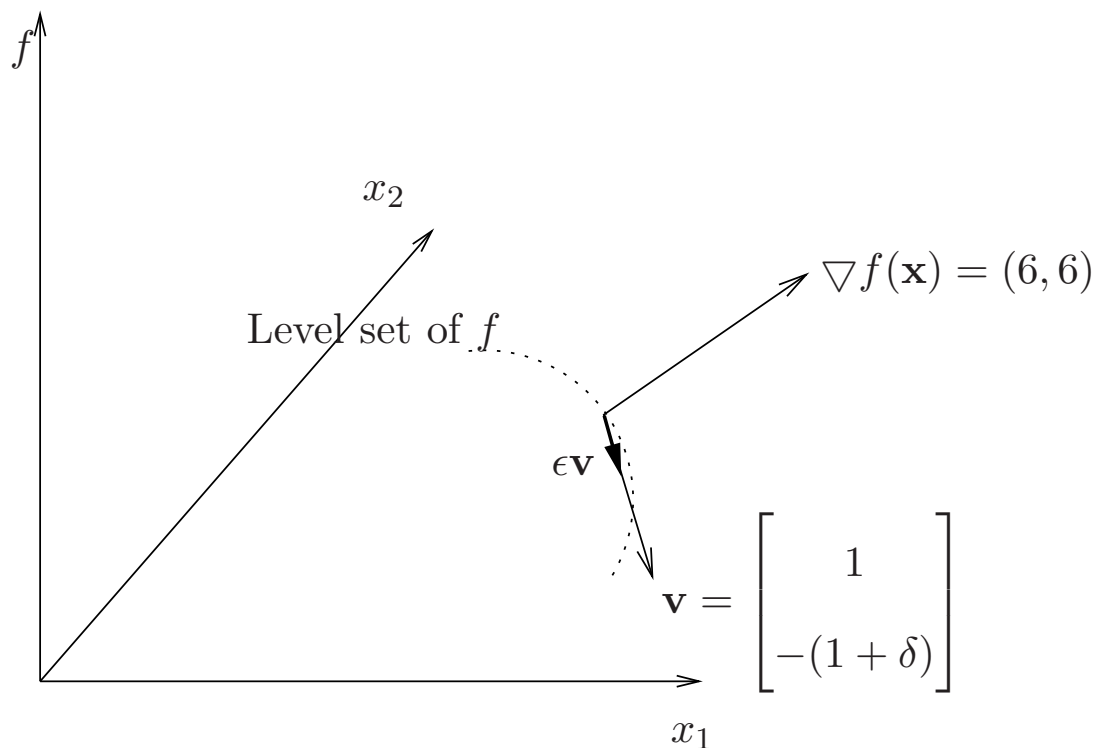
To summarize, the point of the condition “fix $\mathbf{v} \in \mathbb{R}^n$ such that $Tf^k(\mathbf{x}, \mathbf{v}) \neq 0$ ” is just that it ensures that *some* term among the first k terms is non-zero. If it happens, however, that some lower-order terms are non-zero as well and have signs that differ from the k 'th, then once M is large enough that the k 'th order expansion gives the sign for *every* $dx < 1/M$, the sign of the k 'th order term may well be different from the sign of the true difference.

In virtually all the applications we care about, k is either one or two. For example, set $k = 2$, and pick \mathbf{v} in the unit circle. If $\mathbf{v}' Hf(\mathbf{x}) \mathbf{v} \neq 0$ and m is sufficiently large then for $\mathbf{dx}^m = \mathbf{v}/m$, $|\nabla f(\bar{\mathbf{x}})\mathbf{dx}^m + \frac{1}{2}\mathbf{dx}^m'Hf(\bar{\mathbf{x}})\mathbf{dx}^m| > |\text{the remainder term}|$.

For some k 's, in particular the important case of $k = 1$, whether or not the caveat $T(\mathbf{x}, \mathbf{v}) \neq 0$ is satisfied depends on the direction of \mathbf{v} . Indeed, if the domain of f is \mathbb{R}^n , $n > 1$, it will *always* fail to be satisfied for *some* direction(s) \mathbf{v} (since there always exists some \mathbf{v} such that $\nabla f(\bar{\mathbf{x}}) \cdot \mathbf{v} = 0$).

To see the significance of this caveat, consider an unconstrained optimum of the function. In this case, the first order term in the Taylor series is necessarily zero, demonstrating that if you omitted the caveat the theorem would be false for $k = 1$. If $k > 1$, then the theorem goes thru even if the first $k - 1$ terms *are* zero, provided the k 'th term isn't.

Note that there is a difference between saying that the k 'th order *term* in the expansion is nonzero and that the k 'th order derivative is nonzero. Most obviously, the gradient could be nonzero, but the \mathbf{dx} could be orthogonal to the gradient. More generally, it follows that if we want to know when the first k terms in the Taylor expansion dominate the remainder, we must *first* fix the direction

FIGURE 7. 1st order approx “works” if $\epsilon \approx 0$

that the vector \mathbf{dx} points in, *then* take the length of the vector to zero: what we *can't in general* do is find an ϵ in advance that will work for all possible directions at once. More precisely, there will not exist in general an $\epsilon > 0$, such that for all \mathbf{dx} with norm less than ϵ , the absolute magnitude of the first k terms of the Taylor expansion will dominate the abs magnitude of the remainder term.

Illustration of Taylor's theorem for $k = 1$: The purpose of this example is to illustrate, that

- (1) provided the direction of movement \mathbf{dx} isn't orthogonal to the gradient, in which case the caveat of Taylor's theorem would fail for $k = 1$, then the sign of the linear approximation to the change in f will agree with the sign of the true change in f , provided that the magnitude of the shift \mathbf{dx} is sufficiently small.
- (2) there does *not* exist an $\epsilon > 0$ such that for *any* \mathbf{dx} with $\|\mathbf{dx}\| < \epsilon$, the sign of the linear approximation to the change in f will agree with the sign of the true change in f .

i.e., you have to first choose the direction and then the maximum length of the vector \mathbf{dx} . Suppose

that $f(x) = 3x_1^2 + 3x_2^2$, so that $\nabla f(x) = \begin{bmatrix} 6x_1 \\ 6x_2 \end{bmatrix}$ and

$\text{Hf}(x) = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}$. When $\bar{\mathbf{x}} = (1, 1)$, then

$$\begin{aligned} & f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) \\ &= \nabla f(\bar{\mathbf{x}})\mathbf{dx} + \frac{1}{2}\mathbf{dx}'\text{Hf}(\bar{\mathbf{x}})\mathbf{dx} \\ &= \begin{bmatrix} 6 & 6 \end{bmatrix}\mathbf{dx} + \frac{1}{2}\mathbf{dx}'\begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}\mathbf{dx} \\ &= 6(dx_1 + dx_2 + \frac{1}{2}(dx_1^2 + dx_2^2)) \end{aligned}$$

Notice that the entire Taylor expansion has exactly two terms, so that instead of an approximation sign in the display above, you have an equality. That is, when $k = 2$, there *is* no remainder term.

Next note that if $\mathbf{v} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, and $\mathbf{dx} = \epsilon\mathbf{v}$, for some $\epsilon \in \mathbb{R}$, then the *first* term in the Taylor expansion is zero, while the second is $6\epsilon^2$. Thus the first term in the Taylor expansion is dominated in absolute value by the second, regardless of the length of ϵ . Fortunately, however, this doesn't disprove Taylor's theorem, since in the direction \mathbf{v} , the first order term in the Taylor expansion is zero, so that when $k = 1$, the caveat in the theorem about the non-zerosness of the k 'th term is not satisfied.

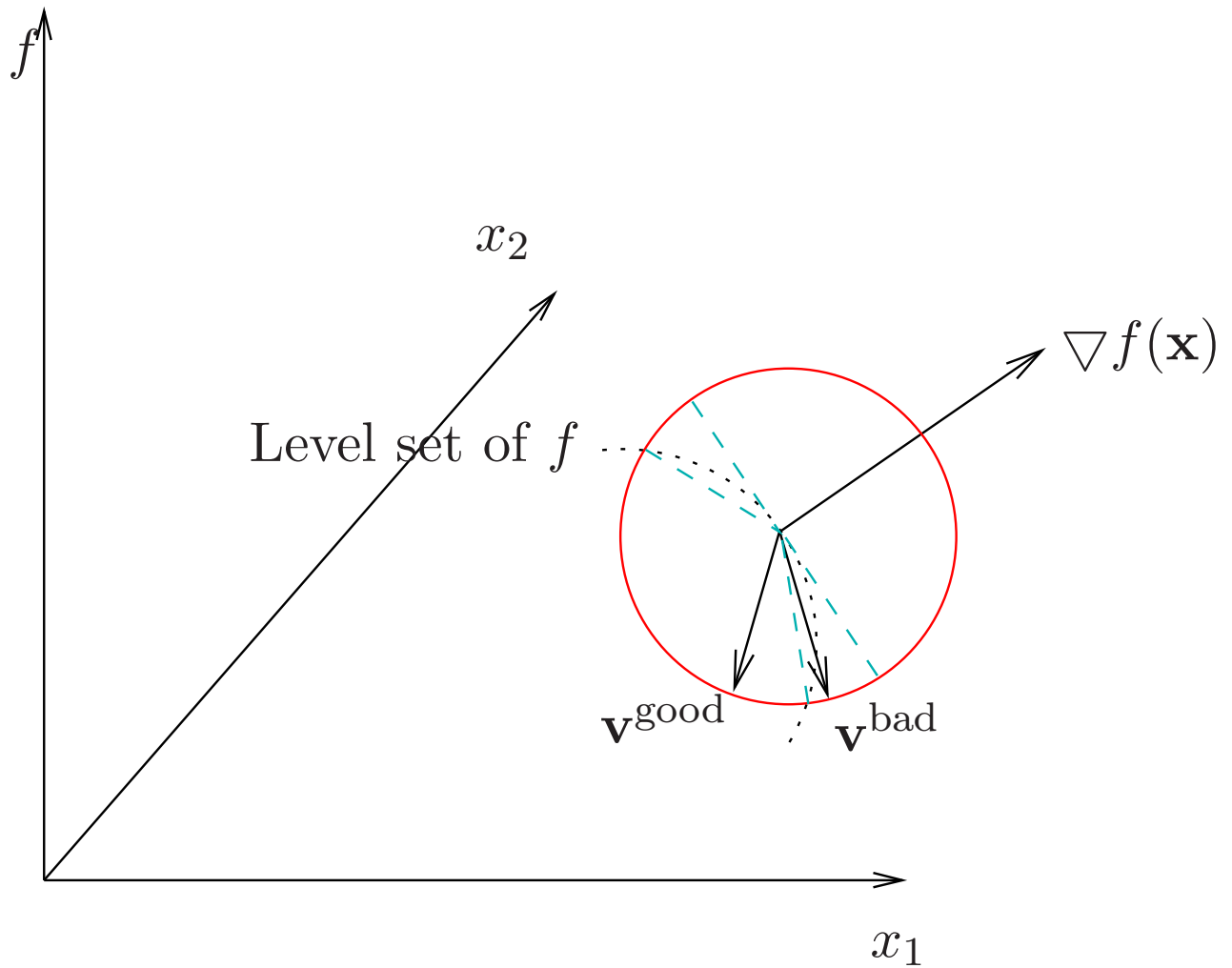
Now fix an arbitrary $\delta > 0$ and consider $\mathbf{v} = \begin{bmatrix} 1 \\ -(1 + \delta) \end{bmatrix}$. With this modification, the first term of the Taylor expansion in the direction \mathbf{v} is $-6\delta < 0$. Thus, the caveat in Taylor's theorem *is* satisfied for $k = 1$, and so the theorem had better work for this k . Indeed, we'll show that there exists $\bar{\epsilon} > 0$ such that if $\epsilon < \bar{\epsilon}$ and $\mathbf{dx} = \epsilon\mathbf{v}$, then $|\nabla f(\bar{\mathbf{x}})\mathbf{dx}| > |\frac{1}{2}\mathbf{dx}'\text{Hf}(\bar{\mathbf{x}})\mathbf{dx}|$, or, in other words, the sign of $f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}})$ will agree with the sign of $-6\epsilon\delta$.

Let $\mathbf{dx} = \epsilon \mathbf{v}$, for $\epsilon > 0$. Observe that the first term in the Taylor expansion is negative ($-6\delta\epsilon < 0$), while

$$\begin{aligned} f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}}) &= 6 \left(dx_1 + dx_2 + \frac{1}{2}(dx_1^2 + dx_2^2) \right) \\ &= 6 \left(-\epsilon\delta + \frac{1}{2}[\epsilon^2 + \epsilon^2(1 + \delta)^2] \right) \\ &= 6\epsilon \left(-\delta + \epsilon[\delta + 1 + \delta^2/2] \right) \\ &= 6\epsilon\delta \left(-1 + \underbrace{\epsilon[1 + 1/\delta + \delta/2]}_{\rightarrow \infty \text{ as } \delta \rightarrow 0} \right) \end{aligned}$$

Note that if $\epsilon > 0$ is, say greater than unity, then $f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}})$ is *positive*. On the other hand, provided that $\epsilon < \bar{\epsilon} = \frac{1}{1+1/\delta+\delta/2}$ then $f(\bar{\mathbf{x}} + \mathbf{dx}) - f(\bar{\mathbf{x}})$ will be negative, just like the first term in the Taylor expansion!

We now illustrate why it is impossible to pick an $\epsilon > 0$ such that the first order Taylor expansion will give the correct sign for *all* vectors of length not greater than ϵ . In Fig. 8, the circle centered at \mathbf{x} is of radius ϵ . Note that if you consider a vector of length ϵ that points into one of the two dashed cones emanating from \mathbf{dx} (for example \mathbf{v}^{bad}), then it will pass thru the lower contour set of f corresponding to \mathbf{x} and out the other side into the *upper* contour set. On the other hand, since \mathbf{v}^{bad} makes an obtuse angle with $\nabla f(\mathbf{x})$, the differential $\nabla f \mathbf{x} \mathbf{dx}$ is negative, i.e., gives the incorrect sign. For a vector such as \mathbf{v}^{good} , which lies outside the dashed cones, the sign of the differential is the same as the sign of the actual difference $f(\mathbf{x} + \mathbf{dx}) - f(\mathbf{x})$.

FIGURE 8. No $\epsilon > 0$ works for all directions