3. Univariate and Multivariate Differentiation (cont) 1

3.4. Multivariate Calculus: functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ 1

3.5. Four graphical examples. 3

3.6. Taylor’s Theorem 14

3. Univariate and Multivariate Differentiation (cont)

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

We’ll now generalize what we did last time to a function $f : \mathbb{R}^n \to \mathbb{R}^m$. 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 particular kind of function from \( \mathbb{R}^n \) to \( \mathbb{R}^n \) that we care about is the gradient function. Specifically, think of the gradient as being \( n \) functions from \( \mathbb{R}^n \to \mathbb{R} \), all stacked on top of each other. The gradient of the gradient is the matrix constructed by stacking the gradients of each of these functions \textit{viewed as row vectors} on top of each other. E.g., the first row will be the derivative of the first partial, i.e., \( \nabla f_1(\cdot) \). The derivative of the derivative of a function is called the \textit{Hessian} of that function. The Hessian of \( f \) is, of course, the Jacobian of the gradient of \( f \).

To visualize the derivative and differential associated with \( f : \mathbb{R}^n \to \mathbb{R}^m \), it is helpful to think, as usual, of \( f \) as a vertical stack of \( m \) functions \( f^i : \mathbb{R}^n \to \mathbb{R} \), all stacked on top of each other. It is then natural to think of the derivative of \( f \) as a vertical stack of all the derivatives of the \( f^i \)'s.

That is, \( f'(\cdot) \equiv Jf(\cdot) = \begin{bmatrix} \nabla f^1(\cdot)' \\ \nabla f^2(\cdot)' \\ \vdots \\ \nabla f^m(\cdot)' \end{bmatrix} \), where each \( \nabla f^i(\cdot) \) is a column vector consisting of the partials of \( f^i \). In the special case of \( \nabla f : \mathbb{R}^n \to \mathbb{R}^n \), we have \( \nabla f'(\cdot) \equiv J\nabla f(\cdot) \equiv \text{Hf}(\cdot) = \begin{bmatrix} \nabla f_1(\cdot)' \\ \nabla f_2(\cdot)' \\ \vdots \\ \nabla f_n(\cdot)' \end{bmatrix} \), where each \( \nabla f_i(\cdot) \) is the gradient of the \( i \)’th partial of \( f \).
Now, returning to a general function \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \), think of the differential of \( \nabla f \), i.e., \( df^x(\cdot) = Jf(x)(\cdot) \) as a vertical stack consisting of the differentials of the \( f^i \)'s at \( x \), i.e.,

\[
\begin{bmatrix}
\nabla f^1(x) \cdot dx \\
\nabla f^2(x) \cdot dx \\
\vdots \\
\nabla f^m(x) \cdot dx
\end{bmatrix}
\]

3.5. **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 $f : \mathbb{R}^2 \to \mathbb{R}$. 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 $dx$ in the domain.

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

$$f(x) = \left(\frac{x_1^2}{2} - \frac{x_1^3}{3}\right)\left(\frac{x_2^3}{3} - \frac{x_2^2}{2}\right)$$

whose gradient is

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

so that $d\nabla f(x)\cdot dx = Hf(x) \cdot dx$, where

$$\nabla f_1(x) = \begin{bmatrix} (1 - 2x_1)\left(\frac{x_2^3}{3} - \frac{x_2^2}{2}\right) \\ \left(x_1 - x_1^2\right)\left(x_2 - x_2\right) \end{bmatrix}$$

and

$$\nabla f_2(x) = \begin{bmatrix} \left(x_2 - x_2\right)\left(x_1 - x_1^2\right) \\ \left(2x_2 - 1\right)\left(x_1^2/2 - \frac{x_1^3}{3}\right) \end{bmatrix}$$

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

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig1.png}
\caption{Graph of $f$}
\end{figure}
Plugging in the numbers, we obtain

\[
\nabla f(x) = \begin{bmatrix} -0.0274 \\ -0.0274 \end{bmatrix}; \quad \nabla f(x + dx) = \begin{bmatrix} -0.0401 \\ -0.0075 \end{bmatrix}
\]

so that

\[
\nabla f(x + dx) - \nabla f(x) = \begin{bmatrix} -0.0127 \\ 0.0199 \end{bmatrix}
\]

i.e., the first partial becomes more negative while the second becomes less so. Evaluating the differential of \( \nabla f \) at \( x \) at the magnitude of the change we obtain

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

Note that when we evaluate the differential, the second component of the approximation is much closer to the second component of the true change in \( \nabla f \) than is the first element.
To see the 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

![Graph of $f_1$ and $f_2$](image)

![Differential of $f_1$ and $f_2$](image)

**Figure 2.** The differential approximation to a change in gradient
much easier to understand Fig. 2 if you 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:

- The *level* of $\nabla f(x)$ is indicated by pink lines;
- The *level* of $\nabla f(x + dx)$ 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 $f_2(x+dx)$ is almost perfect, while the first order linear approximation to $f_1(x + dx)$ is much less so. This is reflected in the bottom right panel, where there is a big gap between $(f_1(x + dx) - f_1(x))$ and $df_1(x)(dx)$ and a negligible one between $(f_2(x + dx) - f_2(x))$ and $df_2(x)(dx)$.

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

**Second example** (see Fig. 3): $f$ 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 $y = \alpha x$, for some scalar $\alpha \in \mathbb{R}_+$, then the slope of the level set of $f$ through $y$ is equal to the slope of the level set of $f$ through $x$. Since gradient vectors are perpendicular to level sets, this implies that the gradients of $f$ at both $x$ and $y$ must point in the same direction. Let’s check that this is true for $f(x) = x_1x_2$. 
\[ \nabla f(x) = \begin{bmatrix} x_2 & x_1 \end{bmatrix} \]

\[ \text{Hf}(x) = J \nabla f(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \]

so the differential of \( \nabla f \) at \( x \) is

\[
\begin{align*}
    d\nabla f(x)(dx) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} \\
    &= \begin{bmatrix} \lambda x_2 \\ \lambda x_1 \end{bmatrix}
\end{align*}
\]

In this case \( J \nabla f(x) \) 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 \( dx \) along the ray through the origin passing through \( x \), i.e., choose \( dx = \lambda x \), for some scalar \( \lambda > 0 \). In this case, we have

\[
\begin{align*}
    d\nabla f(x)(dx) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix} \\
    &= \begin{bmatrix} \lambda x_2 \\ \lambda x_1 \end{bmatrix}
\end{align*}
\]
so that, taking a first order approximation to $\nabla f(x + \lambda x)$:

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

But in this case, we can replace the approximation symbol with an equality. That is, the gradient of $f$ at $(1 + \lambda)x$ is a scalar multiple of the gradient of $f$ at $x$, 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.)

**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 $|f_1(x)|/|f_2(x)|$. 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 $x$ and consider a north-west movement in the domain, orthogonal to the gradient of $f$. Since, $\nabla f(x) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}$ an north-west movement orthogonal to this vector would be $dx = (-\lambda x_1, \lambda x_2)$. (Observe that $\nabla f(x) \cdot dx = -\lambda x_1 x_2 + \lambda x_1 x_2 = 0$, so that indeed $dx$ and $\nabla f(x)$
are orthogonal to each other). Now
\[
d\nabla f(x)(dx) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\lambda x_1 \\ \lambda x_2 \end{bmatrix} = \begin{bmatrix} \lambda x_2 \\ -\lambda x_1 \end{bmatrix}
\]
so that
\[
\nabla f(x + \lambda x) = \nabla f(x) + d\nabla f(x)(dx) = \begin{bmatrix} (1 + \lambda)x_2 \\ (1 - \lambda)x_1 \end{bmatrix}
\]
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: \(\xi(p) = \begin{bmatrix} \xi^1(p) \\ \vdots \\ \xi^n(p) \end{bmatrix}\). The Jacobian of this function is written as \(J\xi(\cdot)\). 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 \(\bar{p}\). Want to know the effect of a change in the price vector from \(\bar{p}\) to \(p\):
\[
\xi(p) - \xi(\bar{p}) \\
\approx d\xi \\
= J\xi(\bar{p})(p - \bar{p})
\]
Explain that $J\xi(\cdot)$ is the matrix constructed by stacking on top of each other the gradients of each of the demand functions. i.e.,

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

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

$$d\xi = \begin{pmatrix} d\xi^1 \\ d\xi^2 \end{pmatrix}$$

$$= \begin{bmatrix} \nabla \xi^1(\bar{p})' \\ \nabla \xi^2(\bar{p})' \end{bmatrix} \begin{pmatrix} dp_1 \\ dp_2 \end{pmatrix}$$

$$= \begin{bmatrix} \xi^1(\bar{p}) & \xi^1(\bar{p}) \\ \xi^2(\bar{p}) & \xi^2(\bar{p}) \end{bmatrix} \begin{pmatrix} dp_1 \\ dp_2 \end{pmatrix}$$

$$= \begin{pmatrix} \xi^1(\bar{p})dp_1 + \xi^1(\bar{p})dp_2 \\ \xi^2(\bar{p})dp_1 + \xi^2(\bar{p})dp_2 \end{pmatrix}$$

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.

\[
\xi(p) = \begin{pmatrix}
\frac{y}{2p_1} \\
\frac{y}{2p_2}
\end{pmatrix}
\]

\[
J\xi(\cdot) = \begin{bmatrix}
\frac{-y}{2p_1^2} & 0 \\
0 & \frac{-y}{2p_2^2}
\end{bmatrix}
\]

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

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

Calculate the approximation:

\[
d\xi(\cdot) = \begin{bmatrix}
\frac{-y}{2p_1^2} & 0 \\
0 & \frac{-y}{2p_2^2}
\end{bmatrix} 
\begin{pmatrix}
0.1 \\
0.1
\end{pmatrix}
\]

\[
= \begin{bmatrix}
\frac{-8000}{32} & 0 \\
0 & \frac{-8000}{32}
\end{bmatrix} 
\begin{pmatrix}
0.1 \\
0.1
\end{pmatrix}
\]

\[
= \begin{bmatrix}
-250 & 0 \\
0 & -250
\end{bmatrix} 
\begin{pmatrix}
0.1 \\
0.1
\end{pmatrix}
\]

\[
= (-25, -25)
\]

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 \(dp\) under the linear function defined by the Jacobian
matrix. Fig. 5 shows the change in price in $dp$ space, the pair of gradient vectors, the image of $dp$ 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 $dp$, the vertical axis is the second.
3.6. 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.

\[
f(\bar{x} + dx) - f(\bar{x}) \approx f'(\bar{x})dx
\]

\[
f(\bar{x} + dx) - f(\bar{x}) \approx f'(\bar{x})dx + \frac{1}{2} f''(\bar{x})dx^2
\]

even better

\[
f(\bar{x} + dx) - f(\bar{x}) = f'(\bar{x})dx + \frac{1}{2} f''(\bar{x})dx^2 + \text{a remainder term}
\]

where the “remainder term” consists of the third order derivative of \( f \), evaluated at at a point somewhere between \( \bar{x} \) and \( \bar{x} + dx \): the \( k' \)th term in the series is \( f^{(k)}(\bar{x})dx^k / k! \), where \( f^{(k)} \) denotes the \( k \)'th derivative of \( f \) (e.g., \( f^{(3)} = f''' \)) and \( k! \) denotes “\( n \)-factorial,” i.e., \( k! = k \times (k-1) \times (n-2) \times \ldots \times 2 \).
Similarly, if $f : \mathbb{R}^n \to \mathbb{R}$ and is twice continuously differentiable, then

\[
f(\bar{x} + dx) - f(\bar{x}) \approx \nabla f(\bar{x}) \cdot dx
\]

\[
f(\bar{x} + dx) - f(\bar{x}) \approx \nabla f(\bar{x}) dx + \frac{1}{2} dx' Hf(\bar{x}) dx
\]
even better

\[
f(\bar{x} + dx) - f(\bar{x}) = \nabla f(\bar{x}) dx + \frac{1}{2} dx' Hf(\bar{x}) dx + \text{a remainder term}
\]

We’ll refer to the last line as a second order Taylor expansion of $f$ about $\bar{x}$ in the direction $dx$. 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 $\kappa$, simply define $Tf_\kappa(x, dx) = \begin{cases} \nabla f(\bar{x}) \cdot dx & \text{if } \kappa = 1 \\ dx' \cdot Hf(\bar{x}) \cdot dx & \text{if } \kappa = 2 \\ \text{the analogous hypermatrix term} & \text{if } \kappa > 2 \end{cases}$.

Next, define the $k$’th order Taylor expansion of $f$ about $x$ in the direction $dx$, to be the following weighted sum of the $Tf_\kappa$’s:

\[
T^k(f, x, dx) = \sum_{\kappa=1}^{k} \frac{Tf_\kappa(\bar{x}, dx)}{\kappa!}
\]

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

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

\[
f(\bar{x} + dx) - f(\bar{x}) = T^k(f, x, dx) + \frac{Tf_{k+1}(x + \lambda dx, dx)}{(k + 1)!}
\]

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 $\bar{x}$ and $\bar{x} + dx$. A priori, we have no idea of the value of $\lambda$. For example, you should verify that if $f$ is an $K$’th order polynomial, then regardless of the
value of $x$ and $dx$, the remainder term for the $K-1$'th order Taylor expansion must be evaluated at $\lambda = 0$.

You might think that if the $(k+1)$'th derivative of $f$ at $\bar{x}$ 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 $dx$ is small enough. The caveat is that for small enough $dx$’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 $dx$. Indeed, if the domain of $f$ is $\mathbb{R}^n$, $n > 1$, it will always fail to be satisfied for some direction(s) $dx$ (since there always exists $dx$ such that $\nabla f(x) \cdot dx = 0$).

(A version of) Taylor-Young’s Theorem: Consider a $K + 1$ times continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$, and any $k \leq K$. Consider also a sequence $(dx^m)$ such that $\lim_m ||dx^m|| \to 0$ but $\lim_m \frac{|T^k(f, x, dx^m)|}{||dx^m||^k} > 0$. Then there exists $M \in \mathbb{N}$ such that for $m > M$, $|T^k(f, x, dx^m)|$ strictly exceeds the absolute value of the remainder term.

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

A less general version of Taylor-Young’s Theorem: Consider a $K + 1$ times continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$, and any $k \leq K$. Fix $v \in \mathbb{R}^n$ such that $Tf_k(x, v) \neq 0$. Then

---

\footnote{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.}
there exists $M \in \mathbb{N}$ such that for $m > M$, $|T^k(f, x, v/m)|$ 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, $f^{(k)}dx^k/k!$, is nonzero. By the Taylor Lagrange theorem, we have in this case that for some $x' \in [x, x + dx]$,

$$f(x + dx) - f(x) = \underbrace{\frac{f^{(k)}(\bar{x})dx^k}{k!}}_{k'th\ \text{order\ Taylor\ expansion}} + \underbrace{\frac{f^{(k+1)}(x')dx^{k+1}}{(k+1)!}}_{\text{Remainder term}}$$

$$= \frac{dx^k}{k!} \left( f^{(k)}(\bar{x}) + dx \frac{f^{(k+1)}(x')}{(k+1)} \right) \quad (2)$$

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 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 (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 $\kappa$'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

\[
\begin{align*}
    f'(x) &= 13 - 18x + 6x^2 \\
    f''(x) &= -18 + 12x \\
    f'''(x) &= 12
\end{align*}
\]

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 \( T f_k(\mathbf{x}, \mathbf{v}) \neq 0 \)” 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. Now consider a sequence $(\mathbf{d}x^m)$ defined by, for each $m$, $\mathbf{d}x^m = \mathbf{v}/m$, and assume that for all $m$ greater than some $M$, $\frac{|\nabla f(\bar{x})\mathbf{d}x^m + \frac{1}{2}\mathbf{d}x^m \mathbf{H}(\bar{x})\mathbf{d}x^m|}{(||\mathbf{v}||/m)^2} > 0$. (You should check that a sufficient condition for this property to be satisfied is that $\nabla f(\bar{x})\mathbf{v} \neq 0$. It is a little trickier to check this, but an alternative sufficient condition is that $\mathbf{H}(\bar{x})$ is a definite matrix.) In this case, if $m$ is sufficiently large then

$$|\nabla f(\bar{x})\mathbf{d}x^m + \frac{1}{2}\mathbf{d}x^m \mathbf{H}(\bar{x})\mathbf{d}x^m| > |\text{the remainder term}|.$$ 

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

To see the significance of the caveat, consider an unconstrained optimum of the function. In this case, the first order term in the series is 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 $\mathbf{d}x$ 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 that the vector $\mathbf{d}x$ points in, then take the length of the vector to zero: what we can’t in general do is find an $\epsilon$ 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 $\mathbf{d}x$ with norm less than $\epsilon$, absolute magnitude of the first $k$ terms of the Taylor expansion dominates the abs magnitude of the remainder term.
Illustration of Taylor’s theorem for $k = 1$: The purpose of this example is to illustrate numerically that 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.

Suppose that $f(x) = 3x_1^2 + 3x_2^2$, so that

$$\nabla f(x) = \begin{bmatrix} 6x_1 \\ 6x_2 \end{bmatrix} \quad \text{and} \quad H f(x) = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}.$$ When $x = (1, 1)$, then

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

$$= \begin{bmatrix} 6 & 6 \end{bmatrix} \mathbf{dx} + \frac{1}{2} \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} \mathbf{dx}' \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} \mathbf{dx}$$

$$= 6 (dx_1 + dx_2) + \frac{1}{2} (dx_1^2 + dx_2^2))$$

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 $v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, and $\mathbf{dx} = \epsilon v$, for some $\epsilon \in \mathbb{R}$, then the first term in the Taylor expansion

\[\text{Figure 6. 1st order approx “works” if } \epsilon \approx 0\]
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 $\epsilon$. 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-zeroness of the sum of the first $k$ terms 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 $d\mathbf{x} = \epsilon \mathbf{v}$, then $|\nabla f(\bar{x})d\mathbf{x}| > |\frac{1}{2}d\mathbf{x}'Hf(\bar{x})d\mathbf{x}|$, or, in other words, the sign of $f(\bar{x} + d\mathbf{x}) - f(\bar{x})$ will agree with the sign of $-6\epsilon\delta$.

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

$$
\begin{align*}
f(\bar{x} + d\mathbf{x}) - f(\bar{x}) &= 6 \left( dx_1 + dx_2 + \frac{1}{2}(dx_1^2 + dx_2^2) \right) \\
&= 6 \left( -\epsilon \delta + \frac{1}{2}[\epsilon^2 + \epsilon^2(1 + \delta)^2] \right) \\
&= 6\epsilon \left( -\delta + \epsilon[(1 + \delta) + \delta^2/2] \right)
\end{align*}
$$

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