EECS 16B Designing Information Devices and Systems II
Fall 2021 Discussion Worksheet Discussion 12B

The following notes are useful for this discussion:

1. Quadratic Approximation and Vector Differentiation

As shown in the previous discussion, a common way to approximate a non-linear high-dimensional functions is to perform linearization near a point. In the case of a two-dimensional function $f(x, y)$ with scalar output, the linear approximation of $f(x, y)$ at a point $(x_\star, y_\star)$ is given by

$$f(x,y) \approx f(x_\star, y_\star) + f_x(x_\star, y_\star)(x - x_\star) + f_y(x_\star, y_\star)(y - y_\star) \tag{1}$$

where as in the previous section,

$$f_x(x_\star, y_\star) = \left.\frac{\partial f(x,y)}{\partial x}\right|_{(x_\star, y_\star)} \quad \text{and} \quad f_y(x_\star, y_\star) = \left.\frac{\partial f(x,y)}{\partial y}\right|_{(x_\star, y_\star)}. \tag{2}$$

In vector form, this can be written as:

$$f(\vec{x}) \approx f(\vec{x}_\star) + \left[D_{\vec{x}} f|_{\vec{x}_\star} \right] (\vec{x} - \vec{x}_\star). \tag{3}$$

Recall from the previous discussion that $D_{\vec{x}} f$ is a row-vector filled with the partial derivatives $\frac{\partial f(\vec{x})}{\partial x_i}$:

$$D_{\vec{x}} f = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \cdots & \frac{\partial f(\vec{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_{x_1}(\vec{x}) & \cdots & f_{x_n}(\vec{x}) \end{bmatrix}. \tag{4}$$

Our goal is to extend this idea to a quadratic approximation. To do this, we need some notion of a second derivative. For this discussion, we will only be considering these types of functions from $\mathbb{R}^n \to \mathbb{R}$, since that is the typical form for a cost function used during optimization.

(a) **Given the function $f(x) = \mathrm{e}^{-2x}$, find the first and second derivatives, and write out its quadratic approximation at $x = x_\star$.** *Hint: Use Taylor's theorem.*

Solution:

$$f'(x) = -2\mathrm{e}^{-2x} \tag{5}$$
$$f''(x) = 4\mathrm{e}^{-2x}. \tag{6}$$

As shown in the previous discussion, a linear approximation of $f(x)$ around $x_\star$ is:

$$f(x) \approx f(x_\star) + f'(x_\star) \cdot (x - x_\star) \tag{7}$$
$$\approx \mathrm{e}^{-2x_\star} - 2\mathrm{e}^{-2x_\star}(x - x_\star) \tag{8}$$

To create a quadratic approximation, we can add the following term to the equation:

$$f(x) \approx f(x_\star) + f'(x_\star) \cdot (x - x_\star) + \frac{1}{2} f''(x_\star) \cdot (x - x_\star)^2 \tag{9}$$

$$\approx e^{-2x_\star} - 2e^{-2x_\star}(x - x_\star) + 2e^{-2x_\star}(x - x_\star)^2. \tag{10}$$

How did we get this? Recall that the Taylor expansion around point $x_\star$ of scalar function f is:

$$f(x) = \sum_{k=0}^{\infty} f^{(k)}(x_\star) \cdot \frac{(x - x_\star)^k}{k!} \tag{11}$$

$$= f(x_\star) + f'(x_\star) \cdot (x - x_\star) + \frac{1}{2} f''(x_\star) \cdot (x - x_\star)^2 + \cdots \tag{12}$$

Linearization takes the first two terms, which give a linear function; quadratic approximation just takes the extra quadratic term, giving a quadratic function.

(b) To write second partial derivatives compactly, we will introduce a new notation that builds off the notation f_x and f_y introduced previously. To compute f_{xy}, we first take the derivative in x, then in y:

$$f_{xy}(x_\star, y_\star) = \left.\frac{\partial f_x(x, y)}{\partial y}\right|_{(x_\star, y_\star)} = \left.\frac{\partial^2 f(x, y)}{\partial y \partial x}\right|_{(x_\star, y_\star)}. \tag{13}$$

Given the function $f(x, y) = x^2 y^2$, find all of the first and second partial derivatives.

Solution: We have

$$f_x(x, y) = \frac{\partial f(x, y)}{\partial x} = 2xy^2 \tag{14}$$

$$f_y(x, y) = \frac{\partial f(x, y)}{\partial y} = 2x^2 y \tag{15}$$

for the first partial derivatives. To find the second derivatives, we need to take partials with respect to x and y for each of the above first partial derivatives, giving us 4 different equations.

$$f_{xx}(x, y) = \frac{\partial f_x(x, y)}{\partial x} = \frac{\partial}{\partial x}\left[2xy^2\right] = 2y^2 \tag{16}$$

$$f_{xy}(x, y) = \frac{\partial f_x(x, y)}{\partial y} = \frac{\partial}{\partial y}\left[2xy^2\right] = 4xy \tag{17}$$

$$f_{yx}(x, y) = \frac{\partial f_y(x, y)}{\partial x} = \frac{\partial}{\partial x}\left[2x^2 y\right] = 4xy \tag{18}$$

$$f_{yy}(x, y) = \frac{\partial f_y(x, y)}{\partial y} = \frac{\partial}{\partial y}\left[2x^2 y\right] = 2x^2. \tag{19}$$

(c) To find the quadratic approximation of $f(x, y)$ near $(x_\star, y_\star)$, we plug in $f(x_\star + \Delta x, y_\star + \Delta y)$ and drop the terms that are higher order than quadratic:

$$f(x_\star + \Delta x, y_\star + \Delta y) = (x_\star + \Delta x)^2 (y_\star + \Delta y)^2 \tag{20}$$

$$= (x_\star^2 + 2x_\star \Delta x + (\Delta x)^2)(y_\star^2 + 2y_\star \Delta y + (\Delta y)^2) \tag{21}$$

$$\approx x_\star^2 y_\star^2 + 2x_\star y_\star^2 \Delta x + 2x_\star^2 y_\star \Delta y \tag{22}$$

$$+ y_\star^2 (\Delta x)^2 + 4x_\star y_\star (\Delta x)(\Delta y) + x_\star^2 (\Delta y)^2 \tag{23}$$

$$= f(x_\star, y_\star) + f_x(x_\star, y_\star)\Delta x + f_y(x_\star, y_\star)\Delta y \tag{24}$$

$$+ \frac{1}{2} f_{xx}(x_\star, y_\star)(\Delta x)^2 + \frac{1}{2} f_{yy}(x_\star, y_\star)(\Delta y)^2 \tag{25}$$

$$+ f_{xy}(x_\star, y_\star)(\Delta x)(\Delta y). \tag{26}$$

This is slightly different from the expression we get via the Taylor series expansion. **How would we rewrite this expression, so that *all* second derivatives are involved, each with a coefficient of $\frac{1}{2}$?**

Solution: We note that

$$f_{xy}(x, y) = f_{yx}(x, y) \tag{27}$$

so we can write

$$f_{xy}(x, y) = \frac{1}{2} f_{xy}(x, y) + \frac{1}{2} f_{yx}(x, y) \tag{28}$$

and plug this into the previous set of equations to get

$$\begin{aligned} f(x + \Delta x, y + \Delta y) &= f(x_\star, y_\star) + f_x(x_\star, y_\star)\Delta x + f_y(x_\star, y_\star)\Delta y & (29)\\ &+ \frac{1}{2} f_{xx}(x_\star, y_\star)(\Delta x)^2 + \frac{1}{2} f_{yy}(x_\star, y_\star)(\Delta y)^2 & (30)\\ &+ \frac{1}{2} f_{yx}(x_\star, y_\star)(\Delta y)(\Delta x) + \frac{1}{2} f_{xy}(x_\star, y_\star)(\Delta x)(\Delta y). & (31) \end{aligned}$$

As expected from the scalar case, there is a component of the second partial derivative along the x direction, multiplied by $(\Delta x)^2$, and similarly for the y direction. There are also two different cross terms, both multiplied by $(\Delta x)(\Delta y)$.

(d) Just as we created the derivative row vector to hold all the first partial derivatives to help in writing linearization in matrix/vector form:

$$D_{\vec{x}} f = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \cdots & \frac{\partial f(\vec{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_{x_1}(\vec{x}) & \cdots & f_{x_n}(\vec{x}) \end{bmatrix} \tag{32}$$

we would like to create a matrix to hold all the second partial derivatives to help in writing quadratic approximation in matrix/vector form:

$$H_{\vec{x}} f = \begin{bmatrix} \frac{\partial^2 f(\vec{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\vec{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} f_{x_1 x_1}(\vec{x}) & \cdots & f_{x_n x_1}(\vec{x}) \\ \vdots & \ddots & \vdots \\ f_{x_1 x_n}(\vec{x}) & \cdots & f_{x_n x_n}(\vec{x}) \end{bmatrix} \tag{33}$$

This matrix is the *Hessian* of f. Note that this quantity is different from the *Jacobian* matrix that was covered in the previous discussion. In contrast to the Hessian, which is the matrix of second partial derivatives of a *scalar-valued vector-input* function $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian is the matrix of first partial derivatives of a *vector-valued vector-input* function $\vec{f}\colon \mathbb{R}^n \to \mathbb{R}^k$.

In fact, the Hessian is the (Jacobian) derivative of the derivative; if we let $\vec{g}(\vec{x}) = (D_{\vec{x}} f)^\top$ (so that it's a column vector and the dimensions work out), then $H_{\vec{x}} f = D_{\vec{x}} \vec{g}$. **To get a feel for the Hessian of f, find $H_{(x,y)} f$ for the f above, that is, $f(x, y) = x^2 y^2$.**

Solution: We know f is a function in two variables, so

$$\begin{aligned} H_{(x,y)} f &= \begin{bmatrix} f_{xx}(x, y) & f_{yx}(x, y) \\ f_{xy}(x, y) & f_{yy}(x, y) \end{bmatrix} & (34)\\ &= \begin{bmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{bmatrix}. & (35) \end{aligned}$$

(e) **Using the Hessian, write out the general formula for the quadratic approximation of a scalar-valued function f of a vector $\vec{x}$ in vector/matrix form.**

Solution: Before, in our linearization, we saw that the only way to get a scalar from the first derivative $D_{\vec{x}} f|_{\vec{x}_\star}$ and the increment $\Delta\vec{x}$ is to multiply them:

$$f(\vec{x}_\star + \Delta\vec{x}) \approx f(\vec{x}_\star) + \left[D_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}). \tag{36}$$

Now we are tasked to make a scalar from the second derivative $H_{\vec{x}} f|_{\vec{x}_\star}$ and $\Delta\vec{x}$. Remember that a scalar can be obtained by a row vector times a column vector, and that a matrix times a column vector is a column vector. So the following gives us a column vector:

$$\left[H_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}) \tag{37}$$

and the following gives us a scalar:

$$(\Delta\vec{x})^\top \left[H_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}). \tag{38}$$

Finally, remember that the Taylor series gives us a factor of $\frac{1}{2}$ on the second derivative, so the full quadratic approximation is

$$f(\vec{x}_\star + \Delta\vec{x}) \approx f(\vec{x}_\star) + \left[D_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}) + \frac{1}{2}(\Delta\vec{x})^\top \left[H_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}). \tag{39}$$

If we would like to verify this calculation, note that the solution to the previous problem generalizes to the following quadratic approximation:

$$f(\vec{x}_\star + \Delta\vec{x}) \approx f(\vec{x}_\star) + \sum_{i=1}^{n} f_{x_i}(\vec{x}_\star)\Delta x_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{x_i x_j}(\vec{x}_\star)(\Delta x_i)(\Delta x_j). \tag{40}$$

So we should expect to get the same thing by expanding our matrix-vector form. Notice that

$$\left[D_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}) = \sum_{i=1}^{n} \left[D_{\vec{x}} f|_{\vec{x}_\star}\right]_i [\Delta\vec{x}]_i = \sum_{i=1}^{n} f_{x_i}(\vec{x}_\star)\Delta x_i \tag{41}$$

and

$$(\Delta\vec{x})^\top \left[H_{\vec{x}} f|_{\vec{x}_\star}\right](\Delta\vec{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \left[H_{\vec{x}} f|_{\vec{x}_\star}\right]_{ij} [\Delta\vec{x}]_i [\Delta\vec{x}]_j = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{x_i x_j}(\vec{x}_\star)(\Delta x_i)(\Delta x_j). \tag{42}$$

So we get the exact same thing in both ways, and our vector/matrix formulation works.

(f) **[Practice]: Show that the quadratic approximation for the scalar-valued function $f(\vec{w}) = \mathrm{e}^{\vec{x}^\top \vec{w}}$ around $\vec{w} = \vec{w}_\star$ is**

$$f(\vec{w}_\star + \Delta\vec{w}) \approx \mathrm{e}^{\vec{x}^\top \vec{w}_\star}\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}\left(\vec{x}^\top(\Delta\vec{w})\right)^2\right). \tag{43}$$

assuming that $\vec{x}$ is just some constant, given vector.

Hint: You can compute the following partial derivatives:

$$f_{w_i}(\vec{w}) = x_i f(\vec{w}) \tag{44}$$

$$f_{w_i w_j}(\vec{w}) = x_i x_j f(\vec{w}). \tag{45}$$

Now compute $D_{\vec{w}}f$ and $H_{\vec{w}}f$, and plug it into the quadratic approximation formula.

Solution: To begin with, here's how we compute the first and second partial derivatives.

$$\begin{aligned}
f_{w_i}(\vec{w}) &= \frac{\partial}{\partial w_i} \mathrm{e}^{\vec{x}^\top \vec{w}} && (46)\\
&= \frac{\partial}{\partial w_i} \mathrm{e}^{\sum_{i=1}^n x_i w_i} && (47)\\
&= x_i \mathrm{e}^{\sum_{i=1}^n x_i w_i} && (48)\\
&= x_i \mathrm{e}^{\vec{x}^\top \vec{w}} && (49)\\
&= x_i f(\vec{w}). && (50)\\
f_{w_i w_j}(\vec{w}) &= \frac{\partial f_{w_i}(\vec{w})}{\partial w_j} && (51)\\
&= \frac{\partial}{\partial w_j}(x_i f(\vec{w})) && (52)\\
&= x_i \left(\frac{\partial}{\partial w_j} f(\vec{w}) \right) && (53)\\
&= x_i x_j f(\vec{w}) && (54)
\end{aligned}$$

Now we would like to compute $D_{\vec{w}}f$ and $H_{\vec{w}}f$.

$$\begin{aligned}
D_{\vec{w}}f &= \begin{bmatrix} f_{w_1}(\vec{w}) & \cdots & f_{w_n}(\vec{w}) \end{bmatrix} && (55)\\
&= \begin{bmatrix} x_1 f(\vec{w}) & \cdots & x_n f(\vec{w}) \end{bmatrix} && (56)\\
&= f(\vec{w}) \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} && (57)\\
&= f(\vec{w}) \vec{x}^\top && (58)\\
H_{\vec{w}}f &= \begin{bmatrix} f_{w_1 w_1}(\vec{w}) & \cdots & f_{w_1 w_n}(\vec{w}) \\ \vdots & \ddots & \vdots \\ f_{w_n w_1}(\vec{w}) & \cdots & f_{w_n w_n}(\vec{w}) \end{bmatrix} && (59)\\
&= \begin{bmatrix} x_1^2 f(\vec{w}) & \cdots & x_1 x_n f(\vec{w}) \\ \vdots & \ddots & \vdots \\ x_n x_1 f(\vec{w}) & \cdots & x_n^2 f(\vec{w}) \end{bmatrix} && (60)\\
&= f(\vec{w}) \begin{bmatrix} x_1^2 & \cdots & x_1 x_n \\ \vdots & \ddots & \vdots \\ x_n x_1 & \cdots & x_n^2 \end{bmatrix} && (61)\\
&= f(\vec{w}) \vec{x} \vec{x}^\top. && (62)
\end{aligned}$$

Using these expressions, the quadratic approximation of f at $\vec{w}$ is

$$f(\vec{w}_\star + \Delta\vec{w}) \approx f(\vec{w}_\star) + \left[D_{\vec{w}} f|_{\vec{w}_\star}\right](\Delta\vec{w}) + \frac{1}{2}(\Delta\vec{w})^\top \left[H_{\vec{w}} f|_{\vec{w}_\star}\right](\Delta\vec{w}) \tag{63}$$

$$= f(\vec{w}_\star) + f(\vec{w}_\star)\vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}(\Delta\vec{w})^\top \left(f(\vec{w}_\star)\vec{x}\vec{x}^\top\right)(\Delta\vec{w}) \tag{64}$$

$$= f(\vec{w}_\star)\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}(\Delta\vec{w})^\top \vec{x}\vec{x}^\top(\Delta\vec{w})\right) \tag{65}$$

$$= \mathrm{e}^{\vec{x}^\top \vec{w}_\star}\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}\left((\Delta\vec{w})^\top \vec{x}\right)\left(\vec{x}^\top(\Delta\vec{w})\right)\right) \tag{66}$$

$$= \mathrm{e}^{\vec{x}^\top \vec{w}_\star}\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}\left(\vec{x}^\top(\Delta\vec{w})\right)^\top\left(\vec{x}^\top(\Delta\vec{w})\right)\right) \tag{67}$$

$$= \mathrm{e}^{\vec{x}^\top \vec{w}_\star}\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}\left(\vec{x}^\top(\Delta\vec{w})\right)^2\right). \tag{68}$$

(g) Using the result in the previous subpart, **use linearity to give the quadratic approximation for the function $\sum_{i=1}^m \mathrm{e}^{\vec{x}_i^\top \vec{w}}$ around $\vec{w} = \vec{w}_\star$. Here, assume that the $\vec{x}_i$ are just some given vectors.**

Solution: By linearity, we can take the quadratic approximations from each of the components and add them together to form the approximation for the above summation:

$$f(\vec{w}_\star + \Delta\vec{w}) \approx \sum_{i=1}^m \mathrm{e}^{\vec{x}_i^\top \vec{w}_\star}\left(1 + \vec{x}_i^\top(\Delta\vec{w}) + \frac{1}{2}\left(\vec{x}_i^\top(\Delta\vec{w})\right)^2\right) \tag{69}$$

(h) **[Practice]:** The second derivative also has an interpretation as the derivative of the derivative. However, we saw that the derivative of a scalar-valued function with respect to a vector is naturally a row. **If you wanted to approximate how much the derivative changed by moving a small amount $\Delta\vec{w}$, how would you get such an estimate using your expression for the second derivative?**

Solution: We can view this problem as if we were trying to get a linear approximation of the first derivative using the second derivative. From the previous problems with linearization, we would like to take the derivative of the first derivative row vector, evaluate it at $\vec{w} = \vec{w}_\star$ and have it act upon the incremental vector $\Delta\vec{w}$ to get the resulting change in the first derivative. In the scalar case, this looks like

$$f'(w_\star + \Delta w) = f'(w_\star) + f''(w_\star)\Delta w \tag{70}$$

But in the vector case, since the first derivative is a row vector, we need to produce a row vector from the second derivative term.

There is no way to get a matrix that multiplies a column vector and returns a row vector. Matrix multiplication just doesn't work that way based on the definitions that we have made so far from 16A. We have a choice — we can define a new kind of operator that does this, or we can figure out a way to shoehorn this into standard matrix multiplication. We choose the latter. To get a row out of a matrix using the increment $\Delta\vec{w}$ we just compute $(\Delta\vec{w})^\top\left[H_{\vec{w}} f|_{\vec{w}_\star}\right]$. Notice that in this row times a matrix, the resulting i-th element in the row indeed has the change in the i-th coordinate from all j possible partial second derivatives:

$$\left[(\Delta\vec{w})^\top\left[H_{\vec{w}} f|_{\vec{w}_\star}\right]\right]_i = \sum_{j=1}^n \left[H_{\vec{w}} f|_{\vec{w}_\star}\right]_{ji} \Delta w_j \tag{71}$$

$$= \sum_{j=1}^{n} f_{w_j w_i}(\vec{w}) \Delta w_j \tag{72}$$

$$= \left[D_{\vec{w}}(f_{w_i})\big|_{\vec{w}_\star} \right] (\Delta \vec{w}) \tag{73}$$

So

$$D_{\vec{w}} f|_{\vec{w}_\star + \Delta \vec{w}} \approx D_{\vec{w}} f|_{\vec{w}_\star} + (\Delta \vec{w})^\top \left[H_{\vec{w}} f|_{\vec{w}_\star} \right]. \tag{74}$$

Contributors:

- Neelesh Ramachandran.
- Druv Pai.
- Pavan Bhargava.
- Anant Sahai.