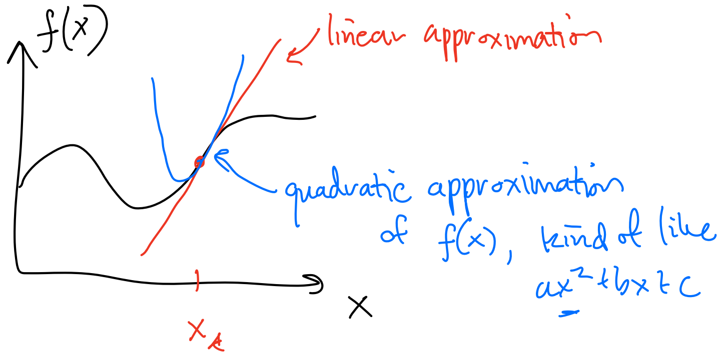
EECS16B DIS12B remote

Learning Objectives

① How to write second order/second derivative/quadratic approximations for functions

↳ Substkill: Writing/computing the Hessian matrix



- Energy functions in physics often have or deal with x^2
- Least squares: $\|A\vec{x}-\vec{b}\|^2$ ← function we want to minimize

$$f(x) \approx f(x_A) + \left.\frac{df}{dx}\right|_{x_A}(x-x_A) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_A}(x-x_A)^2$$

second degree/quadratic taylor series expansion for $f(x)$

$f(\vec{x})$

vector input, scalar output

$x \in \mathbb{R}^n$ $\quad$ $f(\vec{x}) \in \mathbb{R}$

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

The following notes are useful for this discussion: Note 19

1. Quadratic Approximation and Vector Differentiation

As shown in the previous discussion, a common way to locally 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

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

In vector form, this can be written as:

derivative row vector

$$f(\vec{x}) \approx f(\vec{x}_\star) + \left[D_{\vec{x}} f|_{\vec{x}_\star} \right] (\vec{x} - \vec{x}_\star). \tag{2}$$

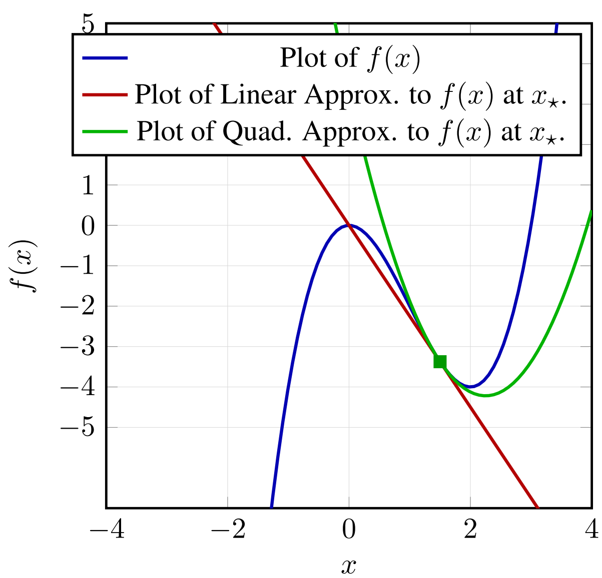
column vector

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

$$D_{\vec{x}} f = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \cdots & \frac{\partial f(\vec{x})}{\partial x_n} \end{bmatrix}. \tag{3}$$

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 functions from $\mathbb{R}^n \to \mathbb{R}$, since that is the typical form for a cost function used during optimization.

Why could a quadratic approximation be useful? Consider the plot from last week's discussion, below, which has been annotated with linear and quadratic approximations (the second of which we will learn to calculate). The quadratic approximation tracks the original curve more closely, reducing the error at larger values of δ away from $x_\star$.



$$f(x) = f(x_\star) + \frac{\partial f}{\partial x}\Big|_{x_\star}(\quad) + \cdots$$

(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 a Taylor series expansion.* ← (involves many derivative)

$$f(x) \approx \underbrace{f(x_\star)}_{} + \underbrace{\frac{df}{\partial x}\Big|_{x_\star}}_{}(x - x_\star) + \frac{1}{2}\underbrace{\frac{\partial^2 f}{\partial x^2}\Big|_{x_\star}}_{}(x - x_\star)^2$$

$$\boxed{e^{-2x_\star} - 2e^{-2x_\star}(x - x_\star) + \frac{1}{2}\cdot 4e^{-2x_\star}(x - x_\star)^2}$$

$$\frac{\partial f}{\partial x} = -2e^{-2x} \qquad \frac{\partial^2 f}{\partial x^2} = 4e^{-2x}$$

$\frac{d}{dx}$ vs $\frac{\partial}{\partial x}$
no difference in scalar case

(b) Given a multivariable function, we can take second partial derivatives. The first derivative can take the partial derivative with respect to x or y. For each of these first derivatives, we can again take a partial derivative on x or y, yielding 4 total second partial derivatives. For example, the notation below indicates taking a partial on x, then a partial on y (for multiple partials, we read/apply the denominator right-to-left, or inside-to-outside by convention).

$$f(x,y) = f(\vec{x}) \qquad \vec{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x} \leftarrow \tag{4}$$

Given the function $f(x,y) = x^2y^2$, find all of the first and second partial derivatives. Specifically, these are $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial x \partial x}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$, $\frac{\partial^2 f}{\partial y \partial y}$. ← involved in quadratic approximation

$f = x^2y^2$

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2y^2) = y^2\frac{\partial}{\partial x}(x^2) = \boxed{2xy^2} \qquad \frac{\partial^2}{\partial x \partial y}f = \boxed{4xy} \qquad \frac{\partial^2}{\partial x^2}f = \frac{\partial}{\partial x}(2xy^2) = \boxed{2y^2}$$

$$\frac{\partial f}{\partial y} = \boxed{2yx^2} \qquad \frac{\partial^3}{\partial y \partial x}f = \boxed{4xy} \qquad \frac{\partial^2}{\partial y^2}f = \boxed{2x^2}$$

(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. Notice that Δ here means the same thing as δ did in previous discussions (this can change purely depending on preference.)

$x_\star \Delta x + x_\star \Delta x$ $\qquad$ $y_\star \Delta y + y_\star \Delta y$

$$f(x_\star + \Delta x, y_\star + \Delta y) = (x_\star + \Delta x)^2(y_\star + \Delta y)^2 \tag{5}$$
$$= (x_\star^2 + 2x_\star\Delta x + (\Delta x)^2)(y_\star^2 + 2y_\star\Delta y + (\Delta y)^2) \tag{6}$$
$$\approx x_\star^2y_\star^2 + 2x_\star y_\star^2\Delta x + 2x_\star^2 y_\star \Delta y \tag{7}$$
$$+ y_\star^2(\Delta x)^2 + \boxed{4x_\star y_\star(\Delta x)(\Delta y)} + x_\star^2(\Delta y)^2 \tag{8}$$
$$= f(x_\star, y_\star) + \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y \tag{9}$$
$$+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(\Delta x)^2 + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(\Delta y)^2 \tag{10}$$
$$+ \boxed{\frac{\partial^2 f}{\partial x \partial y}(\Delta x)(\Delta y)} \tag{11}$$

can we put $\frac{\partial^2 f}{\partial y \partial x}$ in expression in lines ~~10~~ 9 to 11

$2x_\star\Delta x \cdot 2y_\star \Delta y$

This is slightly different from the expression we get via the Taylor series expansion. **How would we rewrite this expression, so that *all* 4 second derivatives are involved, each with a coefficient of $\frac{1}{2}$?** *Hint: what was the relationship we found between $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$?*

$$\frac{\partial^2 f}{\partial x \partial y}(\Delta x)(\Delta y) = \underbrace{\frac{1}{2}\frac{\partial^2 f}{\partial x \partial y}\Delta x \Delta y + \frac{1}{2}\frac{\partial^2 f}{\partial y \partial x}\Delta x \Delta y}$$

For a multivariable $f(\vec{x})$, we want to involve all second order derivatives in our quadratic approximation

(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:

$$D_{\vec{x}} f = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \cdots & \frac{\partial f(\vec{x})}{\partial x_n} \end{bmatrix} \tag{12}$$

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

Is $(H_{\vec{x}} f)\vec{x}$ a scalar?

Second derivative involves all different combinations of variables when taking derivatives twice

$$H_{\vec{x}} f = \begin{bmatrix} \frac{\partial^2 f(\vec{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\vec{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial x_n^2} \end{bmatrix} \tag{13}$$

$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

$f(x) \rightarrow \frac{\partial^2}{\partial x^2} f(x)$

scalar

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 same f as above.** $f(x,y) = x^2y^2$. **[Practice]:** For additional practice, try computing the hessian of $f_2(x,y,z) = x^2y^2z + x^3z^2 + y$.

$f = x^2y^2$ $\quad x_1 = x, x_2 = y$

$$H_{(x,y)} f = \underbrace{\begin{array}{c|c} & \\ x & \\ y & \end{array}\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}}_{2\times 2} = \begin{bmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{bmatrix}$$

2 variables x, y

Hessian matrix is always $n \times n$ where we have n variables

(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.** That is, we want to have an expression of the following form:

column

$$f(\vec{x}_\star + \Delta\vec{x}) \approx f(\vec{x}_\star) + \underbrace{\left[D_{\vec{x}} f|_{\vec{x}_\star}\right]}_{\text{row vector}} (\Delta\vec{x}) + \boxed{\frac{1}{2}\Delta\vec{x}^T (H_{\vec{x}} f) \Delta\vec{x}} \quad (14)$$

What populates the underlined space?

$H_{\vec{x}} f$ matrix

$\Delta\vec{x} = (\vec{x} - \vec{x}_\star)$

$\cancel{H_{\vec{x}} f (\Delta\vec{x})}$ vector

Hessian is the change in derivative (derivative of derivative)

$\underbrace{D_{\vec{x}} f}_{\text{row}} + \underbrace{\cancel{H_{\vec{x}} f}\ \Delta\vec{x}}_{\text{row or col?}}$

$\underset{1\times n}{D_{\vec{x}} f} + \underbrace{\overset{1\times n}{\Delta\vec{x}^T}\overset{n\times n}{(H_{\vec{x}} f)}}_{\substack{\text{row or col?} \\ 1\times n}}$

we want something like

$\underbrace{\underset{1\times n}{\Delta\vec{x}^T}\ \underset{n\times n}{(H_{\vec{x}} f)}}\ \underbrace{\underset{n\times 1}{\Delta\vec{x}}}$

1×1 scalar

$\underbrace{\frac{\partial^2 f}{\partial x \partial y} \delta y\, \delta x}$ $\qquad \frac{\partial^2 f}{\partial x_i \partial x_j} \Delta x_j \Delta x_i$

(f) 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 first derivative changed by moving a small amount $\Delta\vec{w}$, how would you get such an estimate using your expression for the second derivative?**

$\Delta\vec{x}$

$$D_{\vec{x}} f(\vec{x}) = D_{\vec{x}} f\Big|_{\vec{x}_\star} + \Delta\vec{x}^T \left[(H_{\vec{x}} f)\Big|_{\vec{x}_\star}\right]$$

function of $\vec{x}$

$\Delta\vec{x}^T = (\vec{x}^T - \vec{x}_\star^T)$

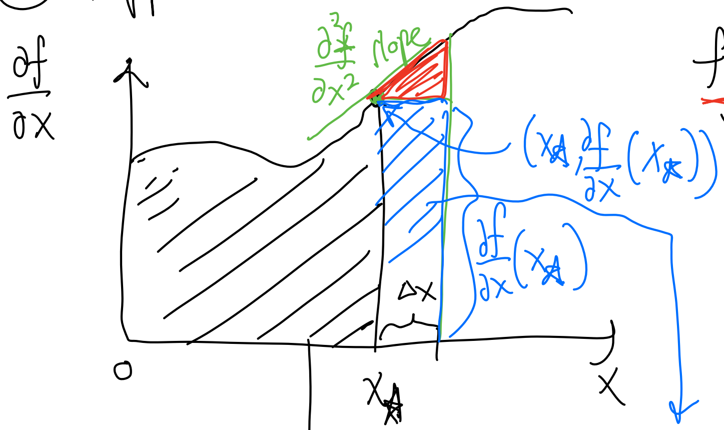
think about linearizing derivative because it's also a function of $\vec{x}$

Why the response to ⓔ? Why the $\frac{1}{2}$ term

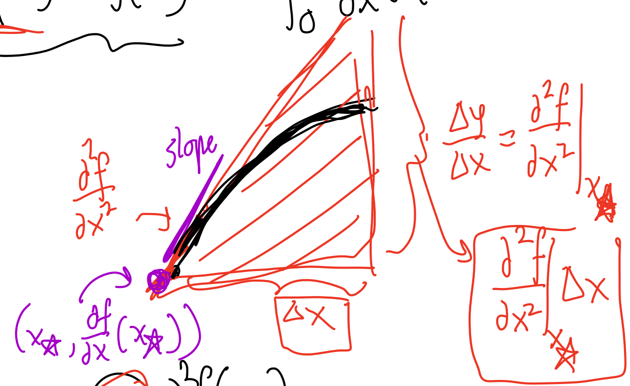
① The math works out w/ taylor expansion ←

② Units perspective ←

③ Approximation of derivative perspective

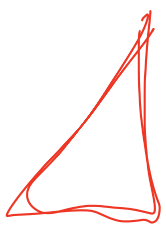


$$f(x) - f(0) = \int_0^x \frac{\partial}{\partial x} f(x)\, dx$$



$$f(x) = f(x_\star) + \frac{\partial f(x_\star)}{\partial x}\overbrace{(x - x_\star)}^{\Delta x} + \frac{1}{2}\underbrace{\frac{\partial^2 f(x_\star)}{\partial x^2}(x - x_\star)^2}$$

Area of triangle: $\frac{1}{2} b \cdot h$ ← this is where the half term comes from



$h = \frac{\partial^2 f}{\partial x^2}\Big|_{x_\star} \Delta x$

$b = \Delta x$

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

$$f(\vec{w}_\star + \Delta\vec{w}) \approx e^{\vec{x}^\top \vec{w}_\star}\left(1 + \vec{x}^\top(\Delta\vec{w}) + \frac{1}{2}\left(\vec{x}^\top(\Delta\vec{w})\right)^2\right). \tag{15}$$

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

Hint: You can compute the following partial derivatives:

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$\frac{\partial f}{\partial w_i} = x_i f(\vec{w}) \tag{16}$$

$$\boxed{\frac{\partial^2 f}{\partial w_j \partial w_i} = x_i x_j f(\vec{w}).} \tag{17}$$

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

$$\frac{\partial}{\partial w_i} f = \frac{\partial}{\partial w_i}\left(e^{\vec{x}^T \vec{w}}\right) = e^{\vec{x}^T\vec{w}} \frac{\partial}{\partial w_i}\left(\vec{x}^T \vec{w}\right)$$

$$= e^{\vec{x}^T\vec{w}} \frac{\partial}{\partial w_i}\left(\sum_{j=1}^{n} w_j x_j\right) = e^{\vec{x}^T\vec{w}}\left[0 + \frac{\partial}{\partial w_i}(w_i x_i) + 0 + 0\right]$$

$$1 \times n$$

$$\rightarrow D_{\vec{x}} f = \left[e^{\vec{x}^T\vec{w}} x_1 \ \cdots \ e^{\vec{x}^T\vec{w}} x_n\right] = e^{\vec{x}^T\vec{w}}\left[x_1 \ \cdots \ x_n\right] = e^{\vec{x}^T\vec{w}}\vec{x}^T$$

$$= e^{\vec{x}^T\vec{w}} x_i$$

$$\rightarrow H_{\vec{x}} f = \begin{bmatrix} e^{\vec{x}^T\vec{w}} x_1 \cdot x_1 & e^{\vec{x}^T\vec{w}} x_1 \cdot x_2 & \cdots \\ \vdots & & \\ \vdots & & \end{bmatrix} = e^{\vec{x}^T\vec{w}} \begin{bmatrix} x_1 x_1 & x_1 x_2 & x_1 x_3 & \cdots \\ \vdots & & & \\ x_n x_1 & x_n x_2 & \cdots & x_n x_n \end{bmatrix} = e^{\vec{x}^T\vec{w}} \vec{x}\vec{x}^T$$

$$n \times n$$

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

$$f(\vec{x}) \approx e^{\vec{x}^T\vec{w}_\star} + e^{\vec{x}^T\vec{w}_\star}\vec{x}^T\Delta\vec{w} + \frac{1}{2}\Delta\vec{w}^T\left(e^{\vec{x}^T\vec{w}_\star}\vec{x}\vec{x}^T\right)\Delta\vec{w}$$

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

Quadratic approximation for $f(\vec{w}) = \sum_{i=1}^{m} e^{\vec{x}_i^T\vec{w}}$

will be sum of quadratic approximations of $e^{\vec{x}_i^T\vec{w}}$ separately

(derivatives are linear / distribute over things that add)

$$f(\vec{w}) \approx \sum_{i=1}^{m} e^{\vec{x}_i^T\vec{w}}\left(1 + \vec{x}_i^T\Delta\vec{w} + \frac{1}{2}\left(\Delta\vec{w}^T\vec{x}_i\right)^2\right)$$

Contributors:

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

