Section 03: Solutions

This week in section, we'll be focusing on vector calculus. See this week's section solutions on the course website for more content related to the bias-variance tradeoff discussed in lecture this week.

Solution:

Section Plan

- Reminders (5 min)
- Explain Gradients/Jacobians/Hessians. Slides exist for this, but teach however you see fit (15 min)
- Students do 1.2 a, b (10 min)
- Talk about approximations: 1.2 c, d.
- Introduce 1.2 g and have the students do it (5 min)

1. Vector Calculus

1.1. Definitions

Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $g : \mathbb{R}^n \to \mathbb{R}^m$. The **gradient** of f(with respect to x) evaluated at x is the vector of partial derivatives:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

The **Jacobian** of g(with respect to x) evaluated at x is the matrix of partial derivatives:

$$\nabla_x g(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_x^T g_1(x) \\ \vdots \\ \nabla_x^T g_m(x) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Sometimes the Jacobian is denoted by $J_g(x)$, but we use $\nabla_x g(x)$ to highlight that the Jacobian is nothing more than the generalization of the gradient to functions which have a vector output.

The **Hessian** of f(with respect to x) evaluated at x is the matrix of partial derivatives:

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Sometimes the Hessian is denoted by $H_f(x)$, but we use $\nabla_x^2 f(x)$ to highlight that the Hessian is the Jacobian of the gradient of f.

1.2. Estimation

What the gradient and Jacobian at a point do is express how the output of a function changes when the input is changed by a small amount. Thus, they can be used to approximate the values of a function close to the point at which they are evaluated. Let's see how we can do this for one variable. Let $f : \mathbb{R} \to \mathbb{R}$:

$$\frac{df}{dx}(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow \frac{df}{dx}(x) \approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow f(x+\epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x)$$

Let us now extend this to multiple dimensions and derive the definition of the gradient starting from this approximation view point. Suppose we have a function $f : \mathbb{R}^n \to \mathbb{R}$ and we want to determine how the function changes around a point $x \in \mathbb{R}^n$. First we will determine how the function changes when we slightly vary its first coordinate:

$$f(x_1 + \epsilon_1, \ldots, x_n) \approx f(x_1, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n)$$

Now, let us slightly vary the first two coordinates:

$$\begin{aligned} f(x_1 + \epsilon_1, x_2 + \epsilon_2, \ldots, x_n) &\approx f(x_1, x_2 + \epsilon_2, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2 + \epsilon_2, \ldots, x_n) \\ &\approx f(x_1, x_2, \ldots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n) + \\ &+ \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) + \epsilon_1 \epsilon_2 \frac{\partial f}{\partial x_2} \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) \\ &\approx f(x_1, x_2, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n) \end{aligned}$$

where we eliminate the term where $\epsilon_1 \epsilon_2$ because it would be very small compared to the others. Repeating the process for all n dimensions we obtain the approximation:

$$f(x_1 + \epsilon_1, \ldots, x_n + \epsilon_n) \approx f(x_1, \ldots, x_n) + \sum_{i=1}^{n} \epsilon_i \frac{\partial f}{\partial x_i}(x_1, x_2, \ldots, x_n)$$

Let $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$ and $x = [x_1, \ldots, x_n]^T$, then we can rewrite the above as:

$$f(x + \epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon$$

Questions:

(a) Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 e^{x_1 x_2} \\ x_1 e^{x_1 x_2} + \frac{2}{x_2} \end{bmatrix} \text{ and } \nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix}$$

(b) Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

Solution:

$$\nabla_x (\nabla_x f)(x) = \begin{bmatrix} \frac{\partial (\nabla_x f)_1(x)}{\partial x_1} & \frac{\partial (\nabla_x f)_1(x)}{\partial x_2} \\ \frac{\partial (\nabla_x f)_2(x)}{\partial x_1} & \frac{\partial (\nabla_x f)_2(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix} = \nabla_x^2 f(x)$$

(c) The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

Solution:

The Jacobian also offers the best linear approximation of g around a point x, but now it approximates a vector, instead of a scalar,

$$g(x + \epsilon) \approx g(x) + \nabla_x g(x) \epsilon$$

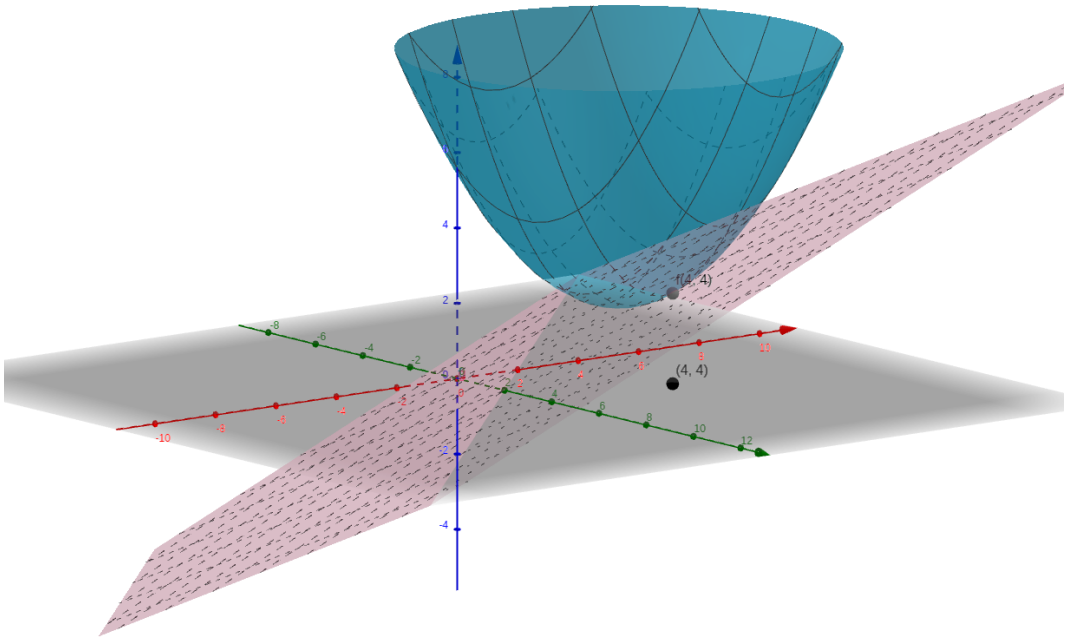

Figure 1: Graph of the function f and the tangent plane.

where $\nabla_x g(x)\epsilon$ is a matrix multiplication instead of a dot product.

(d) If we use the gradient and the Hessian of $f : \mathbb{R}^n \to \mathbb{R}$, what type of an approximation for the function f around a point x can we create.

Solution:

Using the two, we can create the best quadratic approximation of f, given by:

$$f(x+\epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon + \frac{1}{2}\epsilon^T \nabla_x^2 f(x)\epsilon$$

(e) Consider the function $f(x_1, x_2) = 2 + 0.2(x_1 - 3)^2 + 0.2(x_2 - 3)^2$ which is graphed in figure 1. The pink plane is the tangent plane for the point $x = (4, 4)$ and it represents the graph of the best linear approximation of f around the point x. What is the function describing the tangent plane:

Solution:

$$f((4,4)) = 2.4 \text{ and } \nabla_x f(x) = \begin{bmatrix} 0.4(x_1 - 3) \\ 0.4(x_2 - 3) \end{bmatrix} \text{ and } \nabla_x f((4,4)) = \begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix}$$

Tangent plane function:

$$\hat{f}(y) = f(x) + \nabla_x f(x)^T (y - x) = 2.4 + \begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix}^T \left(y - \begin{bmatrix} 4 \\ 4 \end{bmatrix}\right)$$

(f) One thing to note is that the linear approximation becomes very poor once we move away from x. Suppose we want a better approximation. For this purpose, we can use the Hessian as explained in part 2. Write down this approximation for an arbitrary x. How good would this approximation be?

Solution:

$$f(y) = f(x) + \nabla_x f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla_x^2 f(x)(y - x)$$

We would be able to recreate f perfectly since f is quadratic.

(g) Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

Solution:

- d_1: Value of approximation goes up.
- d_2: Value of approximation goes down.
- d_3: Value of approximation stays the same.

The same can be said for f, but only in the immediate vicinity of the point x.

1.3. Algebra

Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$, . Below is a list of important gradient properties:

- **Gradient of constant:** $\nabla_x c = 0 \in \mathbb{R}^n$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule**: $\nabla_x(fg)(x) = \nabla_x f(x) \cdot g(x) + \nabla_x g(x) \cdot f(x)$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^m \to \mathbb{R}^k$, $l : \mathbb{R}^m \to \mathbb{R}$. Below is a list of important Jacobian properties:

- **Jacobian of constant:** $\nabla_x c = 0 \in \mathbb{R}^{n \times m}$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule**: $\nabla_x(f^T g)(x) = [\nabla_x f(x)]^T g(x) + [\nabla_x g(x)]^T f(x)$.
- **Chain rule**: $\nabla_x(h \circ g)(x) = \nabla_{g(x)} h(g(x)) \nabla_x g(x)$ and $\nabla_x(l \circ g)(x) = \left[[\nabla_{g(x)} l(g(x))]^T \nabla_x g(x)\right]^T$.

Questions:

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = v^T x$ for $v \in \mathbb{R}^n$. Using the definition of the gradient, write out $\nabla_x f(x)$ and specify its dimensions.

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial v^T x}{\partial x_1} \\ \vdots \\ \frac{\partial v^T x}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_i v_i x_i \\ \vdots \\ \frac{\partial}{\partial x_n} \sum_i v_i x_i \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = v \in \mathbb{R}^n$$

(b) Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be $f(x) = x$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \dots & \frac{\partial x_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_m}{\partial x_1} & \dots & \frac{\partial x_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} = I \in \mathbb{R}^{n \times n}$$

(c) Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be $f(x) = Ax$ for $A \in \mathbb{R}^{m \times n}$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial (Ax)_1}{\partial x_1} & \dots & \frac{\partial (Ax)_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial (Ax)_m}{\partial x_1} & \dots & \frac{\partial (Ax)_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_k A_{1k} x_k & \dots & \frac{\partial}{\partial x_n} \sum_k A_{1k} x_k \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} \sum_k A_{mk} x_k & \dots & \frac{\partial}{\partial x_n} \sum_k A_{mk} x_k \end{bmatrix}$$
$$= \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \dots & A_{mn} \end{bmatrix} = A \in \mathbb{R}^{m \times n}$$

(d) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = \alpha v^T x + \beta w^T x$ where $\alpha, \beta \in \mathbb{R}$ and $v, w \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

$$\nabla_x f(x) = \nabla_x \left(\alpha v^T x + \beta w^T x \right) = \alpha \nabla_x v^T x + \beta \nabla_x w^T x = \alpha v + \beta w$$

(e) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = x^T A x$ and $A \in \mathbb{R}^{n \times n}$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

$$\nabla_x f(x) = \nabla_x \left(x^T A x \right) = (\nabla_x x)^T (Ax) + (\nabla_x Ax)^T x = IAx + A^T x = (A + A^T)x$$

where we used the product rule and split $x^T Ax$ into $g(x)^T h(x)$ where $g, h : \mathbb{R}^n \to \mathbb{R}^n$ and $g(x) = x$ and $h(x) = Ax$

(f) With f defined as in the previous part, what is the Hessian of f. Only use previously proven facts and recall that the Hessian is the Jacobian of the gradient.

Solution:

$$\nabla_x^2 f(x) = \nabla_x (\nabla_x f)(x) = \nabla_x (A + A^T)x = A + A^T$$

where we used part 3 in the last step.

(g) Let $f : \mathbb{R}^m \to \mathbb{R}$ be $f(x) = (Ax - y)^T W (Ax - y)$ and $A \in \mathbb{R}^{m \times n}, W \in \mathbb{R}^{n \times n}, y \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

Let $f = h \circ g$ where $g(x) = Ax - y$ and $h(z) = z^T W z$. Using the chain rule and parts 3 and 5, we can derive:

$$\begin{aligned}\nabla_x f(x) = \nabla_x (h \circ g)(x) &= \left[[\nabla_{g(x)} h(g(x))]^T \nabla_x g(x) \right]^T \\ &= \left[[(W + W^T)(Ax - y)]^T A \right]^T \\ &= A^T (W + W^T)(Ax - y)\end{aligned}$$