Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

7.X.4 The Gradient: a vector input, scalar output

Our plan is to look at every derivative combination of scalars, vectors, and matrices we can.

First, we consider:

$$\frac{\partial(\text{Scalar})}{\partial(\text{Vector})} = \frac{\partial s}{\partial v} \tag{1}$$

We'll take s to be our scalar, and v to be our vector. So, our input is a **vector**, and our output is a **scalar**.

$$\Delta v \longrightarrow \boxed{f} \longrightarrow \Delta s \tag{2}$$

How do we make sense of this? Well, let's write Δv_i explicitly:

$$\overbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}^{\Delta v} \longrightarrow \Delta s \tag{3}$$

We can see that we have m different **inputs** we can change in order to change our **one** output.

So, our derivative needs to have m different **elements**: one for each element v_i.

7.X.5 Finding the scalar/vector derivative

But how do we shape our matrix? Let's look at our **rule**.

$$\Delta s \approx \frac{\partial s}{\partial v} \star \Delta v \qquad \text{or} \qquad \Delta s \approx \frac{\partial s}{\partial v} \star \overbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}^{\Delta v} \tag{4}$$

How do we get $\Delta reds$? We have so many variables. Let's focus on them one at a time: breaking Δv into Δv_i, so we'll try to consider each v_i **separately**.

It's usually possible to change each v_i, so we have to look at every one of them.

One problem, though: how can we treat each **derivative** separately? Each Δv_i will move our position, which can change a different derivative v_k: they can **affect** each other.

7.X.6 Review: Planar Approximation

We'll resolve this the same way we did in chapter 3, **gradient** descent: by taking advantage of the "planar approximation".

The solution is this: assume your function is **smooth**. The **smaller** a step you take, the **less** your derivative has a chance to change.

This isn't true for big steps, but eventually, if your step is small enough, then the derivative will barely change.

Example: Take $f(x) = x^2$.

- If we go from $x = 1 \to 2$, then our derivative goes from $f'(x) = 2 \to 4$.
- Let's **shrink** our step. We go from $x = 1 \to 1.01$, our derivative goes from $f'(x) = 2 \to 2.02$.
  - Our derivative is almost the same!

if we take a small enough step Δv_i, then, if our function is **smooth**, then the derivative will hardly change!

You could imagine repeatedly shrinking the size of our step, until the change in the derivatives is basically unnoticeable.

So, if we zoom in enough (shrink the scale of change), then we can **pretend** the derivative is **constant**.

Concept 1

If you have a **smooth function**, then...

If you take sufficiently **small steps**, then you can treat the derivatives as **constant**.

Clarification

This section is **optional**.

We can describe "sufficiently small steps" in a more mathematical way:

Our goal is for $f'(x)$ to be **basically constant**: it doesn't change much. $\Delta f'(x)$ is **small**.

Let's say it can't change more then δ.

If you want

- $\Delta f'(x)$ to be very small ($|\Delta f'(x)| < \delta$)
- It has been proven that...
  - can take a small enough step $|\Delta x| < \epsilon$, and to get that result.

One way to describe this is to say that our function is (locally) **flat**: it looks like some kind of plane/hyperplane.

The word "locally" represents the small step size: we stay in the "local area".

Clarification 2

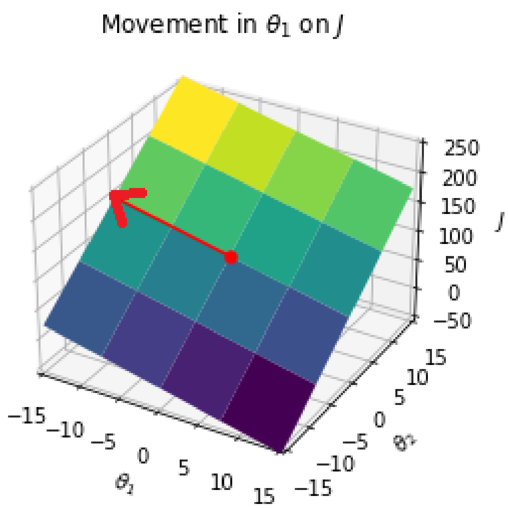
Why is this **true**? Because a **hyperplane** can be represented using our **linear** function

$$f(x) \approx \theta^T x + \theta_0 = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

If we take a derivative:

$$\frac{\partial f}{\partial x_i} = \theta_i$$

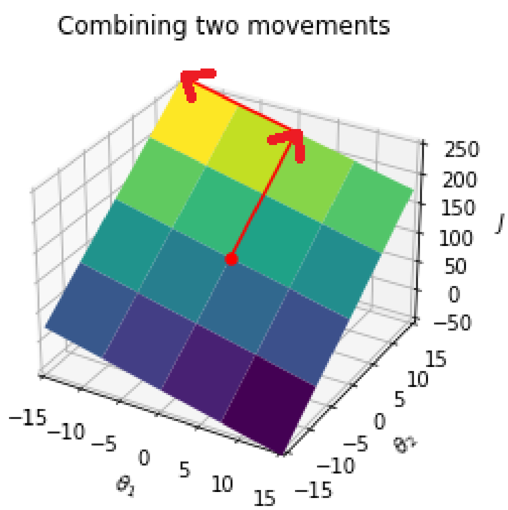
That derivative is a **constant**! It's doesn't change based on **position**.

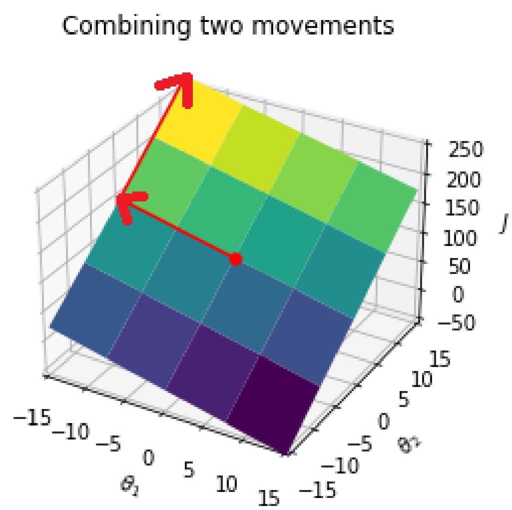


If we take very small steps, we can approximate our function as **flat**.

Why does this help? If our derivative doesn't **change**, we can combine multiple steps You can take multiple steps Δv_i and the order doesn't matter.

So, you can combine your steps or separate them easily.





We can break up our big step into two smaller steps that are truly independent: order doesn't matter.

With that, we can add up all of our changes:

$$\Delta s = \Delta s_{\text{from } v_1} + \Delta s_{\text{from } v_2} + \cdots + \Delta s_{\text{from } v_m} \tag{5}$$

7.X.7 Our scalar/vector derivative

From this, we can get an **approximated** version of the MV chain rule.

Definition 3

The **multivariable chain rule approximation** looks similar to the multivariable chain rule, but for finite changes Δx_i.

In 3-D, we get

$$\Delta f = \overbrace{\frac{\partial f}{\partial x}\Delta x}^{x \text{ component}} + \overbrace{\frac{\partial f}{\partial y}\Delta y}^{y \text{ component}} + \overbrace{\frac{\partial f}{\partial z}\Delta z}^{z \text{ component}}$$

In general, we have

$$\Delta f = \sum_{i=1}^{m} \overbrace{\frac{\partial f}{\partial x_i}\Delta x_i}^{x_i \text{ component}}$$

This function lets us add up the effect each component has on our output, using **derivatives**.

This gives us what we're looking for:

$$\Delta s \approx \sum_{i=1}^{m} \frac{\partial s}{\partial v_i} \Delta v_i \tag{6}$$

If we circle back around to our original approximation:

$$\sum_{i=1}^{m} \frac{\partial s}{\partial v_i} \Delta v_i = \frac{\partial s}{\partial v} \star \overbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}^{\Delta v} \tag{7}$$

When we look at the left side, we're multiplying pairs of components, and then adding them. That sounds similar to a **dot product**.

$$\sum_{i=1}^{m} \frac{\partial s}{\partial v_i} \Delta v_i = \overbrace{\begin{bmatrix} \partial s/\partial v_1 \\ \partial s/\partial v_2 \\ \vdots \\ \partial s/\partial v_m \end{bmatrix}}^{\partial s/\partial v} \cdot \overbrace{\begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \vdots \\ \Delta v_m \end{bmatrix}}^{\Delta v} \tag{8}$$

This gives us our derivative: it contains all of the **element-wise** derivatives we need, and in a **useful** form!

Definition 4

If s is a **scalar** and v is an $(m \times 1)$ **vector**, then we define the **derivative** or **gradient** $\partial s / \partial v$ as fulfilling:

$$\Delta s = \frac{\partial s}{\partial v} \cdot \Delta v$$

Or, equivalently,

$$\Delta s = \left(\frac{\partial s}{\partial v}\right)^T \Delta v$$

Thus, our derivative must be an $(m \times 1)$ vector

$$\frac{\partial s}{\partial v} = \begin{bmatrix} \partial s / \partial v_1 \\ \partial s / \partial v_2 \\ \vdots \\ \partial s / \partial v_m \end{bmatrix} = \begin{bmatrix} \frac{\partial s}{\partial v_1} \\ \frac{\partial s}{\partial v_2} \\ \vdots \\ \frac{\partial s}{\partial v_m} \end{bmatrix}$$

We can see the shapes work out in our matrix multiplication:

$$\overbrace{\Delta s}^{(1\times 1)} = \overbrace{\left(\frac{\partial s}{\partial v}\right)^T}^{(1\times m)} \overbrace{\Delta v}^{(m\times 1)} \tag{9}$$