Partial Derivatives and Gradients

Last modified: 2026-11-24 09:19:45 (PST)

(adapted from Fieller [1], Section 7.2)

This section covers derivatives of functions of vectors and matrices. Its linear algebra prerequisites, such as vectors, matrices, transposes, dot products, and quadratic forms, are covered in Linear Algebra.

Let $\underset{\sim}{x}$ and $\underset{\sim}{\beta}$ be column vectors of length p (see column vector and dot product).

Definition 0.1 (Partial derivative)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ and $\underset{\sim}{x} = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$. The **partial derivative** [2] of f with respect to x_j at $\underset{\sim}{x}$ is

$$\frac{\partial}{\partial x_j} f(\underset{\sim}{x}) \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{f(x_1, \dots, x_j + h, \dots, x_p) - f(x_1, \dots, x_p)}{h},$$

when this limit exists. It is the derivative at x_j of the one-variable function $t \mapsto f(x_1, \dots, t, \dots, x_p)$, which varies the jth coordinate and holds the other coordinates fixed.

Example 0.1 (Partial derivatives of $x_1^2 x_2 + 3x_2$)**.** Let $f(x_1, x_2) = x_1^2 x_2 + 3x_2$. At $\underset{\sim}{x} = (1, 2)^\top$,

$$\begin{aligned} f(1, 2) &= 1 \cdot 2 + 6 \\ &= 8, \end{aligned}$$

and for $h \neq 0$

$$
\begin{aligned}
\frac{f(1+h,2)-f(1,2)}{h} &= \frac{(1+h)^2\cdot 2+6-8}{h} && \text{(substitute into } f\text{)}\\
&= \frac{(1+h)^2\cdot 2-2}{h} && (6-8=-2)\\
&= \frac{(1+2h+h^2)\cdot 2-2}{h} && \text{(expand } (1+h)^2\text{)}\\
&= \frac{2+4h+2h^2-2}{h} && \text{(distributive law)}\\
&= \frac{4h+2h^2+2-2}{h} && \text{(reorder the terms)}\\
&= \frac{4h+2h^2}{h} && (2-2=0)\\
&= 4+2h, && \text{(divide by } h\neq 0\text{)}
\end{aligned}
$$

which tends to 4 as $h \to 0$, so $\frac{\partial}{\partial x_1} f(1,2) = 4$. Holding x_2 fixed and differentiating in x_1 gives the same answer: $\frac{\partial}{\partial x_1} f(\underset{\sim}{x}) = 2x_1x_2$, which is $2\cdot 1\cdot 2 = 4$ at $(1,2)^\top$. Holding x_1 fixed instead gives $\frac{\partial}{\partial x_2} f(\underset{\sim}{x}) = x_1^2 + 3$, which is $1+3=4$ at $(1,2)^\top$.

Exercise 0.1 (First-order partial derivatives)**.** Adapted from Miller [3], Question 1.1.42. Let $f(x,y) = x^2y + e^x + \sin(xy)$. Find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

Solution

Solution 0.1. To compute $\frac{\partial f}{\partial x}$, treat y as a constant:

$$\frac{\partial f}{\partial x} = 2xy + e^x + y\cos(xy)$$

To compute $\frac{\partial f}{\partial y}$, treat x as a constant:

$$
\begin{aligned}
\frac{\partial f}{\partial y} &= x^2 + 0 + x\cos(xy)\\
&= x^2 + x\cos(xy)
\end{aligned}
$$

Exercise 0.2 (Partial derivatives of the normal density with respect to parameters)**.** Adapted from Miller [3], Question 1.1.43.
Let

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Find $\frac{\partial f}{\partial \mu}$ and $\frac{\partial f}{\partial \sigma}$.

Solution

Solution 0.2. **1. Derivative with respect to μ:** Treat σ as constant. Differentiating the exponential factor via the chain rule:

$$\begin{aligned}\frac{\partial f}{\partial \mu} &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\cdot\left[-\frac{2(x-\mu)(-1)}{2\sigma^2}\right] \\ &= \frac{x-\mu}{\sigma^2}f(x;\mu,\sigma)\end{aligned}$$

2. Derivative with respect to σ: Write $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ and apply the product rule:

$$\begin{aligned}\frac{\partial f}{\partial \sigma} &= -\frac{1}{\sqrt{2\pi}}\sigma^{-2}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} + \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\cdot\left[\frac{(x-\mu)^2}{\sigma^3}\right] \\ &= \left[-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right]f(x;\mu,\sigma)\end{aligned}$$

Remark: Setting $\frac{\partial f}{\partial \mu} = 0$ and $\frac{\partial f}{\partial \sigma} = 0$ for a sample of independent observations leads directly to the values that make the sample most probable, which statistics calls the maximum likelihood estimates: $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Theorem 0.1 (Partial derivatives of the normal density with respect to its parameters)**.** *Let $\mu \in \mathbb{R}$ and $\sigma > 0$, and let*

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Then for every $x \in \mathbb{R}$,

$$\begin{aligned}\frac{\partial f}{\partial \mu} &= \frac{x-\mu}{\sigma^2}f(x;\mu,\sigma) \\ \frac{\partial f}{\partial \sigma} &= \left[-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right]f(x;\mu,\sigma)\end{aligned}$$

Proof

Proof. This is the solution to Exercise 0.2, worked in Solution 0.2. The argument has two steps.

1. For $\frac{\partial f}{\partial \mu}$, treat σ as constant. The prefactor is constant, so the chain rule on the exponential gives $\frac{\partial f}{\partial \mu} = f(x;\mu,\sigma) \cdot \left[-\frac{2(x-\mu)(-1)}{2\sigma^2}\right] = \frac{x-\mu}{\sigma^2} f(x;\mu,\sigma)$.
2. For $\frac{\partial f}{\partial \sigma}$, write $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$. The product rule gives one term from differentiating σ^{-1}, namely $-\sigma^{-1} f(x;\mu,\sigma)$, and one from the chain rule on the exponential, namely $\frac{(x-\mu)^2}{\sigma^3} f(x;\mu,\sigma)$. Adding them gives $\left[-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right] f(x;\mu,\sigma)$.

□

Exercise 0.3 (Partial derivatives with exponential of quadratic form)**.** Adapted from Miller [3], Question 1.1.44.
Find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ for

$$f(x,y) = x\exp\{x^2+y^2\}$$

Solution

Solution 0.3. With respect to x, apply the product rule:

$$\begin{aligned}\frac{\partial f}{\partial x} &= (1)\exp\{x^2+y^2\} + x\cdot[2x\exp\{x^2+y^2\}] \\ &= (1+2x^2)\exp\{x^2+y^2\}\end{aligned}$$

With respect to y, treat x as constant:

$$\begin{aligned}\frac{\partial f}{\partial y} &= x\cdot[2y\exp\{x^2+y^2\}] \\ &= 2xy\exp\{x^2+y^2\}\end{aligned}$$

Exercise 0.4 (Using symmetry to compute partial derivatives)**.** Adapted from Miller [3], Question 1.1.45.
Find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ for

$$f(x,y) = e^{xy} - \log(x^2+y^2)$$

on $\mathbb{R}^2 \setminus \{(0,0)\}$.

Solution

Solution 0.4. Holding y constant, differentiate with respect to x:

$$\frac{\partial f}{\partial x} = ye^{xy} - \frac{2x}{x^2+y^2}$$

Notice that the function is symmetric in x and y: $f(x,y) = f(y,x)$. Interchanging x and y gives $\frac{\partial f}{\partial y}$ immediately:

$$\frac{\partial f}{\partial y} = xe^{xy} - \frac{2y}{x^2+y^2}$$

Remark: Recognizing symmetry in multi-variable functions eliminates redundant calculations.

Definition 0.2 (Vector derivative, gradient)**.** If $f(\underset{\sim}{\beta})$ is a scalar-valued function of a $p \times 1$ vector $\underset{\sim}{\beta}$, such as $f(\underset{\sim}{\beta}) = \underset{\sim}{x} \cdot \underset{\sim}{\beta}$, then its **vector derivative** is the column vector of its partial derivatives (Definition 0.1):

$$\frac{\partial}{\partial \underset{\sim}{\beta}} f(\underset{\sim}{\beta}) = \begin{bmatrix} \frac{\partial}{\partial \beta_1} f(\underset{\sim}{\beta}) \\ \frac{\partial}{\partial \beta_2} f(\underset{\sim}{\beta}) \\ \vdots \\ \frac{\partial}{\partial \beta_p} f(\underset{\sim}{\beta}) \end{bmatrix}$$

The vector derivative is also called the **gradient** [4] of f with respect to $\underset{\sim}{\beta}$, written $\nabla_{\underset{\sim}{\beta}} f(\underset{\sim}{\beta})$.

Example 0.2 (A vector derivative)**.** Let $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$ and $f(\underset{\sim}{\beta}) = 3\beta_1 + 5\beta_2^2$. Then

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} f(\underset{\sim}{\beta})}_{2\times 1} = \begin{bmatrix} \frac{\partial}{\partial \beta_1}(3\beta_1 + 5\beta_2^2) \\ \frac{\partial}{\partial \beta_2}(3\beta_1 + 5\beta_2^2) \end{bmatrix}$$
$$= \begin{bmatrix} 3 \\ 10\beta_2 \end{bmatrix},$$

and at $\underset{\sim}{\beta} = (1,2)^\top$ the vector derivative is $(3,20)^\top$.

Video lecture

Hutchinson's Gradients Refresher (17 min) covers gradients and partial derivatives, the ideas behind this section [5]. The login for the video site is posted on Canvas.

Videos

Jon Krohn's "Calculus for Machine Learning" YouTube playlist has videos on partial derivatives:

- What Partial Derivatives Are (Hands-on Introduction)
- Partial Derivative Exercises
- Advanced Partial Derivatives
- Advanced Partial-Derivative Exercises
- Partial Derivative Notation

Checking a gradient by its shape

The gradient always has the same shape as the input it is taken with respect to, because it holds exactly one partial derivative per component of that input:

Table 1: The shape of a gradient follows the shape of what it differentiates with respect to.

f is a function of	its gradient is
a single number	a single number — the ordinary derivative
a vector in $\mathbb{R}^p$	a vector in $\mathbb{R}^p$
a matrix in $\mathbb{R}^{m\times n}$	a matrix in $\mathbb{R}^{m\times n}$

The table is the cheapest check there is on a gradient just worked out by hand: if the shapes disagree, something went wrong. It is the shape rule of matrix multiplication again, applied to calculus.

Definition 0.1 (Linear approximation of a function of a vector, first-order Taylor approximation of a function of a vector)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ have a gradient at β. The **linear approximation** of f at β [6] (also called the **first-order Taylor approximation** or the **first-order Taylor expansion**) is the function of a step $\underset{\sim}{\delta} \in \mathbb{R}^p$

$$\hat{f}_{\underset{\sim}{\beta}}(\underset{\sim}{\delta}) \stackrel{\text{def}}{=} f(\underset{\sim}{\beta}) + \nabla f(\underset{\sim}{\beta}) \cdot \underset{\sim}{\delta}. \tag{1}$$

Example 0.1 (A linear approximation of a function of two variables)**.** Take $f(\beta) = 3\beta_1 + 5\beta_2^2$ from Example 0.2, whose gradient at $\underset{\sim}{\beta} = (1, 2)^\top$ is $\nabla f(\underset{\sim}{\beta}) = (3, 20)^\top$. At that point,

$$\begin{aligned} f(\underset{\sim}{\beta}) &= 3 \cdot 1 + 5 \cdot 2^2 && \text{(substitute } \beta_1 = 1 \text{ and } \beta_2 = 2) \\ &= 23, && (3 + 20 = 23) \end{aligned}$$

so the linear approximation at $\underset{\sim}{\beta}$ is

$$\hat{f}_{\underset{\sim}{\beta}}(\underset{\sim}{\delta}) = 23 + 3\delta_1 + 20\delta_2.$$

For the step $\underset{\sim}{\delta} = (0.1, 0.1)^\top$, the approximation is

$$\begin{aligned} \hat{f}_{\underset{\sim}{\beta}}(\underset{\sim}{\delta}) &= 23 + 3 \cdot 0.1 + 20 \cdot 0.1 && \text{(substitute } \delta_1 = \delta_2 = 0.1) \\ &= 25.3, && (23 + 0.3 + 2 = 25.3) \end{aligned}$$

and the value of f at $\underset{\sim}{\beta} + \underset{\sim}{\delta} = (1.1, 2.1)^\top$ is

$$\begin{aligned} f(\underset{\sim}{\beta} + \underset{\sim}{\delta}) &= 3 \cdot 1.1 + 5 \cdot 2.1^2 && \text{(substitute } \beta_1 = 1.1 \text{ and } \beta_2 = 2.1) \\ &= 25.35. && (3.3 + 22.05 = 25.35) \end{aligned}$$

The approximation misses by 0.05, which is $5\delta_2^2$, the second-order term of f that the approximation leaves out. That term is never negative, and is zero only when $\delta_2 = 0$, so the graph of the approximation in Figure 1 is a plane that touches the graph of f along the line $\beta_2 = 2$ and lies below it everywhere else.

(This 3D surface is interactive; see the HTML or slide version of this page to view it.)

Figure 1: The graph of $f(\underset{\sim}{\beta}) = 3\beta_1 + 5\beta_2^2$ (blue) and its linear approximation at $(1, 2)^\top$ (vermilion), drawn as a function of the point $(1, 2)^\top + \underset{\sim}{\delta}$. The two surfaces touch along the line $\beta_2 = 2$.

i Remark (The one-variable linear approximation is the case $p = 1$)

Remark 0.1 (The one-variable linear approximation is the case $p = 1$). For $p = 1$, the gradient of f is its derivative, so Equation 1 is the one-variable linear approximation, with the step $\underset{\sim}{\delta}$ in place of ε.

Definition 0.2 (Direction)**.** A **direction** in $\mathbb{R}^p$ is given by a nonzero vector $\underset{\sim}{d} \in \mathbb{R}^p$: moving from a point $\underset{\sim}{x}$ in the direction $\underset{\sim}{d}$ means moving to the points $\underset{\sim}{x} + t\underset{\sim}{d}$ with $t > 0$. For any $c > 0$, $c\underset{\sim}{d}$ reaches the same points (with t/c in place of t), so $\underset{\sim}{d}$ and $c\underset{\sim}{d}$ give the same direction.

Example 0.2 (Moving from $(2, -1)^\top$ in the direction $(1, 1)^\top$)**.** From $\underset{\sim}{x} = (2, -1)^\top$, moving in the direction $\underset{\sim}{d} = (1, 1)^\top$ reaches the points $\underset{\sim}{x} + t\underset{\sim}{d} = (2 + t, -1 + t)^\top$ for

$t > 0$, such as $(3,0)^\top$ at $t = 1$. The vector $(2,2)^\top$ gives the same direction: it reaches $(3,0)^\top$ at $t = \frac{1}{2}$. The vector $(-1,-1)^\top$ gives the opposite direction, reaching points such as $(1,-2)^\top$.

Definition 0.3 (Level set)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$, and let c be a real number. The **level set** [7] of f at c is the preimage of the single value c under f:

$$f^{-1}(\{c\}) \stackrel{\text{def}}{=} f^{-1}(\{c\}).$$

Theorem 0.1 (A level set is the set of points where the function takes the value)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$, and let c be a real number. The level set of f at c is the set of points where f takes the value c:*

$$f^{-1}(\{c\}) = \{\underset{\sim}{x} \in \mathbb{R}^p : f(\underset{\sim}{x}) = c\}.$$

Proof

Proof. Unfold the preimage:

$$\begin{aligned} f^{-1}(\{c\}) &= f^{-1}(\{c\}) && \text{(Definition 0.3)} \\ &= \{\underset{\sim}{x} \in \mathbb{R}^p : f(\underset{\sim}{x}) \in \{c\}\} && \text{(definition of the preimage)} \\ &= \{\underset{\sim}{x} \in \mathbb{R}^p : f(\underset{\sim}{x}) = c\}. && (\{c\} \text{ has the one element } c) \end{aligned}$$

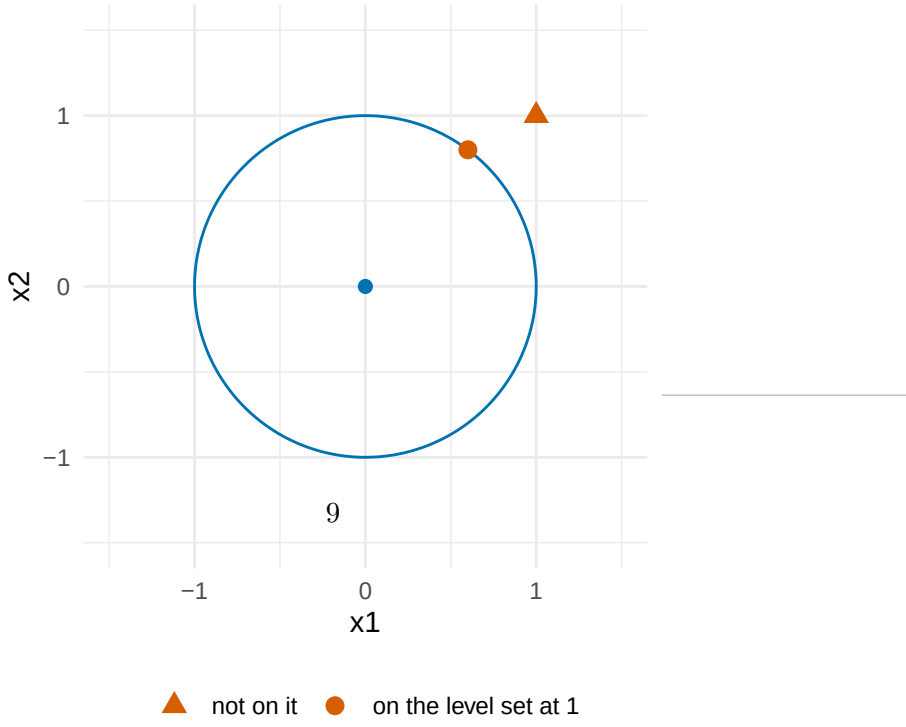
□

Example 0.3 (Level sets of $x_1^2 + x_2^2$)**.** Let $f(\underset{\sim}{x}) = x_1^2 + x_2^2$ for $\underset{\sim}{x} = (x_1, x_2)^\top$. By Theorem 0.1, each level set of f holds the points where f takes that value:

- The level set $f^{-1}(\{1\})$ is the circle of radius 1 around $\underset{\sim}{0}$. For example, $(0.6, 0.8)^\top$ is on it, since $0.36 + 0.64 = 1$, and $(1,1)^\top$ is not, since $\underset{\sim}{1} + 1 = 2$.
- The level set $f^{-1}(\{0\})$ is the single point $\underset{\sim}{0}$.
- The level set $f^{-1}(\{-1\})$ is empty, since $x_1^2 + x_2^2 \geq 0$.

Figure 2 draws the first two.

```r
ls_angle_ls <- seq(0, 2 * pi, length.out = 201)
ls_x1_ls <- cos(ls_angle_ls)
ls_x2_ls <- sin(ls_angle_ls)
ls_circle_ls <- tibble::tibble(x1 = ls_x1_ls, x2 = ls_x2_ls)
ls_origin_ls <- tibble::tibble(x1 = 0, x2 = 0)
ls_kinds_ls <- c("on the level set at 1", "not on it")
ls_points_x1_ls <- c(0.6, 1)
ls_points_x2_ls <- c(0.8, 1)
ls_points_ls <- tibble::tibble(
  x1 = ls_points_x1_ls,
  x2 = ls_points_x2_ls,
  kind = ls_kinds_ls
)
ls_points_aes_ls <- ggplot2::aes(shape = kind)
ls_shapes_ls <- c("on the level set at 1" = 16, "not on it" = 17)
ls_limits_ls <- c(-1.5, 1.5)
ls_circle_ls |>
  ggplot2::ggplot() +
  ggplot2::aes(x = x1, y = x2) +
  ggplot2::geom_path(colour = "#0072B2") +
  ggplot2::geom_point(data = ls_origin_ls, colour = "#0072B2", size = 2) +
  ggplot2::geom_point(
    data = ls_points_ls, mapping = ls_points_aes_ls,
    colour = "#D55E00", size = 3
  ) +
  ggplot2::scale_shape_manual(values = ls_shapes_ls) +
  ggplot2::coord_equal(xlim = ls_limits_ls, ylim = ls_limits_ls) +
  ggplot2::labs(shape = NULL) +
  ggplot2::theme_minimal() +
  ggplot2::theme(legend.position = "bottom")
```



Remark (Level sets in the plane)

Remark 0.2 (Level sets in the plane). For $p = 2$, a level set is often a curve in the plane, such as the circle $f^{-1}(\{1\})$ in Example 0.3.

Definition 0.4 (Level curve)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ and let c be a real number. The **level curve** of f at c is the level set of f at c when it is a curve in the plane. It is the special case $p = 2$ of a level set. A contour line is a level curve drawn on a plot of f.

Example 0.4 (A level curve of $x_1^2 + x_2^2$)**.** Let $f(\underset{\sim}{x}) = x_1^2 + x_2^2$ for $\underset{\sim}{x} = (x_1, x_2)^\top$. The level curve at 4 is the circle of radius 2 around $\underset{\sim}{0}$.

Theorem 0.2 (Level sets of a linear function)**.** *Let $h(\underset{\sim}{x}) = \beta \cdot \underset{\sim}{x} + \beta_0$ with $\beta \neq 0$. Every level set $h^{-1}(\{c\})$ is a hyperplane with normal vector $\underset{\sim}{\beta}$, parallel to the level set $h^{-1}(\{0\})$. Its dimension is $p - 1$: a point for $p = 1$, a line for $p = 2$, and a two-dimensional plane for $p = 3$.*

Proof

Proof. For any c,

$$\begin{aligned} h^{-1}(\{c\}) &= \{\underset{\sim}{x} : \underset{\sim}{\beta} \cdot \underset{\sim}{x} + \beta_0 = c\} \\ &= \{\underset{\sim}{x} : \underset{\sim}{\beta} \cdot \underset{\sim}{x} + (\beta_0 - c) = 0\}. \end{aligned}$$

This is the hyperplane with normal vector $\underset{\sim}{\beta}$ and offset $\beta_0 - c$. Changing c changes only the offset, so the level sets share the normal vector and are parallel. A hyperplane in $\mathbb{R}^p$ has dimension $p - 1$ by the hyperplane theorem, which gives the point, line and plane cases for $p = 1, 2, 3$. □

Example 0.5 (Parallel level lines)**.** Let $h(\underset{\sim}{x}) = 2x_1 + x_2 + 1$, so $\underset{\sim}{\beta} = (2, 1)^\top$ and $\beta_0 = 1$. The level set at c is the line $x_2 = c - 1 - 2x_1$. For $c = -1, 0, 1$ the lines are $x_2 = -2 - 2x_1$, $x_2 = -1 - 2x_1$ and $x_2 = -2x_1$. They all have slope -2 and differ only in the intercept, so they are parallel.

Theorem 0.3 (Stepping along the normal vector)**.** *Let $h(\underset{\sim}{x}) = \underset{\sim}{\beta} \cdot \underset{\sim}{x} + \beta_0$ with $\underset{\sim}{\beta} \neq \underset{\sim}{0}$. For every $\underset{\sim}{x} \in \mathbb{R}^p$ and $t \in \mathbb{R}$,*

$$h(\underset{\sim}{x} + t\underset{\sim}{\beta}) = h(\underset{\sim}{x}) + t\|\underset{\sim}{\beta}\|^2.$$

So a step along $\underset{\sim}{\beta}$ with $t > 0$ increases h.

Proof

Proof. By linearity of the dot product,

$$\begin{aligned} h(\underset{\sim}{x} + t\underset{\sim}{\beta}) &= \underset{\sim}{\beta} \cdot \underset{\sim}{x} + t\underset{\sim}{\beta} + \beta_0 \\ &= \underset{\sim}{\beta} \cdot \underset{\sim}{x} + t\underset{\sim}{\beta} \cdot \underset{\sim}{\beta} + \beta_0 \\ &= h(\underset{\sim}{x}) + t\|\underset{\sim}{\beta}\|^2. \end{aligned}$$

□

Example 0.6 (A step along the normal vector)**.** Let $h(\underset{\sim}{x}) = 2x_1 + x_2 + 1$, so $\beta = (2, 1)^\top$ and $\|\beta\|^2 = 5$. At $\underset{\sim}{x} = \underset{\sim}{0}$, $h(\underset{\sim}{0}) = 1$. A step of $t = 1$ along $\underset{\sim}{\beta}$ lands at $(2, \widetilde{1})^\top$, where $h\big((2, \widetilde{1})^\top\big) = 4 + \widetilde{1} + 1 = \widetilde{6} = \widetilde{1} + 5$.

Theorem 0.4 (The normal vector is the steepest direction)**.** *Let $h(\underset{\sim}{x}) = \beta \cdot \underset{\sim}{x} + \beta_0$ with $\beta \neq \underset{\sim}{0}$. Among unit vectors $\underset{\sim}{u}$ ($\|\underset{\sim}{u}\| = 1$), the change $h(\underset{\sim}{x} + t\underset{\sim}{u}) - h(\underset{\sim}{x})$ with $\widetilde{t} > 0$ is largest for $\underset{\sim}{u} = \underset{\sim}{\beta}/\|\underset{\sim}{\beta}\|$.*

Proof

Proof. For a unit vector $\underset{\sim}{u}$,

$$h(\underset{\sim}{x} + t\underset{\sim}{u}) - h(\underset{\sim}{x}) = t\underset{\sim}{\beta} \cdot \underset{\sim}{u}.$$

Since $t > 0$, maximizing this change is the same as maximizing $\beta \cdot \underset{\sim}{u}$. By the Cauchy-Schwarz inequality, $\beta \cdot \underset{\sim}{u} \leq \|\beta\|\|\underset{\sim}{u}\| = \|\beta\|$. Equality holds exactly when $\underset{\sim}{u} = \alpha\underset{\sim}{\beta}$ for some $\alpha > 0$. The condition $\|\underset{\sim}{u}\| = \widetilde{1}$ then forces $\alpha\|\underset{\sim}{\beta}\| = 1$, so

$$\underset{\sim}{u} = \underset{\sim}{\beta}/\|\underset{\sim}{\beta}\|.$$

□

Example 0.7 (Two unit steps compared)**.** Let $h(\underset{\sim}{x}) = 2x_1 + x_2 + 1$, so $\beta = (2, 1)^\top$ and $\|\beta\| = \sqrt{5}$. A step of size 1 along the unit vector $\beta/\sqrt{5}$ raises h by $\underset{\sim}{\beta} \cdot \underset{\sim}{\beta}/\sqrt{5} = \sqrt{5} \approx 2.24$. A step of size 1 along $(1, 0)^\top$ raises it by only $\underset{\sim}{\beta} \cdot (1, 0)^\top = 2$.

The course website's version of this page has an interactive plot here.

The course website's version of this page has an interactive 3D plot here.

Exercise 0.1 (Take a gradient by hand)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be

$$f(\underset{\sim}{w}) = 3w_1w_2 + w_2^3 - 5w_1$$

1. Compute $\partial f/\partial w_1$ (Definition 0.1), treating w_2 as a constant.
2. Compute $\partial f/\partial w_2$, treating w_1 as a constant.
3. Assemble $\nabla_{\underset{\sim}{w}} f$ and evaluate it at $\underset{\sim}{w} = \begin{bmatrix} 2 & -1 \end{bmatrix}^\top$.
4. Check the shape of your answer against Table 1.

Solution

Solution 0.1. **1.** Differentiate with respect to w_1, holding w_2 fixed. The first term is the constant $3w_2$ times w_1, so it contributes $3w_2$. The second term has no w_1 in it, so it contributes 0. The third term is -5 times w_1, so it contributes -5:

$$\begin{aligned} \frac{\partial f}{\partial w_1} &= 3w_2 + 0 - 5 \\ &= 3w_2 - 5 \end{aligned}$$

2. Now with respect to w_2, holding w_1 fixed. The first term is the constant $3w_1$ times w_2, contributing $3w_1$. The second term contributes $3w_2^2$. The third has no w_2 in it, contributing 0:

$$\begin{aligned} \frac{\partial f}{\partial w_2} &= 3w_1 + 3w_2^2 + 0 \\ &= 3w_1 + 3w_2^2 \end{aligned}$$

3. Stack the two, in the order the coordinates are numbered:

$$\nabla_{\underset{\sim}{w}} f(\underset{\sim}{w}) = \begin{bmatrix} 3w_2 - 5 \\ 3w_1 + 3w_2^2 \end{bmatrix}$$

The gradient is a vector-valued *function* of $\underset{\sim}{w}$, not a single vector. At $\underset{\sim}{w} = \begin{bmatrix} 2 & -1 \end{bmatrix}^\top$,

$$\begin{aligned} \nabla_{\underset{\sim}{w}} f &= \begin{bmatrix} 3(-1) - 5 \\ 3(2) + 3(-1)^2 \end{bmatrix} \\ &= \begin{bmatrix} -3 - 5 \\ 6 + 3 \end{bmatrix} \\ &= \begin{bmatrix} -8 \\ 9 \end{bmatrix} \end{aligned}$$

4. The input was a vector in $\mathbb{R}^2$ and so is the answer, as Table 1 requires. Note the signs: from this point, increasing w_1 *decreases* f while increasing w_2 increases it, so the uphill direction is neither axis.

[1] N. Fieller, *Basics of matrix algebra for statistics with R*. Chapman and Hall/CRC, 2016. doi: 10.1201/9781315370200.

[2] Wikipedia contributors, "Partial derivative — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Partial_derivative

[3] S. J. Miller, *The probability lifesaver: Calculus review problems*. 2016. Available: https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/index.htm#:~:text=http%3A//web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/supplementalchap_calcreview.pdf

[4] Wikipedia contributors, "Gradient — Wikipedia, the free encyclopedia." Available: https://en.wikipedia.org/wiki/Gradient

[5] B. Hutchinson, "DATA 471/571 (machine learning) and CSCI 481/581 (deep learning) video lectures." Western Washington University. Accessed: Sep. 28, 2026. [Online]. Available: https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/

[6] Wikipedia contributors, "Linear approximation — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Linear_approximation

[7] Wikipedia contributors, "Level set — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Level_set