Critical Points and Optimality Conditions

Last modified: 2026-10-31 15:22:31 (PDT)

Critical points and optimization

Definition 0.1 (Flat point (stationary point))**.** Let f be differentiable at c. If $f'(c) = 0$, then c is a **flat point** of f (also called a **stationary point**) [1].

Example 0.1 (Flat points, and a point that is not one)**.**

- $f(w) = w^3$ has $f'(w) = 3w^2$, which is 0 at $w = 0$, so 0 is a flat point. Yet $f(0) = 0$ is neither the minimum nor the maximum of the values f takes on any open interval around 0: $f(w) < 0$ for $w < 0$, and $f(w) > 0$ for $w > 0$.
- $h(w) = w^2$ has $h'(1) = 2 \neq 0$, so 1 is not a flat point.

Exercise 0.1 (Differentiate a quadratic)**.** A flat point of a function is where its derivative is zero. The next exercises use the parabola $f(w) = w^2 - 4w + 7$ to find where f is flat and to compare a linear approximation with the exact value.
Differentiate f.

Solution

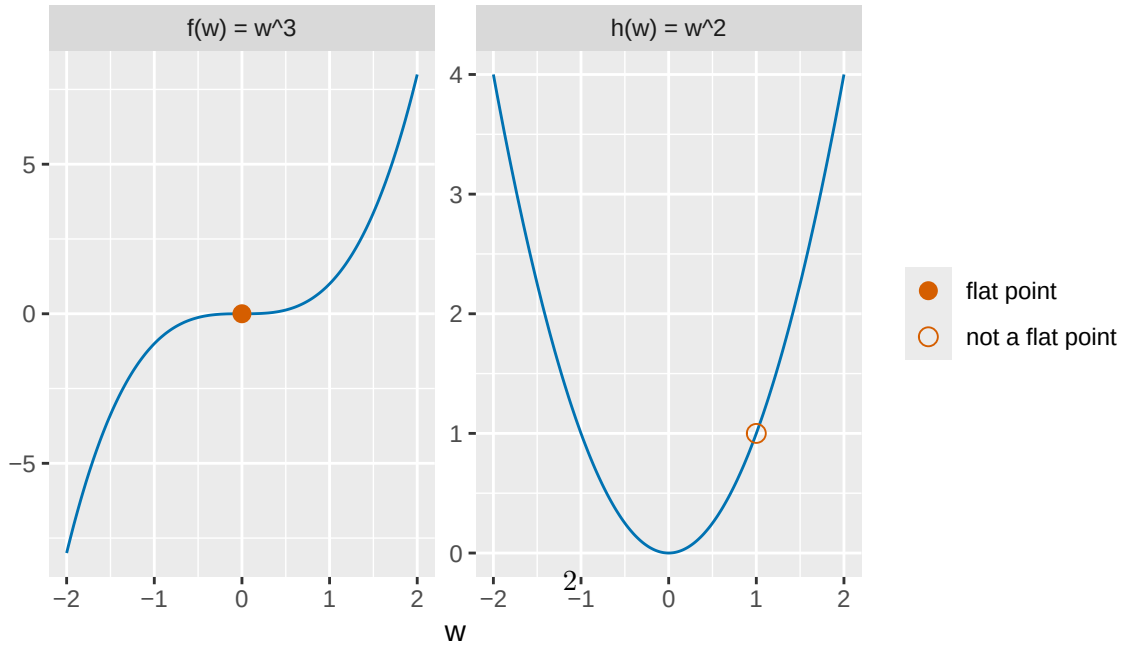
Solution 0.1. Differentiating each term:

$$\frac{df}{dw} = 2w - 4$$

Exercise 0.2 (Find the flat point)**.** In the setting of Exercise 0.1, find the flat point w of f.

R

```r
fp_w_ff <- seq(-2, 2, length.out = 201)
fp_cube_ff <- tibble::tibble(
  w = fp_w_ff, value = fp_w_ff^3, panel = "f(w) = w^3"
)
fp_square_ff <- tibble::tibble(
  w = fp_w_ff, value = fp_w_ff^2, panel = "h(w) = w^2"
)
fp_curves_ff <- dplyr::bind_rows(fp_cube_ff, fp_square_ff)
fp_marks_w_ff <- c(0, 1)
fp_marks_panel_ff <- c("f(w) = w^3", "h(w) = w^2")
fp_marks_kind_ff <- c("flat point", "not a flat point")
fp_marks_ff <- tibble::tibble(
  w = fp_marks_w_ff, value = fp_marks_w_ff,
  panel = fp_marks_panel_ff, kind = fp_marks_kind_ff
)
fp_aes_ff <- ggplot2::aes(w, value)
fp_marks_aes_ff <- ggplot2::aes(shape = kind)
fp_shapes_ff <- c("flat point" = 16, "not a flat point" = 1)
ggplot2::ggplot(fp_curves_ff, fp_aes_ff) +
  ggplot2::geom_line(colour = "#0072B2") +
  ggplot2::geom_point(
    data = fp_marks_ff, mapping = fp_marks_aes_ff,
    colour = "#D55E00", size = 3
  ) +
  ggplot2::scale_shape_manual(values = fp_shapes_ff) +
  ggplot2::facet_wrap(~panel, scales = "free_y") +
  ggplot2::labs(x = "w", y = NULL, shape = NULL)
```



Python

```python
import matplotlib.pyplot as plt
import numpy as np
```

Solution

Solution 0.2. The derivative is zero where $2w - 4 = 0$, which gives $w = 2$.

Exercise 0.3 (Classify the flat point)**.** In the setting of Exercise 0.1, say whether $f(w)$ at the flat point is the minimum or the maximum of the values of f.

Solution

Solution 0.3. The flat point is $w = 2$, from Solution 0.2. Evaluating f there gives

$$\begin{aligned} f(2) &= 4 - 8 + 7 \\ &= 3. \end{aligned}$$

This is the minimum of the values of f. Completing the square, $f(w) = (w-2)^2 + 3$, since

$$\begin{aligned} (w-2)^2 + 3 &= w^2 - 4w + 4 + 3 \\ &= w^2 - 4w + 7, \end{aligned}$$

and $(w-2)^2 \geq 0$, so $f(w) \geq 3 = f(2)$ for every w. Equivalently, the derivative is negative below $w = 2$ and positive above it, so the function falls into that point and rises out of it.

Exercise 0.4 (Slope at $w = 1$)**.** In the setting of Exercise 0.1, evaluate the derivative at $w = 1$.

Solution

Solution 0.4. At $w = 1$ the derivative is $2(1) - 4 = -2$, so f is falling there.

Exercise 0.5 (Predict a nearby value)**.** In the setting of Exercise 0.1, use the linear approximation and the derivative at $w = 1$ from Exercise 0.4 to predict $f(1.01)$.

Solution

Solution 0.5. With $\varepsilon = 0.01$, the linear approximation predicts a change of $(0.01)(-2) = -0.02$, from

$$\begin{aligned} f(1) &= 1 - 4 + 7 \\ &= 4 \end{aligned}$$

to 3.98.

Exercise 0.6 (Exact value)**.** In the setting of Exercise 0.1, find the exact value of $f(1.01)$.

Solution

Solution 0.6. The exact value is

$$\begin{aligned} f(1.01) &= (1.01)^2 - 4(1.01) + 7 \\ &= 1.0201 - 4.04 + 7 \\ &= 3.9801 \end{aligned}$$

a change of -0.0199 from $f(1) = 4$.

Exercise 0.7 (Check the prediction)**.** Compare the prediction in Exercise 0.5 with the exact value of $f(1.01)$ from Exercise 0.6.

Solution

Solution 0.7. The prediction 3.98 (Solution 0.5) and the exact value 3.9801 (Solution 0.6) differ by 0.0001, which is ε^2. For this f, the difference between $f(w+\varepsilon)$ and its linear approximation is, at any w, exactly

$$f(w+\varepsilon) - [f(w) + f'(w)\,\varepsilon] = \varepsilon^2,$$

so halving the step quarters that difference.
A step of gradient descent with step length s changes w by $\varepsilon = -sf'(w)$. The linear approximation predicts the change $-s\,f'(w)^2$ in f, and the difference between $f(w+\varepsilon)$ and its linear approximation is $\varepsilon^2 = s^2 f'(w)^2$. The actual change in f is therefore $-s\,f'(w)^2\,(1-s)$. This change is negative exactly when $0 < s < 1$, for $f'(w) \neq 0$, so f decreases only for those steps. For this f, the difference is smaller than the predicted decrease only when $s < 1$: the step must be short enough.

Definition 0.2 (Critical point)**.** Let f be a function defined on an open interval containing c. The point c is a **critical point** of f if either $f'(c) = 0$ or f is not differentiable at c (Definition 6 in Derivatives and Taylor Series).

Example 0.2 (Critical points that are and are not flat points)**.**

- $g(w) = |w|$ has no derivative at $w = 0$: its one-sided derivatives there are 1 and -1 (worked example). So the difference quotient $\frac{|h|-0}{h}$ has no limit as $h \to 0$, because the limits from the left and from the right (one-sided limits) differ. So 0 is a critical point of g but not a flat point.

- $f(w) = w^3$ has

$$\begin{aligned} f'(0) &= 3 \cdot 0^2 \\ &= 0, \end{aligned}$$

so 0 is a flat point of f, and hence a critical point.

- $h(w) = w^2$ is differentiable everywhere, with $h'(w) = 2w$, which is 0 only at $w = 0$. So 0 is the only critical point of h. For example, $h'(1) = 2 \neq 0$, so 1 is not one.

Theorem 0.1 (Every flat point is a critical point)**.** *If c is a flat point of f (Definition 0.1), then c is a critical point of f (Definition 0.2).*

Proof

Proof. A flat point c has f differentiable at c and $f'(c) = 0$. The condition "$f'(c) = 0$" is the first alternative in Definition 0.2, so c is a critical point. □

Example 0.3 (The flat points of a cubic, and a critical point that is not flat)**.** For $f(w) = w^3 - 3w$, the derivative is $f'(w) = 3w^2 - 3$. The code differentiates f with `D()` and evaluates f' at several points. For $g(w) = |w|$, it computes the difference quotients at 0 from both sides.

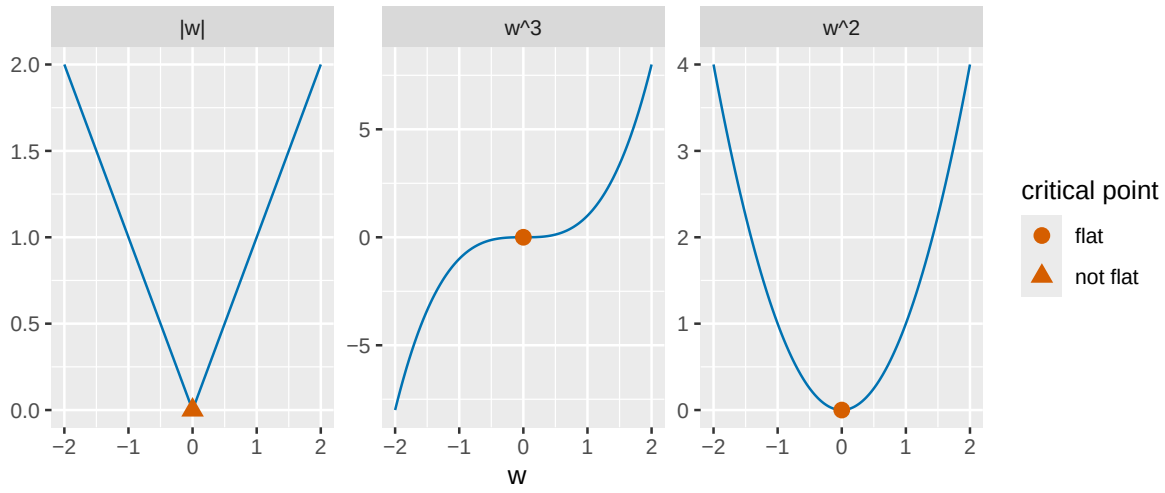
```
f_prime <- D(quote(w^3 - 3 * w), "w")
points <- c(-2, -1, 0, 1, 2)
derivs <- data.frame(w = points, f_prime = eval(f_prime, list(w = points)))
derivs
#>    w f_prime
#> 1 -2       9
#> 2 -1       0
#> 3  0      -3
#> 4  1       0
#> 5  2       9
flat <- derivs$w[derivs$f_prime == 0]

h <- c(-1e-3, 1e-3)
quotients <- abs(h) / h
quotients
#> [1] -1  1
```

The derivative is 0 at 2 of the 5 points, namely $w = -1$ and $w = 1$. Both points are flat

R

```r
cp_w_fc <- seq(-2, 2, length.out = 201)
cp_abs_value_fc <- abs(cp_w_fc)
cp_abs_fc <- tibble::tibble(w = cp_w_fc, value = cp_abs_value_fc, panel = "|w|")
cp_cube_fc <- tibble::tibble(w = cp_w_fc, value = cp_w_fc^3, panel = "w^3")
cp_square_fc <- tibble::tibble(w = cp_w_fc, value = cp_w_fc^2, panel = "w^2")
cp_curves_fc <- dplyr::bind_rows(cp_abs_fc, cp_cube_fc, cp_square_fc)
cp_levels_fc <- c("|w|", "w^3", "w^2")
cp_curves_fc$panel <- factor(cp_curves_fc$panel, levels = cp_levels_fc)
cp_kinds_fc <- c("not flat", "flat", "flat")
cp_marks_fc <- tibble::tibble(
  w = 0, value = 0, panel = cp_levels_fc, kind = cp_kinds_fc
)
cp_marks_fc$panel <- factor(cp_marks_fc$panel, levels = cp_levels_fc)
cp_aes_fc <- ggplot2::aes(w, value)
cp_marks_aes_fc <- ggplot2::aes(shape = kind)
cp_shapes_fc <- c("flat" = 16, "not flat" = 17)
ggplot2::ggplot(cp_curves_fc, cp_aes_fc) +
  ggplot2::geom_line(colour = "#0072B2") +
  ggplot2::geom_point(
    data = cp_marks_fc, mapping = cp_marks_aes_fc,
    colour = "#D55E00", size = 3
  ) +
  ggplot2::scale_shape_manual(values = cp_shapes_fc) +
  ggplot2::facet_wrap(~panel, scales = "free_y") +
  ggplot2::labs(x = "w", y = NULL, shape = "critical point")
```



Python

```python
import matplotlib.pyplot as plt
import numpy as np

cp_w_fc = np.linspace(-2, 2, 201)
cp_curves_fc = [("|w|", np.abs(cp_w_fc), "^"),
                ("w^3", cp_w_fc**3, "o"),
                ("w^2", cp_w_fc**2, "o")]
fig, axes = plt.subplots(1, 3, figsize=(7, 3), layout="constrained")
```

points of f, since the derivative is 0 there, so by Theorem 0.1 they are critical points. The difference quotients of g at 0 are -1 from the left and 1 from the right, so g has no derivative at 0. Thus 0 is a critical point of g (Definition 0.2), but not a flat point.

Theorem 0.2 (Second derivative test)**.** *Let c be a flat point of f, where the second derivative $f''(c)$ is defined.*

- *If $f''(c) > 0$, then $f(x) > f(c)$ for every $x \neq c$ in some open interval around c.*
- *If $f''(c) < 0$, then $f(x) < f(c)$ for every $x \neq c$ in some open interval around c.*

Proof

Proof. Suppose $f''(c) > 0$. Because $f'(c) = 0$, the derivative of f' at c is

$$f''(c) = \lim_{x \to c} \frac{f'(x)}{x - c}.$$

A limit that is positive forces the quotient $f'(x)/(x - c)$ to be positive for x near c. So $f'(x)$ has the same sign as $x - c$ on some open interval around c. Hence $f' < 0$ to the left of c and $f' > 0$ to the right of c, so f is decreasing to the left of c and increasing to the right of c. Therefore $f(x) > f(c)$ for every $x \neq c$ in that open interval.
For $f''(c) < 0$, apply the same reasoning to $-f$, since $(-f)''(c) = -f''(c) > 0$. Then $-f(x) > -f(c)$ near c, which says $f(x) < f(c)$ near c. □

Example 0.4 (Second derivative test for powers of x)**.** The function $f(x) = x^2$ has the flat point $c = 0$ and second derivative $f''(0) = 2 > 0$, so 0 is a local minimizer by Theorem 0.2.
The functions x^3 and x^4 both have the flat point 0 and second derivative 0 there, so the test does not decide. Yet x^4 is smallest at 0 among nearby points, while x^3 is smaller to the left of 0 and larger to the right of 0.

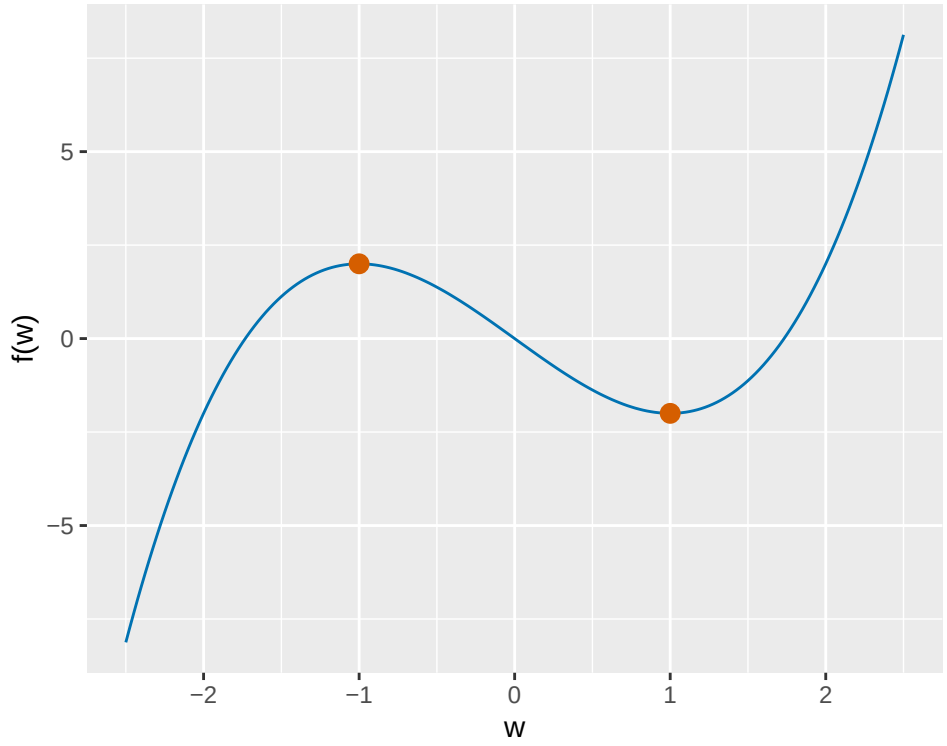
Remark (The second derivative test can fail to decide)

Remark 0.1 (The second derivative test can fail to decide). If $f''(c) = 0$ at a flat point c, the second derivative test does not decide: f may be smallest at c among nearby points, largest there, or neither (Example 0.4).

Definition 0.3 (Inflection point)**.** An **inflection point** of a function f is a point c such that, on some open interval around c, the second derivative f'' is positive on one side of c and negative on the other.

R

```r
fpc_w_fk <- seq(-2.5, 2.5, length.out = 201)
fpc_curve_fk <- tibble::tibble(w = fpc_w_fk, value = fpc_w_fk^3 - 3 * fpc_w_fk)
fpc_flat_w_fk <- c(-1, 1)
fpc_flat_value_fk <- fpc_flat_w_fk^3 - 3 * fpc_flat_w_fk
fpc_flat_fk <- tibble::tibble(w = fpc_flat_w_fk, value = fpc_flat_value_fk)
fpc_aes_fk <- ggplot2::aes(w, value)
ggplot2::ggplot(fpc_curve_fk, fpc_aes_fk) +
  ggplot2::geom_line(colour = "#0072B2") +
  ggplot2::geom_point(data = fpc_flat_fk, colour = "#D55E00", size = 3) +
  ggplot2::labs(x = "w", y = "f(w)")
```



Python

```python
import matplotlib.pyplot as plt
import numpy as np

fpc_w_fk = np.linspace(-2.5, 2.5, 201)
fpc_flat_fk = np.array([-1, 1])
fig, ax = plt.subplots(figsize=(5, 4), layout="constrained")
ax.plot(fpc_w_fk, fpc_w_fk**3 - 3 * fpc_w_fk, color="#0072B2")
ax.plot(fpc_flat_fk, fpc_flat_fk**3 - 3 * fpc_flat_fk, "o",
        color="#D55E00", markersize=8)
ax.set_xlabel("w")
ax.set_ylabel("f(w)")
plt.show()
```




Remark (Inflection points change the curvature)

Remark 0.2 (Inflection points change the curvature). At an inflection point, the graph changes between curving up and curving down.

Example 0.5 (An inflection point of x^3)**.** For $f(x) = x^3$, the second derivative is $f''(x) = 6x$, which is negative for $x < 0$ and positive for $x > 0$. So 0 is an inflection point of f.

Exercise 0.8 (Maximizing a gamma-family kernel)**.** Adapted from Miller [2], Question 1.1.10.
The gamma distribution with shape $a > 0$ has the kernel $x^{a-1}e^{-x}$ on $[0, \infty)$. A kernel has the same mode as its density (why), so maximizing it finds the mode without computing the normalizing constant.
For shape $a = 5$, the kernel is

$$f(x) = x^4 e^{-x}.$$

Find the maximum of f on $[0, \infty)$.

Solution

Solution 0.8. Differentiate using the product rule:

$$\begin{aligned} f'(x) &= 4x^3 e^{-x} - x^4 e^{-x} \\ &= x^3 e^{-x}(4 - x) \end{aligned}$$

On $[0, \infty)$, $e^{-x} > 0$, so $f'(x) = 0$ only at $x = 0$ and $x = 4$.
Evaluate f at the boundary and critical points:

- At $x = 0$: $f(0) = 0$.
- At $x = 4$:

$$\begin{aligned} f(4) &= 4^4 e^{-4} \\ &= 256 e^{-4} \\ &\approx 4.6888. \end{aligned}$$

- As $x \to \infty$: exponential decay dominates polynomial growth, so $\lim_{x\to\infty} x^4 e^{-x} = 0$.

Since $f(x) \geq 0$ for all $x \geq 0$ and $f(4) > 0$, the global maximum occurs at $x = 4$, with maximum value $256/e^4$.

Exercise 0.9 (Classifying critical points with the second derivative test)**.** Adapted from Miller [2], Question 1.1.11.
An optimization algorithm such as Newton's method or gradient descent can stop at a critical point, which may be a local maximum, a local minimum, or neither. The second derivative test tells these apart.
Consider the cubic

$$f(x) = 4x^3 - 3x^2.$$

Find its critical points, and decide whether each is a local maximum, a local minimum, or an inflection point.

Solution

Solution 0.9. Compute the first derivative:

$$\begin{aligned} f'(x) &= 12x^2 - 6x \\ &= 6x(2x - 1) \end{aligned}$$

Setting $f'(x) = 0$ gives critical points at $x = 0$ and $x = 1/2$.
Compute the second derivative:

$$f''(x) = 24x - 6$$

Apply the second derivative test at each critical point:

- At $x = 0$: $f''(0) = -6 < 0$, so $x = 0$ is a strict local maximum.
- At $x = 1/2$:

$$\begin{aligned} f''(1/2) &= 24(1/2) - 6 \\ &= 6 \\ &> 0, \end{aligned}$$

  so $x = 1/2$ is a strict local minimum.

Inflection points occur where the second derivative changes sign ($f''(x) = 0$ at $x = 1/4$). Because f'' is non-zero at both critical points, neither is an inflection point.

Optimality conditions in several variables

Source

This section is adapted from the multivariate calculus and optimality-condition parts of Zhou [3], used under the MIT License (see the license text in Linear Algebra). The source states the second-order Taylor approximation and the optimality conditions without proof; here they are proved from the one-variable Taylor theorem, and the approximation is made exact by evaluating the Hessian at a point between $\underset{\sim}{z}$ and $\underset{\sim}{z}+\underset{\sim}{h}$. The source calls points with zero gradient critical points; here they are stationary points, because Definition 0.2 gives "critical point" a wider meaning. The source states these results for differentiable functions without saying how smooth; here the first-order results assume continuous first partial derivatives, and the second-order results continuous second partial derivatives. These parts of the source are not part of this section:

- its remark that maximizing f is minimizing $-f$, and its classification of constraints
- its figure of local, strict local and global minima
- its one-variable Taylor example and its plots
- matrix calculus
- convexity
- Lagrange multipliers
- Newton's method and gradient descent

A local minimizer of a differentiable function on an open interval is a flat point (Definition 0.1), and, at a flat point, a positive second derivative guarantees a strict local minimum. This section extends both facts to a function f of a $p \times 1$ vector $\underset{\sim}{x}$: the gradient $\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x})$ (gradient) takes the place of the first derivative, and a $p \times p$ matrix of second partial derivatives takes the place of the second; both facts are proved in this section, for every $p \geq 1$. Throughout, $\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z})$ means the gradient evaluated at $\underset{\sim}{x} = \underset{\sim}{z}$.

Theorem 0.3 (A minimizer has derivative zero)**.** *Let g be a real-valued function on an open interval containing c, and let g be differentiable at c. If c is a local minimizer of g, then $g'(c) = 0$. [4]*

Proof

Proof. Take $\delta > 0$ as in the definition of a local minimizer, shrunk if needed so that $(c-\delta, c+\delta)$ lies inside the interval; then $g(c) \leq g(c+h)$ whenever $|h| < \delta$. For $0 < h < \delta$, the difference quotient is

$$\frac{g(c+h) - g(c)}{h} \geq 0,$$

because its numerator is at least 0 and its denominator is positive. As $h \to 0$ from the right, these quotients tend to $g'(c)$ (the two-sided limit exists, so the one-sided limit equals it), and a limit of numbers that are all at least 0 is at least 0; so $g'(c) \geq 0$. For $-\delta < h < 0$, the numerator is still at least 0 but the denominator is negative, so the quotient is at most 0, and in the same way, with $h \to 0$ from the left, $g'(c) \leq 0$. Together, $g'(c) = 0$. □

Example 0.6 (Applying the test, and its limits)**.**

- $g(x) = (x-2)^2$ has its minimizer at $c = 2$, and

$$\begin{aligned} g'(2) &= 2(2-2) \\ &= 0. \end{aligned}$$

- The converse fails: $g(x) = x^3$ has

$$\begin{aligned} g'(0) &= 3 \cdot 0^2 \\ &= 0, \end{aligned}$$

  but 0 is not a local minimizer, since

$$\begin{aligned} g(-h) &= -h^3 \\ &< 0 \\ &= g(0) \end{aligned}$$

  for every $h > 0$.

- Differentiability is needed: $g(x) = |x|$ has its minimizer at 0, but the difference quotient $|h|/h$ is 1 for $h > 0$ and -1 for $h < 0$, so it has no limit and $g'(0)$ does not exist.

Definition 0.4 (Stationary point)**.** Let f be a scalar-valued function of a $p \times 1$ vector $\underset{\sim}{x}$. A point $\underset{\sim}{z}$ at which the gradient exists and

$$\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z}) = \underset{\sim}{0}_{p \times 1}$$

is a **stationary point** of f.

Example 0.7 (A stationary point, and a point that is not one)**.** Let $f(\underset{\sim}{x}) = x_1^2 + x_2^2 - 2x_1$. Its gradient is $(2x_1 - 2,\ 2x_2)^\top$, which is $\underset{\sim}{0}_{2\times 1}$ exactly when $x_1 = 1$ and $x_2 = 0$. So $(1, 0)^\top$ is its only stationary point, and $\underset{\sim}{0}$ is not one: the gradient there is $(-2, 0)^\top$.

Remark (Stationary points in one variable)

Remark 0.3 (Stationary points in one variable). For $p = 1$, a stationary point is a flat point (Definition 0.1).

Theorem 0.4 (First-order necessary condition)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$. If $\underset{\sim}{x}^*$ is a local minimizer of f and the gradient of f exists at $\underset{\sim}{x}^*$, then $\underset{\sim}{x}^*$ is a stationary point of f (Definition 0.4).*

Proof

Proof. Take $\delta > 0$ as in the definition of a local minimizer. Let $i \in \{1, \dots, p\}$, let $\underset{\sim}{e}_i$ be the vector with 1 in entry i and 0 elsewhere, and let $g_i(t) \overset{\text{def}}{=} f(\underset{\sim}{x}^* + t\,\underset{\sim}{e}_i)$ for $t \in \mathbb{R}$. For $|t| < \delta$,

$$\begin{aligned} \|(\underset{\sim}{x}^* + t\,\underset{\sim}{e}_i) - \underset{\sim}{x}^*\| &= |t|\,\|\underset{\sim}{e}_i\| \\ &= |t| \\ &< \delta, \end{aligned}$$

so

$$\begin{aligned} g_i(0) &= f(\underset{\sim}{x}^*) \\ &\le f(\underset{\sim}{x}^* + t\,\underset{\sim}{e}_i) \\ &= g_i(t)\,; \end{aligned}$$

0 is a local minimizer of g_i. Moving $\underset{\sim}{x}$ from $\underset{\sim}{x}^*$ along $\underset{\sim}{e}_i$ changes only x_i, so $g_i'(0)$ is the partial derivative $\frac{\partial}{\partial x_i} f(\underset{\sim}{x}^*)$, which exists because the gradient does. By Theorem 0.3,

$$\begin{aligned} \frac{\partial}{\partial x_i} f(\underset{\sim}{x}^*) &= g_i'(0) \\ &= 0. \end{aligned}$$

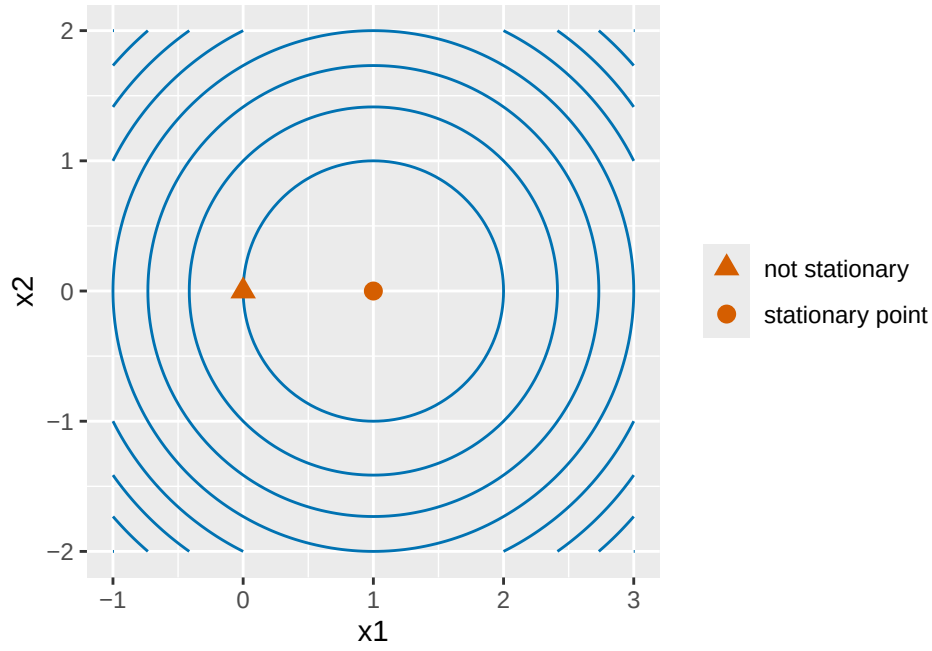
The equation $g_i'(0) = 0$ holds for every i, so every entry of the gradient at $\underset{\sim}{x}^*$ is 0 (vector derivative). □

Example 0.8 (Using the condition to locate, and to rule out, minimizers)**.**

- For $f(\underset{\sim}{x}) = x_1^2 + x_2^2 - 2x_1$ of Example 0.7, any local minimizer must be the stationary

R

```r
sp_x1_fs <- seq(-1, 3, length.out = 101)
sp_x2_fs <- seq(-2, 2, length.out = 101)
sp_grid_fs <- expand.grid(x1 = sp_x1_fs, x2 = sp_x2_fs)
sp_grid_fs$value <- sp_grid_fs$x1^2 + sp_grid_fs$x2^2 - 2 * sp_grid_fs$x1
sp_marks_x1_fs <- c(1, 0)
sp_marks_kind_fs <- c("stationary point", "not stationary")
sp_marks_fs <- tibble::tibble(
  x1 = sp_marks_x1_fs, x2 = 0, kind = sp_marks_kind_fs
)
sp_aes_fs <- ggplot2::aes(x1, x2, z = value)
sp_marks_aes_fs <- ggplot2::aes(x1, x2, shape = kind, z = NULL)
sp_shapes_fs <- c("stationary point" = 16, "not stationary" = 17)
ggplot2::ggplot(sp_grid_fs, sp_aes_fs) +
  ggplot2::geom_contour(colour = "#0072B2") +
  ggplot2::geom_point(
    data = sp_marks_fs, mapping = sp_marks_aes_fs,
    colour = "#D55E00", size = 3, inherit.aes = FALSE
  ) +
  ggplot2::scale_shape_manual(values = sp_shapes_fs) +
  ggplot2::coord_fixed() +
  ggplot2::labs(x = "x1", y = "x2", shape = NULL)
```



Python

```python
import matplotlib.pyplot as plt
import numpy as np

sp_x1_fs = np.linspace(-1, 3, 101)
sp_x2_fs = np.linspace(-2, 2, 101)
```

point $(1,0)^\top$. It is one: completing the square,

$$\begin{aligned} f(\underset{\sim}{x}) &= (x_1-1)^2 + x_2^2 - 1 \\ &\geq -1 \\ &= f(1,0). \end{aligned}$$

- For $f(\underset{\sim}{x}) = e^{2x_1+x_2} - x_1$ of Example, the second entry of the gradient is $e^{2x_1+x_2} > 0$, so f has no stationary point, and so no local minimizer.
- The converse fails: $f(\underset{\sim}{x}) = x_1^2 - x_2^2$ has gradient $(2x_1,\ -2x_2)^\top$, so $\underset{\sim}{0}$ is a stationary point, but

$$\begin{aligned} f(0,t) &= -t^2 \\ &< 0 \\ &= f(\underset{\sim}{0}) \end{aligned}$$

for every $t \neq 0$, so $\underset{\sim}{0}$ is not a local minimizer.

Example 0.9 (A stationary point of the sum of squared errors, on real data)**.** The Auto data record the horsepower and fuel economy (miles per gallon) of 392 car models sold between 1970 and 1982. They come from the StatLib library at Carnegie Mellon University and ship with the book *An Introduction to Statistical Learning*, in the Python package `ISLP` (`ISLP.load_data("Auto")`) and the R package `ISLR2` (`ISLR2::Auto`). This site has a copy at `data/auto.csv`, and both code tabs below read that file.
Let y_i be the miles per gallon and x_i the horsepower of car i. To predict miles per gallon from horsepower with a line, write $\beta = (\beta_0, \beta_1)^\top$ and let $\mathbf{X}$ be the matrix whose row i is $(1, x_i)$. The sum of squared errors is

$$f(\underset{\sim}{\beta}) = \|\underset{\sim}{y} - \mathbf{X}\underset{\sim}{\beta}\|^2,$$

a scalar-valued function of the 2×1 vector β. Its gradient is $-2\,\mathbf{X}^\top(y - \mathbf{X}\beta)$. By Theorem 0.4, a minimizer of f must make this gradient $\underset{\sim}{0}$, which is the equation $\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}^\top y$. The code below solves that equation and evaluates the gradient at the solution and at $\underset{\sim}{\beta} = \underset{\sim}{0}$.

Python

```python
import numpy as np
import pandas as pd

auto = pd.read_csv("data/auto.csv")
y = auto["mpg"].to_numpy(dtype=float)
X = np.column_stack([np.ones(len(auto)), auto["horsepower"].to_numpy(dtype=float)])

def grad(beta):
    return -2 * X.T @ (y - X @ beta)

beta_hat = np.linalg.solve(X.T @ X, X.T @ y)
grad_hat = grad(beta_hat)
grad_zero = grad(np.zeros(2))

pd.DataFrame(
    {"beta": [beta_hat, np.zeros(2)], "gradient": [grad_hat, grad_zero]},
    index=["solution", "zero vector"],
)
#>                                beta                gradient
#> solution     [39.935861, -0.157845]              [0.0, 0.0]
#> zero vector              [0.0, 0.0]  [-18381.6, -1737437.6]
```

A library fit gives the same coefficients:

```python
import numpy as np

# np.polyfit returns the slope first, so reverse to (intercept, slope).
beta_polyfit = np.polyfit(auto["horsepower"], auto["mpg"], deg=1)[::-1]
beta_polyfit
#> array([39.93586102, -0.15784473])
```

R

```r
auto <- read.csv("data/auto.csv")
y <- auto$mpg
X <- cbind(1, auto$horsepower)

grad <- function(beta) -2 * t(X) %*% (y - X %*% beta)

beta_hat <- solve(t(X) %*% X, t(X) %*% y)
grad_hat <- grad(beta_hat)
grad_zero <- grad(c(0, 0))

list(
  solution = list(beta = c(beta_hat), gradient = c(grad_hat)),
  zero_vector = list(beta = c(0, 0), gradient = c(grad_zero))
)
#> $solution
#> $solution$beta
#> [1] 39.9358610 -0.1578447
#>
#> $solution$gradient
#> [1] 9.471535e-12 6.662049e-11
#>
#>
#> $zero_vector
#> $zero_vector$beta
#> [1] 0 0
#>
#> $zero_vector$gradient
#> [1]   -18381.6 -1737437.6
```

A library fit gives the same coefficients:

```r
coef(lm(mpg ~ horsepower, data = auto))
#> (Intercept)  horsepower
#>  39.9358610  -0.1578447
```

The solution is an intercept of 39.9 miles per gallon and a slope of -0.158 miles per gallon per unit of horsepower. The largest entry of the gradient there is 6.7e-11, zero up to rounding error, against 1,737,438 at the zero vector. So the solution is a stationary point (Definition 0.4), and the zero vector is not. The library fits above agree with the solution of that equation.

Theorem 0.5 (Taylor's theorem with a second-order remainder)**.** *Let $a < b$, and let g be a real-valued function on $[a, b]$ whose derivative g' is continuous on $[a, b]$ and whose second derivative g'' exists at every point of (a, b). Then there is a $\tau \in (a, b)$ with [5]*

$$g(b) = g(a) + g'(a)\,(b - a) + \frac{1}{2}\,g''(\tau)\,(b - a)^2.$$

This theorem is the case $n = 2$ of [6], Theorem 5.15; its proof uses the mean value theorem, which these notes do not develop.

Example 0.10 (The remainder point for e^x)**.** Take $g(x) = e^x$ on $[0, 1]$, so

$$\begin{aligned} g' &= g'' \\ &= g. \end{aligned}$$

Theorem 0.5 says $e = 1 + 1 + \frac{1}{2}\,e^\tau$ for some $\tau \in (0, 1)$. Solving,

$$\begin{aligned} e^\tau &= 2(e - 2) \\ &\approx 2 \times 0.71828 \\ &= 1.43656, \end{aligned}$$

so $\tau = \log 1.43656 \approx 0.362$ (natural logarithm), which is in $(0, 1)$.

Theorem 0.6 (Second-order Taylor theorem in several variables)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ have first and second partial derivatives that are continuous on $\mathbb{R}^p$ (continuity), and let $\underset{\sim}{z}, \underset{\sim}{h} \in \mathbb{R}^p$. Then there is a $\tau \in (0, 1)$ with [5]*

$$f(\underset{\sim}{z} + \underset{\sim}{h}) = f(\underset{\sim}{z}) + \left(\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z})\right)^{\top} \underset{\sim}{h} + \frac{1}{2}\,\underset{\sim}{h}^{\top}\,\mathbf{H}_f(\underset{\sim}{z} + \tau \underset{\sim}{h})\,\underset{\sim}{h}.$$

Proof

Proof. Let $g(t) \overset{\text{def}}{=} f(\underset{\sim}{z} + t\underset{\sim}{h})$ for $t \in \mathbb{R}$, the values of f along the line through $\underset{\sim}{z}$ in the direction $\underset{\sim}{h}$. The inner function $\underset{\sim}{y}(t) = \underset{\sim}{z} + t\underset{\sim}{h}$ has entries $y_j = z_j + t h_j$, so $\frac{\partial}{\partial t} y_j = h_j$, and its derivative (vector-valued derivative) is the $1 \times p$ matrix $\underset{\sim}{h}^{\top}$. By the vector chain rule (vector chain rule, with its input t of length 1, its inner function $\underset{\sim}{y}(t) = \underset{\sim}{z} + t\underset{\sim}{h}$, and its outer function f),

$$
\begin{aligned}
g'(t) &= \underset{\sim}{h}^{\top} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z} + t\underset{\sim}{h}) && \text{(vector chain rule)} \\
&= \sum_{j=1}^{p} h_j \frac{\partial}{\partial x_j} f(\underset{\sim}{z} + t\underset{\sim}{h}). && \text{(matrix product)}
\end{aligned}
$$

The second step writes out the matrix product. g' is continuous (continuity): $t \mapsto \underset{\sim}{z} + t\underset{\sim}{h}$ is continuous, since $\|(\underset{\sim}{z} + s\underset{\sim}{h}) - (\underset{\sim}{z} + t\underset{\sim}{h})\| = |s - t| \, \|\underset{\sim}{h}\|$, each $\frac{\partial}{\partial x_j} f$ is continuous by assumption, and compositions, sums and constant multiples of continuous functions are continuous. Each $\frac{\partial}{\partial x_j} f$ has continuous partial derivatives, the second partial derivatives of f, so the same chain-rule computation applies to it: $\frac{\partial}{\partial t}\left[\frac{\partial}{\partial x_j} f(\underset{\sim}{z} + t\underset{\sim}{h})\right] = \sum_{i=1}^{p} h_i \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f(\underset{\sim}{z} + t\underset{\sim}{h})\right)$. So

$$
\begin{aligned}
g''(t) &= \sum_{j=1}^{p} h_j \sum_{i=1}^{p} h_i \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f(\underset{\sim}{z} + t\underset{\sim}{h})\right) && \text{(differentiate each term of } g'(t)\text{)} \\
&= \sum_{i=1}^{p} \sum_{j=1}^{p} h_i \left[\mathbf{H}_f(\underset{\sim}{z} + t\underset{\sim}{h})\right]_{ij} h_j && \text{(definition of the Hessian; reorder the finite sums)} \\
&= \underset{\sim}{h}^{\top} \mathbf{H}_f(\underset{\sim}{z} + t\underset{\sim}{h}) \, \underset{\sim}{h}. && \text{(matrix product)}
\end{aligned}
$$

The last step is again the matrix product. Now apply Theorem 0.5 to g on $[0, 1]$: there is a $\tau \in (0, 1)$ with

$$
\begin{aligned}
f(\underset{\sim}{z} + \underset{\sim}{h}) &= g(1) && \text{(definition of } g\text{)} \\
&= g(0) + g'(0)\,(1 - 0) + \frac{1}{2} g''(\tau)\,(1 - 0)^2 && \text{(Theorem 0.5, with } a = 0, b = 1\text{)} \\
&= g(0) + g'(0) + \frac{1}{2} g''(\tau) && (1 - 0 = 1) \\
&= f(\underset{\sim}{z}) + \underset{\sim}{h}^{\top} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z}) + \frac{1}{2} \underset{\sim}{h}^{\top} \mathbf{H}_f(\underset{\sim}{z} + \tau\underset{\sim}{h}) \, \underset{\sim}{h} && \text{(substitute } g(0), g'(0), g''(\tau)\text{)} \\
&= f(\underset{\sim}{z}) + \left(\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{z})\right)^{\top} \underset{\sim}{h} + \frac{1}{2} \underset{\sim}{h}^{\top} \mathbf{H}_f(\underset{\sim}{z} + \tau\underset{\sim}{h}) \, \underset{\sim}{h}. && \text{(a } 1 \times 1 \text{ matrix equals its transpose)}
\end{aligned}
$$

□

Example 0.11 (For a quadratic form the expansion is exact)**.** Let $f(\underset{\sim}{x}) = \underset{\sim}{x}^{\top} \mathbf{S} \underset{\sim}{x}$ with $\mathbf{S}$ symmetric and constant. Its gradient is $2\mathbf{S}\underset{\sim}{x}$ (derivative of a quadratic form) and its Hessian is $2\mathbf{S}$ at every point (Hessian of a quadratic form), so whatever τ is, Theorem 0.6 reads

$$f(\underset{\sim}{z}+\underset{\sim}{h}) = \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{z} + (2\mathbf{S}\underset{\sim}{z})^\top \underset{\sim}{h} + \frac{1}{2}\,\underset{\sim}{h}^\top (2\mathbf{S})\underset{\sim}{h}$$
$$= \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{z} + 2\,\underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{h} + \underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{h},$$

using

$$(2\mathbf{S}\underset{\sim}{z})^\top = 2\,\underset{\sim}{z}^\top \mathbf{S}^\top$$
$$= 2\,\underset{\sim}{z}^\top \mathbf{S}.$$

Multiplying out directly gives the same:

$$(\underset{\sim}{z}+\underset{\sim}{h})^\top \mathbf{S}(\underset{\sim}{z}+\underset{\sim}{h}) = \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{z} + \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{h} + \underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{z} + \underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{h} \quad \text{(distribute)}$$
$$= \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{z} + 2\,\underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{h} + \underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{h}. \qquad (\underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{z} = (\underset{\sim}{h}^\top \mathbf{S}\underset{\sim}{z})^\top, \text{ which equals } \underset{\sim}{z}^\top \mathbf{S}\underset{\sim}{h})$$

Definition 0.5 (Strict local minimizer)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$. A point $\underset{\sim}{x}^*$ is a **strict local minimizer** of f if there is a number $\delta > 0$ such that $f(\underset{\sim}{x}^*) < f(\underset{\sim}{x})$ for every $\underset{\sim}{x} \neq \underset{\sim}{x}^*$ with $\|\underset{\sim}{x} - \underset{\sim}{x}^*\| < \delta$.

Example 0.12 (Strict and non-strict local minimizers)**.**

- $f(x) = (x-2)^2$ has

$$f(2) = 0$$
$$< (x-2)^2$$
$$= f(x)$$

  for every $x \neq 2$, so 2 is a strict local minimizer (any $\delta > 0$ works).

- A constant function $f(\underset{\sim}{x}) = 0$ has $f(\underset{\sim}{x}^*) \leq f(\underset{\sim}{x})$ for all $\underset{\sim}{x}^*$ and $\underset{\sim}{x}$, so every point is a local minimizer, but no point is a strict one: $f(\underset{\sim}{x}^*) < f(\underset{\sim}{x})$ never holds.

Lemma 0.1 (A quadratic form is at least the smallest eigenvalue times the squared length)**.** *Let* $\mathbf{A}$ *be a symmetric* $p \times p$ *matrix, and let* $\lambda_{\min}$ *be the smallest of its eigenvalues (eigendecomposition). Then* $\underset{\sim}{h}^\top \mathbf{A}\underset{\sim}{h} \geq \lambda_{\min} \|\underset{\sim}{h}\|^2$ *for every* $\underset{\sim}{h} \in \mathbb{R}^p$.

Proof

Proof. Write $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ with $\mathbf{Q}$ orthogonal and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ (spectral theorem), and let $\underset{\sim}{y} \overset{\text{def}}{=} \mathbf{Q}^\top \underset{\sim}{h}$. Then

$$
\begin{aligned}
\underset{\sim}{h}^\top \mathbf{A} \underset{\sim}{h} &= \underset{\sim}{h}^\top \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top \underset{\sim}{h} && \text{(substitute the eigendecomposition)} \\
&= \underset{\sim}{y}^\top \mathbf{\Lambda} \underset{\sim}{y} && (\underset{\sim}{h}^\top \mathbf{Q} = (\mathbf{Q}^\top \underset{\sim}{h})^\top, \text{ transpose of a product}) \\
&= \sum_{i=1}^{p} \lambda_i \, y_i^2 && (\mathbf{\Lambda} \text{ is diagonal}) \\
&\geq \sum_{i=1}^{p} \lambda_{\min} \, y_i^2 && (\lambda_i \geq \lambda_{\min} \text{ and } y_i^2 \geq 0) \\
&= \lambda_{\min} \, \|\underset{\sim}{y}\|^2 && \text{(the squared length is the sum of squares)} \\
&= \lambda_{\min} \, \|\underset{\sim}{h}\|^2. && \text{(orthogonal matrices preserve length)}
\end{aligned}
$$

The second step is the transpose of a product. The last step applies orthogonal matrices preserve length to $\mathbf{Q}^\top$, which is orthogonal because

$$
\begin{aligned}
\mathbf{Q}^\top \mathbf{Q} &= \mathbf{Q}\mathbf{Q}^\top \\
&= \mathbf{I}_p
\end{aligned}
$$

(orthogonal matrix). □

Example 0.13 (Checking the bound). $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ has eigenvalues 3 and 1 (eigenvalue example), so $\lambda_{\min} = 1$. At $\underset{\sim}{h} = (1, 0)^\top$, $\underset{\sim}{h}^\top \mathbf{A} \underset{\sim}{h} = 2 \geq 1 \cdot 1$. At $\underset{\sim}{h} = (1, -1)^\top$,

$$
\begin{aligned}
\underset{\sim}{h}^\top \mathbf{A} \underset{\sim}{h} &= 2 - 1 - 1 + 2 \\
&= 2 \\
&= 1 \cdot 2 :
\end{aligned}
$$

the bound holds with equality, since this $\underset{\sim}{h}$ is an eigenvector for $\lambda_{\min}$.

Lemma 0.2 (A quadratic form with small entries is small). *Let* $\mathbf{E}$ *be a* $p \times p$ *matrix whose entries all satisfy* $|e_{ij}| \leq m$. *Then* $|\underset{\sim}{h}^\top \mathbf{E} \underset{\sim}{h}| \leq p \, m \, \|\underset{\sim}{h}\|^2$ *for every* $\underset{\sim}{h} \in \mathbb{R}^p$.

Proof

Proof. Let $\underset{\sim}{a} \overset{\text{def}}{=} (|h_1|, \ldots, |h_p|)^\top$ and $\underset{\sim}{1} \overset{\text{def}}{=} (1, \ldots, 1)^\top$, both in $\mathbb{R}^p$. Then

$$
\begin{aligned}
\left|\underset{\sim}{h}^{\top}\mathbf{E}\underset{\sim}{h}\right| &= \left|\sum_{i=1}^{p}\sum_{j=1}^{p} h_i\, e_{ij}\, h_j\right| && \text{(matrix product)}\\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}|h_i|\,|e_{ij}|\,|h_j| && \text{(triangle inequality for numbers)}\\
&\le m\sum_{i=1}^{p}\sum_{j=1}^{p}|h_i|\,|h_j| && (|e_{ij}|\le m)\\
&= m\left(\sum_{i=1}^{p}|h_i|\right)^2 && \text{(the double sum factors)}\\
&= m\,(\underset{\sim}{1}\cdot\underset{\sim}{a})^2 && \text{(dot product)}\\
&\le m\,\|\underset{\sim}{1}\|^2\,\|\underset{\sim}{a}\|^2 && \text{(Cauchy-Schwarz, squared)}\\
&= m\,p\,\|\underset{\sim}{h}\|^2. && (\|\underset{\sim}{1}\|^2 = p \text{ and } \|\underset{\sim}{a}\|^2 = \textstyle\sum_i h_i^2, \text{ which equals } \|\underset{\sim}{h}\|^2)
\end{aligned}
$$

The steps use the matrix product, the dot product and the Cauchy-Schwarz inequality. □

Example 0.14 (Checking the bound)**.** Let $\mathbf{E} = \begin{bmatrix} 0.1 & -0.2 \\ -0.2 & 0.1 \end{bmatrix}$, so $m = 0.2$ works, and $\underset{\sim}{h} = (1, 1)^{\top}$, so $\|\underset{\sim}{h}\|^2 = 2$. Then

$$
\begin{aligned}
\underset{\sim}{h}^{\top}\mathbf{E}\underset{\sim}{h} &= 0.1 - 0.2 - 0.2 + 0.1\\
&= -0.2,
\end{aligned}
$$

and

$$
\begin{aligned}
|-0.2| &= 0.2\\
&\le 2 \cdot 0.2 \cdot 2\\
&= 0.8.
\end{aligned}
$$

Theorem 0.7 (Second-order sufficient condition)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ have first and second partial derivatives that are continuous on $\mathbb{R}^p$ (continuity). If $\underset{\sim}{x}^*$ is a stationary point of f (Definition 0.4) and $\mathbf{H}_f(\underset{\sim}{x}^*)$ is positive definite, then $\underset{\sim}{x}^*$ is a strict local minimizer of f (Definition 0.5).*

Proof

Proof. $\mathbf{H}_f(\underset{\sim}{x}^*)$ is positive definite, and so symmetric, so its smallest eigenvalue λ is positive (definiteness and eigenvalues). Each of the p^2 entries of $\mathbf{H}_f$ is continuous at $\underset{\sim}{x}^*$ (continuity): for $\varepsilon = \lambda/(2p)$ there is a $\delta_{ij} > 0$ with $\left|\left[\mathbf{H}_f(\underset{\sim}{y})\right]_{ij} - \left[\mathbf{H}_f(\underset{\sim}{x}^*)\right]_{ij}\right| < \lambda/(2p)$

whenever $\|y - \underset{\sim}{x}^*\| < \delta_{ij}$. Let δ be the smallest of these p^2 numbers.
Now let $0 < \|\underset{\sim}{h}\| < \delta$. By Theorem 0.6 there is a $\tau \in (0,1)$ for which the first step of the next display holds; let $y \overset{\text{def}}{=} \underset{\sim}{x}^* + \tau \underset{\sim}{h}$ and $\mathbf{E} \overset{\text{def}}{=} \mathbf{H}_f(y) - \mathbf{H}_f(\underset{\sim}{x}^*)$. Since $\|y - \underset{\sim}{x}^*\| = \tau \|\underset{\sim}{h}\| < \delta$, every entry of $\mathbf{E}$ has absolute value less than $\lambda/(2p)$, so Lemma 0.2 applies with $m = \lambda/(2p)$:

$$
\begin{aligned}
f(\underset{\sim}{x}^* + \underset{\sim}{h}) - f(\underset{\sim}{x}^*) &= \left(\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^*)\right)^\top \underset{\sim}{h} + \frac{1}{2}\, \underset{\sim}{h}^\top \mathbf{H}_f(\underset{\sim}{y})\, \underset{\sim}{h} && \text{(Theorem 0.6)} \\
&= \frac{1}{2}\, \underset{\sim}{h}^\top \mathbf{H}_f(\underset{\sim}{y})\, \underset{\sim}{h} && (\underset{\sim}{x}^* \text{ is stationary}) \\
&= \frac{1}{2}\, \underset{\sim}{h}^\top \mathbf{H}_f(\underset{\sim}{x}^*)\, \underset{\sim}{h} + \frac{1}{2}\, \underset{\sim}{h}^\top \mathbf{E}\, \underset{\sim}{h} && (\mathbf{H}_f(\underset{\sim}{y}) = \mathbf{H}_f(\underset{\sim}{x}^*) + \mathbf{E}\text{; distribute}) \\
&\geq \frac{1}{2}\, \lambda\, \|\underset{\sim}{h}\|^2 + \frac{1}{2}\, \underset{\sim}{h}^\top \mathbf{E}\, \underset{\sim}{h} && \text{(Lemma 0.1)} \\
&\geq \frac{1}{2}\, \lambda\, \|\underset{\sim}{h}\|^2 - \frac{1}{2}\, \left|\underset{\sim}{h}^\top \mathbf{E}\, \underset{\sim}{h}\right| && \text{(a number is at least minus its absolute value)} \\
&\geq \frac{1}{2}\, \lambda\, \|\underset{\sim}{h}\|^2 - \frac{1}{2}\, p\, \frac{\lambda}{2p}\, \|\underset{\sim}{h}\|^2 && \text{(Lemma 0.2, } m = \lambda/(2p)) \\
&= \frac{\lambda}{4}\, \|\underset{\sim}{h}\|^2 && \text{(arithmetic)} \\
&> 0. && (\lambda > 0 \text{ and } \underset{\sim}{h} \neq \underset{\sim}{0})
\end{aligned}
$$

So $f(\underset{\sim}{x}^*) < f(\underset{\sim}{x})$ whenever $0 < \|\underset{\sim}{x} - \underset{\sim}{x}^*\| < \delta$ (take $\underset{\sim}{h} = \underset{\sim}{x} - \underset{\sim}{x}^*$). □

Example 0.15 (Classifying stationary points)**.**

- $f(\underset{\sim}{x}) = x_1^2 + x_2^2 - 2x_1$ has the stationary point $(1,0)^\top$ (Example 0.7), and its Hessian is

$$
\begin{aligned}
\mathbf{H}_f(\underset{\sim}{x}) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\
&= 2\mathbf{I}_2,
\end{aligned}
$$

  which is positive definite, so $(1,0)^\top$ is a strict local minimizer.

- $f(\underset{\sim}{x}) = x_1^2 - x_2^2$ has the stationary point $\underset{\sim}{0}$ and Hessian $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$, which is not positive definite ($(0,1)^\top$ gives -2), and indeed $\underset{\sim}{0}$ is not a local minimizer (Example 0.8).

Example 0.16 (The condition is sufficient, not necessary)**.** $f(\underset{\sim}{x}) = x_1^2 + x_2^4$ and $g(\underset{\sim}{x}) = x_1^2 - x_2^4$ both have gradient $\underset{\sim}{0}$ at $\underset{\sim}{0}$ and the same Hessian there, $\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$, since

$$\begin{aligned} \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2}(\pm x_2^4)\right) &= \pm 12 x_2^2 \\ &= 0 \end{aligned}$$

at $x_2 = 0$. That Hessian is not positive definite ($(0,1)^\top$ gives 0), so Theorem 0.7 says nothing about either function. In fact $\underset{\sim}{0}$ is a strict local minimizer of f, since $f(\underset{\sim}{x}) > 0 = f(\underset{\sim}{0})$ for $\underset{\sim}{x} \neq \underset{\sim}{0}$, but not a local minimizer of g, since

$$\begin{aligned} g(0,t) &= -t^4 \\ &< 0 \\ &= g(\underset{\sim}{0}) \end{aligned}$$

for $t \neq 0$.

Definition 0.6 (Local maximizer)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$. A point $\underset{\sim}{x}^* \in \mathbb{R}^p$ is a **local maximizer** of f if there is a number $\delta > 0$ such that $f(\underset{\sim}{x}^*) \geq f(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathbb{R}^p$ with $\|\underset{\sim}{x} - \underset{\sim}{x}^*\| < \delta$.

Example 0.17 (A local maximizer)**.** Let $f(x) = 1 - x^2$. For every $x \in \mathbb{R}$,

$$\begin{aligned} f(0) - f(x) &= 1 - (1 - x^2) && \text{(substitute)} \\ &= x^2 && \text{(simplify)} \\ &\geq 0, && \text{(a square)} \end{aligned}$$

so $f(0) \geq f(x)$ for all x, and 0 is a local maximizer of f with any $\delta > 0$.

Definition 0.7 (Saddle point)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$. A stationary point of f that is neither a local minimizer nor a local maximizer of f is a **saddle point** of f [7].

Example 0.18 (A saddle point)**.** Let $f(\underset{\sim}{x}) = x_1^2 - x_2^2$. Its gradient is $(2x_1, -2x_2)^\top$, which is $\underset{\sim}{0}_{2\times 1}$ only at $\underset{\sim}{x} = \underset{\sim}{0}_{2\times 1}$, so $\underset{\sim}{0}_{2\times 1}$ is a stationary point. For every $t \neq 0$,

$$
\begin{aligned}
f(t,0) &= t^2 && \text{(substitute)} \\
&> 0 && (t \neq 0) \\
&= f(\underset{\sim}{0}_{2\times 1}), && \text{(substitute)} \\
f(0,t) &= -t^2 && \text{(substitute)} \\
&< 0 && (t \neq 0) \\
&= f(\underset{\sim}{0}_{2\times 1}). && \text{(substitute)}
\end{aligned}
$$

So every ball around $\underset{\sim}{0}_{2\times 1}$ contains a point with a larger value of f and a point with a smaller one. Hence $\underset{\sim}{0}_{2\times 1}$ is neither a local minimizer nor a local maximizer, and it is a saddle point.

[1] Wikipedia contributors, "Stationary point — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Stationary_point

[2] 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

[3] H. Zhou, "Optimization and multivariate calculus." Lecture notes for Biostat 216, Mathematical Methods for Biostatistics, University of California, Los Angeles, 2024. Accessed: Oct. 05, 2026. [Online]. Available: https://ucla-biostat-216.github.io/2024fall/slides/13-optim/13-optim.html

[4] Wikipedia contributors, "Fermat's theorem (stationary points) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Fermat%27s_theorem_(stationary_points)

[5] Wikipedia contributors, "Taylor's theorem — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Taylor%27s_theorem

[6] W. Rudin, *Principles of mathematical analysis*, 3rd ed. in International series in pure and applied mathematics. New York: McGraw-Hill, 1976.

[7] Wikipedia contributors, "Saddle point — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Saddle_point