

Optimization

Last modified: 2026-10-27 20:53:54 (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). Let $f(w) = w^2 - 4w + 7$. 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 .

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 . The linear approximation drops everything of that order and smaller, so halving the step quarters the error. That trade is the whole bargain of gradient descent, the step-by-step method of fitting models that the optimization page defines. We take a step in the direction the derivative recommends, and the recommendation is trustworthy only as far as the step is small.

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 $h \rightarrow 0$, 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).*

i 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. $\square$

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 $w = 1$. These are flat points of f , 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.

Exercise 0.8 (Maximizing a gamma-family kernel). Adapted from Miller [2], Question 1.1.10.

Find the maximum value of

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

on the interval $[0, \infty)$.

Solution

Solution 0.8. Differentiate using the product rule:

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

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 $x = 4$:

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

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

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.
Find the critical points of

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

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 $\tilde{z}$ and $\tilde{z} + \tilde{h}$. 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 $\tilde{x}$: the gradient $\frac{\partial}{\partial \tilde{x}} f(\tilde{x})$ (**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, $\frac{\partial}{\partial \tilde{x}} f(\tilde{z})$ means the gradient evaluated at $\tilde{x} = \tilde{z}$.

Theorem 0.2 (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]*

i 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 $h \rightarrow 0$ 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 $h \rightarrow 0$ from the left, $g'(c) \leq 0$. Together, $g'(c) = 0$. $\square$

Example 0.4 (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 0.3 (Stationary point). Let f be a scalar-valued function of a $p \times 1$ vector $\tilde{x}$. A point $\tilde{z}$ at which the gradient exists and

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

is a **stationary point** of f .

Example 0.5 (A stationary point, and a point that is not one). Let $f(\tilde{x}) = x_1^2 + x_2^2 - 2x_1$. Its gradient is $(2x_1 - 2, 2x_2)^\top$, which is $\mathbf{0}_{2 \times 1}$ exactly when $x_1 = 1$ and $x_2 = 0$. So $(1, 0)^\top$ is its only stationary point, and $\mathbf{0}$ is not one: the gradient there is $(-2, 0)^\top$.

i Remark (Stationary points in one variable)

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

Theorem 0.3 (First-order necessary condition). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$. If $\tilde{x}^*$ is a **local minimizer** of f and the gradient of f exists at $\tilde{x}^*$, then $\tilde{x}^*$ is a stationary point of f (Definition 0.3).

i Proof

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

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

so

$$\begin{aligned}
g_i(0) &= f(\tilde{x}^*) \\
&\leq f(\tilde{x}^* + t e_i) \\
&= g_i(t) :
\end{aligned}$$

0 is a local minimizer of g_i . Moving $\tilde{x}$ from $\tilde{x}^*$ along e_i changes only x_i , so $g'_i(0)$ is the partial derivative $\frac{\partial}{\partial x_i} f(\tilde{x}^*)$, which exists because the gradient does. By Theorem 0.2,

$$\begin{aligned}
\frac{\partial}{\partial x_i} f(\tilde{x}^*) &= g'_i(0) \\
&= 0.
\end{aligned}$$

The equation $g'_i(0) = 0$ holds for every i , so every entry of the gradient at $\tilde{x}^*$ is 0 ([vector derivative](#)). $\square$

Example 0.6 (Using the condition to locate, and to rule out, minimizers).

- For $f(x) = x_1^2 + x_2^2 - 2x_1$ of Example 0.5, any local minimizer must be the stationary point $(1, 0)^\top$. It is one: completing the square,

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

- For $f(x) = e^{2x_1+x_2} - x_1$ 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: $f(x) = x_1^2 - x_2^2$ has gradient $(2x_1, -2x_2)^\top$, so $\tilde{0}$ is a stationary point, but

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

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

Example 0.7 (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

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

a scalar-valued function of the 2×1 vector β . Its gradient is $-2\mathbf{X}^\top(y - \mathbf{X}\beta)$. By Theorem 0.3, a minimizer of f must make this gradient 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 $\tilde{\beta} = 0$.

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"],
)
#>
#> 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:

```
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

```
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] 1.071498e-11 1.673470e-10
#>
#>
#> $zero_vector
#> $zero_vector$beta
#> [1] 0 0
#>
#> $zero_vector$gradient
#> [1] -18381.6 -1737437.6
```

A library fit gives the same coefficients:

```
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 $1.7\text{e-}10$, zero up to rounding error, against 1,737,438 at the zero vector. So the solution is a stationary point (Definition 0.3), and the zero vector is not. The library fits above agree with the solution of that equation.

Theorem 0.4 (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 0.8 (The remainder point for e^x). Take $g(x) = e^x$ on $[0, 1]$, so

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

Theorem 0.4 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 0.5 (Second-order Taylor theorem in several variables). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first and second partial derivatives that are continuous on $\mathbb{R}^p$ ([continuity](#)), and let $z, h \in \mathbb{R}^p$. Then there is a $\tau \in (0, 1)$ with [5]*

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

i Proof

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

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

The second step writes out the **matrix product**. g' is continuous (**continuity**): $t \mapsto \tilde{z} + t\tilde{h}$ is continuous, since $\|(\tilde{z} + s\tilde{h}) - (\tilde{z} + t\tilde{h})\| = |s - t| \|\tilde{h}\|$, 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: $\frac{\partial}{\partial t} \left[\frac{\partial}{\partial x_j} f(\tilde{z} + t\tilde{h}) \right] = \sum_{i=1}^p h_i \frac{\partial}{\partial x_i} \left(\frac{\partial}{\partial x_j} f(\tilde{z} + t\tilde{h}) \right)$. So

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

The last step is again the **matrix product**. Now apply Theorem 0.4 to g on $[0, 1]$: there is a $\tau \in (0, 1)$ with

$$\begin{aligned} f(\tilde{z} + \tilde{h}) &= g(1) && \text{(definition of } g) \\ &= g(0) + g'(0)(1 - 0) + \frac{1}{2} g''(\tau)(1 - 0)^2 && \text{(Theorem 0.4, with } a = 0, b = 1) \\ &= g(0) + g'(0) + \frac{1}{2} g''(\tau) && (1 - 0 = 1) \\ &= f(\tilde{z}) + \tilde{h}^\top \frac{\partial}{\partial \tilde{x}} f(\tilde{z}) + \frac{1}{2} \tilde{h}^\top \mathbf{H}_f(\tilde{z} + \tau\tilde{h}) \tilde{h} && \text{(substitute } g(0), g'(0), g''(\tau)) \\ &= f(\tilde{z}) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{z}) \right)^\top \tilde{h} + \frac{1}{2} \tilde{h}^\top \mathbf{H}_f(\tilde{z} + \tau\tilde{h}) \tilde{h}. && \text{(a } 1 \times 1 \text{ matrix equals its transpose)} \end{aligned}$$

□

Example 0.9 (For a quadratic form the expansion is exact). Let $f(\tilde{x}) = \tilde{x}^\top \mathbf{S} \tilde{x}$ with $\mathbf{S}$ symmetric and constant. Its gradient is $2\mathbf{S}\tilde{x}$ ([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 0.5 reads

$$\begin{aligned} f(\tilde{z} + \tilde{h}) &= \tilde{z}^\top \mathbf{S} \tilde{z} + (2\mathbf{S}\tilde{z})^\top \tilde{h} + \frac{1}{2} \tilde{h}^\top (2\mathbf{S}) \tilde{h} \\ &= \tilde{z}^\top \mathbf{S} \tilde{z} + 2 \tilde{z}^\top \mathbf{S} \tilde{h} + \tilde{h}^\top \mathbf{S} \tilde{h}, \end{aligned}$$

using

$$\begin{aligned} (2\mathbf{S}\tilde{z})^\top &= 2 \tilde{z}^\top \mathbf{S}^\top \\ &= 2 \tilde{z}^\top \mathbf{S}. \end{aligned}$$

Multiplying out directly gives the same:

$$\begin{aligned} (\tilde{z} + \tilde{h})^\top \mathbf{S} (\tilde{z} + \tilde{h}) &= \tilde{z}^\top \mathbf{S} \tilde{z} + \tilde{z}^\top \mathbf{S} \tilde{h} + \tilde{h}^\top \mathbf{S} \tilde{z} + \tilde{h}^\top \mathbf{S} \tilde{h} \quad (\text{distribute}) \\ &= \tilde{z}^\top \mathbf{S} \tilde{z} + 2 \tilde{z}^\top \mathbf{S} \tilde{h} + \tilde{h}^\top \mathbf{S} \tilde{h}. \quad (\tilde{h}^\top \mathbf{S} \tilde{z} = (\tilde{h}^\top \mathbf{S} \tilde{z})^\top, \text{ which equals } \tilde{z}^\top \mathbf{S} \tilde{h}) \end{aligned}$$

Definition 0.4 (Strict local minimizer). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$. A point $\tilde{x}^*$ is a **strict local minimizer** of f if there is a number $\delta > 0$ such that $f(\tilde{x}^*) < f(\tilde{x})$ for every $\tilde{x} \neq \tilde{x}^*$ with $\|\tilde{x} - \tilde{x}^*\| < \delta$.

Example 0.10 (Strict and non-strict local minimizers).

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

$$\begin{aligned} f(2) &= 0 \\ &< (x - 2)^2 \\ &= f(x) \end{aligned}$$

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

- A constant function $f(\tilde{x}) = 0$ has $f(\tilde{x}^*) \leq f(\tilde{x})$ for all $\tilde{x}^*$ and $\tilde{x}$, so every point is a [local minimizer](#), but no point is a strict one: $f(\tilde{x}^*) < f(\tilde{x})$ 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 $\tilde{h}^\top \mathbf{A} \tilde{h} \geq \lambda_{\min} \|\tilde{h}\|^2$ for every $\tilde{h} \in \mathbb{R}^p$.

i Proof

Proof. Write $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ with $\mathbf{Q}$ orthogonal and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ ([spectral theorem](#)), and let $\tilde{y} \stackrel{\text{def}}{=} \mathbf{Q}^\top \tilde{h}$. Then

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

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 0.11 (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 $\tilde{h} = (1, 0)^\top$, $\tilde{h}^\top \mathbf{A} \tilde{h} = 2 \geq 1 \cdot 1$. At $\tilde{h} = (1, -1)^\top$,

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

the bound holds with equality, since this $\tilde{h}$ 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 $|\tilde{h}^\top \mathbf{E} \tilde{h}| \leq p m \|\tilde{h}\|^2$ for every $\tilde{h} \in \mathbb{R}^p$.*

i Proof

Proof. Let $\tilde{a} \stackrel{\text{def}}{=} (|h_1|, \dots, |h_p|)^\top$ and $\tilde{1} \stackrel{\text{def}}{=} (1, \dots, 1)^\top$, both in $\mathbb{R}^p$. Then

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

The steps use the [matrix product](#), the [dot product](#) and the [Cauchy-Schwarz inequality](#). $\square$

Example 0.12 (Checking the bound). Let $\mathbf{E} = \begin{bmatrix} 0.1 & -0.2 \\ -0.2 & 0.1 \end{bmatrix}$, so $m = 0.2$ works, and $\tilde{h} = (1, 1)^\top$, so $\|\tilde{h}\|^2 = 2$. Then

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

and

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

Theorem 0.6 (Second-order sufficient condition). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first and second partial derivatives that are continuous on $\mathbb{R}^p$ ([continuity](#)). If $\tilde{x}^*$ is a stationary point of f (Definition 0.3) and $\mathbf{H}_f(\tilde{x}^*)$ is [positive definite](#), then $\tilde{x}^*$ is a strict local minimizer of f (Definition 0.4).*

i Proof

Proof. $\mathbf{H}_f(\tilde{x}^*)$ 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 $\tilde{x}^*$ ([continuity](#)): for $\varepsilon = \lambda/(2p)$ there is a $\delta_{ij} > 0$ with $|\mathbf{H}_f(y)_{ij} - \mathbf{H}_f(\tilde{x}^*)_{ij}| < \lambda/(2p)$

whenever $\|y - x^*\| < \delta_{ij}$. Let δ be the smallest of these p^2 numbers. Now let $0 < \|h\| < \delta$. By Theorem 0.5 there is a $\tau \in (0, 1)$ for which the first step of the next display holds; let $y \stackrel{\text{def}}{=} x^* + \tau h$ and $\mathbf{E} \stackrel{\text{def}}{=} \mathbf{H}_f(y) - \mathbf{H}_f(x^*)$. Since $\|y - x^*\| = \tau \|h\| < \delta$, every entry of $\mathbf{E}$ has absolute value less than $\lambda/(2p)$, so Lemma 0.2 applies with $m = \lambda/(2p)$:

$$\begin{aligned}
f(x^* + h) - f(x^*) &= \left(\frac{\partial}{\partial x} f(x^*) \right)^\top h + \frac{1}{2} h^\top \mathbf{H}_f(y) h && \text{(Theorem 0.5)} \\
&= \frac{1}{2} h^\top \mathbf{H}_f(y) h && (x^* \text{ is stationary}) \\
&= \frac{1}{2} h^\top \mathbf{H}_f(x^*) h + \frac{1}{2} h^\top \mathbf{E} h && (\mathbf{H}_f(y) = \mathbf{H}_f(x^*) + \mathbf{E}; \text{distribute}) \\
&\geq \frac{1}{2} \lambda \|h\|^2 + \frac{1}{2} h^\top \mathbf{E} h && \text{(Lemma 0.1)} \\
&\geq \frac{1}{2} \lambda \|h\|^2 - \frac{1}{2} |h^\top \mathbf{E} h| && (\text{a number is at least minus its absolute value}) \\
&\geq \frac{1}{2} \lambda \|h\|^2 - \frac{1}{2} p \frac{\lambda}{2p} \|h\|^2 && \text{(Lemma 0.2, } m = \lambda/(2p)) \\
&= \frac{\lambda}{4} \|h\|^2 && (\text{arithmetic}) \\
&> 0. && (\lambda > 0 \text{ and } h \neq 0)
\end{aligned}$$

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

Example 0.13 (Classifying stationary points).

- $f(x) = x_1^2 + x_2^2 - 2x_1$ has the stationary point $(1, 0)^\top$ (Example 0.5), and its Hessian is

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

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

- $f(x) = x_1^2 - x_2^2$ has the stationary point 0 and Hessian $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$, which is not positive definite ($(0, 1)^\top$ gives -2), and indeed 0 is not a local minimizer (Example 0.6).

Example 0.14 (The condition is sufficient, not necessary). $f(x) = x_1^2 + x_2^4$ and $g(x) = x_1^2 - x_2^4$ both have gradient $\mathbf{0}$ at $\mathbf{0}$ and the same Hessian there, $\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$, since

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

at $x_2 = 0$. That Hessian is not positive definite ($(0,1)^\top$ gives 0), so Theorem 0.6 says nothing about either function. In fact $\mathbf{0}$ is a strict local minimizer of f , since $f(x) > 0 = f(\mathbf{0})$ for $x \neq \mathbf{0}$, but not a local minimizer of g , since

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

for $t \neq 0$.

Definition 0.5 (Local maximizer). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$. A point $x^* \in \mathbb{R}^p$ is a **local maximizer** of f if there is a number $\delta > 0$ such that $f(x^*) \geq f(x)$ for all $x \in \mathbb{R}^p$ with $\|x - x^*\| < \delta$.

Example 0.15 (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 0.6 (Saddle point). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$. 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 0.16 (A saddle point). Let $f(x) = x_1^2 - x_2^2$. Its gradient is $(2x_1, -2x_2)^\top$, which is $\mathbf{0}_{2 \times 1}$ only at $x = \mathbf{0}_{2 \times 1}$, so $\mathbf{0}_{2 \times 1}$ 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.

Convexity in several variables

Source

This section is adapted from the convexity part of Zhou [3], used under the MIT License (see the license text in [Linear Algebra](#)). The source states the first-order test (in one direction) and the second-derivative test for convexity without proof; here both directions of the first-order test are proved, and the second-derivative test is proved from the Taylor theorem of Section . The source defines both concave and strictly concave as “ $-f$ is strictly convex”; here f is concave when $-f$ is convex. The source states that an intersection of convex sets is convex and that a maximum of convex functions is convex for any finite number of them; here each is stated for two and the general case follows by induction. These parts of the source are not part of this section:

- its catalogs of convex sets and convex functions, including the positive semidefinite matrices and $-\log \det$
- intersections of infinitely many convex sets, and suprema of infinitely many convex functions
- its examples, exercises and figures, including the multivariate normal maximum likelihood example

Definition 0.7 (Convex set). A set $K \subseteq \mathbb{R}^p$ is **convex** if [8] for all $\underset{\sim}{x}, \underset{\sim}{y} \in K$ and all $t \in [0, 1]$, the point $t\underset{\sim}{x} + (1-t)\underset{\sim}{y}$ is in K : K contains the whole line segment between any two of its points.

Example 0.17 (Convex sets).

- $\mathbb{R}^p$ is convex, and so is any set with at most one point.

- The closed ball (**closed ball**) $B = \{\tilde{x} : \|\tilde{x}\| \leq r\}$ is convex: for $\tilde{x}, \tilde{y} \in B$ and $t \in [0, 1]$,

$$\begin{aligned}
\|t\tilde{x} + (1-t)\tilde{y}\| &\leq \|t\tilde{x}\| + \|(1-t)\tilde{y}\| && \text{(triangle inequality)} \\
&= t\|\tilde{x}\| + (1-t)\|\tilde{y}\| && \text{(norm of a multiple; } t, 1-t \geq 0) \\
&\leq tr + (1-t)r && (\tilde{x}, \tilde{y} \in B) \\
&= r. && \text{(arithmetic)}
\end{aligned}$$

The first step is the **triangle inequality**, and the second uses **the norm of a multiple**.

Example 0.18 (A set that is not convex). $K = \{x \in \mathbb{R} : |x| \geq 1\}$ is not convex: -1 and 1 are in K , but with $t = \frac{1}{2}$ the point $\frac{1}{2}(-1) + \frac{1}{2}(1) = 0$ is not.

Definition 0.8 (Half-space). Let $a \in \mathbb{R}^p$ be a nonzero vector and c a real number. The set $\{\tilde{x} \in \mathbb{R}^p : a \cdot \tilde{x} \geq c\}$ is a (closed) **half-space**: the points on one side of the boundary $\{\tilde{x} : a \cdot \tilde{x} = c\}$, together with the boundary. In $\mathbb{R}^2$, a half-space is called a **half-plane**.

Example 0.19 (The right half-plane). With $a = (1, 0)^\top$ and $c = 0$, the half-plane $\{\tilde{x} \in \mathbb{R}^2 : x_1 \geq 0\}$ is the set of points on or to the right of the x_2 -axis. $(1, 5)^\top$ is in it, since $1 \geq 0$; $(0, -3)^\top$ is in it, on the boundary; and $(-1, 0)^\top$ is not, since $-1 < 0$.

Theorem 0.7 (An intersection of convex sets is convex). *If K_1 and K_2 are convex subsets of $\mathbb{R}^p$ (Definition 0.7), then $K_1 \cap K_2$ is convex.*

i Proof

Proof. Let $\tilde{x}, \tilde{y} \in K_1 \cap K_2$ and $t \in [0, 1]$. Then $\tilde{x}, \tilde{y} \in K_1$, so $t\tilde{x} + (1-t)\tilde{y} \in K_1$, since K_1 is convex; in the same way $t\tilde{x} + (1-t)\tilde{y} \in K_2$. So $t\tilde{x} + (1-t)\tilde{y} \in K_1 \cap K_2$. $\square$

Example 0.20 (Intersecting a ball with a half-plane). The closed unit ball (**unit ball**) in $\mathbb{R}^2$ is convex (Example 0.17), and so is the half-plane (Definition 0.8) $\{\tilde{x} : x_1 \geq 0\}$: if $x_1 \geq 0$ and $y_1 \geq 0$, then $tx_1 + (1-t)y_1 \geq 0$ for $t \in [0, 1]$. By Theorem 0.7, the closed half-disk $\{\tilde{x} : \|\tilde{x}\| \leq 1, x_1 \geq 0\}$ is convex. A union need not be convex: $\{x \in \mathbb{R} : x \leq -1\}$ and $\{x \in \mathbb{R} : x \geq 1\}$ are convex, but their union is the set of Example 0.18.

Definition 0.9 (Strictly convex function). A function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is **strictly convex** if $f(t\tilde{x} + (1-t)\tilde{y}) < tf(\tilde{x}) + (1-t)f(\tilde{y})$ for all $\tilde{x} \neq \tilde{y}$ in $\mathbb{R}^p$ and all $t \in (0, 1)$.

Definition 0.10 (Concave function). A function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is **concave** if $-f$ is **convex**.

Definition 0.11 (Strictly concave function). A function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is **strictly concave** if $-f$ is strictly convex (Definition 0.9).

Example 0.21 (Strictly convex, and convex but not strictly).

- $f(x) = x^2$ is strictly convex:

$$\begin{aligned} tx^2 + (1-t)y^2 - (tx + (1-t)y)^2 &= tx^2 + (1-t)y^2 - t^2x^2 - 2t(1-t)xy - (1-t)^2y^2 \\ &= t(1-t)x^2 - 2t(1-t)xy + t(1-t)y^2 \\ &= t(1-t)(x-y)^2, \end{aligned}$$

(expand the square)
($t - t^2 = t(1-t)$)
(factor)

which is positive when $x \neq y$ and $t \in (0, 1)$. At $t \in \{0, 1\}$ or $x = y$ it is 0, so x^2 is also convex.

- $f(x) = 2x + 1$ is convex but not strictly convex:

$$\begin{aligned} f(tx + (1-t)y) &= 2tx + 2(1-t)y + 1 \\ &= tf(x) + (1-t)f(y) \end{aligned}$$

for every x, y, t , so the inequality holds, but never strictly. It is also concave.

Theorem 0.8 (The maximum of convex functions is convex). If f_1 and f_2 are **convex** functions on $\mathbb{R}^p$, then $f(\tilde{x}) \stackrel{\text{def}}{=} \max\{f_1(\tilde{x}), f_2(\tilde{x})\}$ is convex.

i Proof

Proof. Let $\tilde{x}, \tilde{y} \in \mathbb{R}^p$ and $t \in [0, 1]$. For $i = 1, 2$,

$$\begin{aligned} f_i(t\tilde{x} + (1-t)\tilde{y}) &\leq tf_i(\tilde{x}) + (1-t)f_i(\tilde{y}) \quad (f_i \text{ is convex}) \\ &\leq tf(\tilde{x}) + (1-t)f(\tilde{y}). \quad (f_i \leq f, \text{ and } t, 1-t \geq 0) \end{aligned}$$

Both f_1 and f_2 at $t\tilde{x} + (1-t)\tilde{y}$ are at most $tf(\tilde{x}) + (1-t)f(\tilde{y})$, so their maximum $f(t\tilde{x} + (1-t)\tilde{y})$ is too. $\square$

Example 0.22 (A maximum is convex; a minimum need not be).

- $f_1(x) = x$ is convex, by the computation for $2x + 1$ in Example 0.21 with 2 and 1 replaced by 1 and 0, and $f_2(x) = x^2$ is convex by the same example, so $\max\{x, x^2\}$ is convex by Theorem 0.8.
- The minimum of two convex functions need not be convex: $g(x) = \min\{x^2, (x - 2)^2\}$ has

$$\begin{aligned} g(0) &= g(2) \\ &= 0, \end{aligned}$$

but at the midpoint

$$\begin{aligned} g(1) &= \min\{1, 1\} \\ &= 1 \\ &> \frac{1}{2}g(0) + \frac{1}{2}g(2) \\ &= 0. \end{aligned}$$

Definition 0.12 (Epigraph). The **epigraph** [9] of $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is the set of points on or over its [graph](#),

$$\text{epi } f \stackrel{\text{def}}{=} \{(\tilde{x}, s) : \tilde{x} \in \mathbb{R}^p, s \in \mathbb{R}, f(\tilde{x}) \leq s\}.$$

A pair (x, s) is treated as the $(p + 1)$ -vector with entries $x_1, \dots, x_p, s$, so $\text{epi } f \subseteq \mathbb{R}^{p+1}$, and sums and multiples of pairs are taken entry by entry.

Example 0.23 (The epigraph of x^2). For $f(x) = x^2$, $\text{epi } f = \{(x, s) : x^2 \leq s\}$, the points (x, s) on the parabola $s = x^2$ or on the side of it where s is larger. $(1, 2)$ is in it, since $1^2 \leq 2$; $(2, 1)$ is not, since $2^2 > 1$.

Theorem 0.9 (Convexity through the epigraph). $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is [convex](#) if and only if its epigraph (Definition 0.12) is convex (Definition 0.7).

i Proof

Proof. If. Let $x, y \in \mathbb{R}^p$ and $t \in [0, 1]$. The points $(\tilde{x}, f(\tilde{x}))$ and $(\tilde{y}, f(\tilde{y}))$ are in $\text{epi } f$, since $f(\tilde{x}) \leq f(\tilde{x})$ and $f(\tilde{y}) \leq f(\tilde{y})$. Since $\text{epi } f$ is convex, it contains

$$t(\tilde{x}, f(\tilde{x})) + (1-t)(\tilde{y}, f(\tilde{y})) = (t\tilde{x} + (1-t)\tilde{y}, tf(\tilde{x}) + (1-t)f(\tilde{y})),$$

and by Definition 0.12 that membership means $f(t\tilde{x} + (1-t)\tilde{y}) \leq tf(\tilde{x}) + (1-t)f(\tilde{y})$.

Only if. Let $(\tilde{x}, s)$ and $(\tilde{y}, u)$ be in $\text{epi } f$, so $f(\tilde{x}) \leq s$ and $f(\tilde{y}) \leq u$, and let $t \in [0, 1]$. Then

$$\begin{aligned} f(t\tilde{x} + (1-t)\tilde{y}) &\leq tf(\tilde{x}) + (1-t)f(\tilde{y}) \quad (f \text{ is convex}) \\ &\leq ts + (1-t)u, \quad (f(\tilde{x}) \leq s, f(\tilde{y}) \leq u, \text{ and } t, 1-t \geq 0) \end{aligned}$$

so $t(\tilde{x}, s) + (1-t)(\tilde{y}, u) = (t\tilde{x} + (1-t)\tilde{y}, ts + (1-t)u)$ is in $\text{epi } f$. $\square$

Example 0.24 (Two epigraphs).

- The epigraph of $f(x) = |x|$ is $\{(x, s) : |x| \leq s\}$, a wedge. It is the intersection of the half-planes $\{(x, s) : s - x \geq 0\}$ and $\{(x, s) : s + x \geq 0\}$. Each half-plane is convex: if $s_1 - x_1 \geq 0$ and $s_2 - x_2 \geq 0$, then $(ts_1 + (1-t)s_2) - (tx_1 + (1-t)x_2) = t(s_1 - x_1) + (1-t)(s_2 - x_2) \geq 0$, and the same with $+$ for the other. So the wedge is convex (Theorem 0.7), and $|x|$ is convex by Theorem 0.9.
- The epigraph of $g(x) = -x^2$ is not convex: $(-1, -1)$ and $(1, -1)$ are in it, since $g(\pm 1) = -1$, but their midpoint $(0, -1)$ is not, since $g(0) = 0 > -1$. So $-x^2$ is not convex.

Definition 0.13 (Tangent plane). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have a gradient at $\tilde{y}$ (gradient). The **tangent plane** [10] to the graph of f at $\tilde{y}$ is the graph of the function

$$\tilde{x} \mapsto f(\tilde{y}) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top (\tilde{x} - \tilde{y}),$$

which takes the value $f(\tilde{y})$ at $\tilde{x} = \tilde{y}$ and has the same gradient as f there.

Example 0.25 (A tangent line and a tangent plane).

- For $f(x) = x^2$ at $y = 1$, $f(1) = 1$ and $f'(1) = 2$, so the tangent line is

$$\begin{aligned} s &= 1 + 2(x - 1) \quad (\text{substitute } f(1) \text{ and } f'(1)) \\ &= 1 + 2x - 2 \quad (\text{distribute}) \\ &= 2x - 1. \quad (1 - 2 = -1) \end{aligned}$$

- For $f(x) = x_1^2 + x_2^2$ at $y = (1, 1)^\top$, $f(\tilde{y}) = 2$ and the gradient is $(2y_1, 2y_2)^\top = (2, 2)^\top$, so the tangent plane is

$$\begin{aligned}
s &= 2 + 2(x_1 - 1) + 2(x_2 - 1) && \text{(substitute } f(\tilde{y}) \text{ and the gradient)} \\
&= 2 + 2x_1 - 2 + 2x_2 - 2 && \text{(distribute)} \\
&= 2x_1 + 2x_2 - 2. && (2 - 2 - 2 = -2)
\end{aligned}$$

i Remark (The tangent plane in one variable)

Remark 0.2 (The tangent plane in one variable). For $p = 1$ the tangent plane is the **tangent line** $s = f(y) + f'(y)(x - y)$, the line through $(y, f(y))$ with slope $f'(y)$.

Theorem 0.10 (First-order characterization of convexity). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first partial derivatives that are continuous on $\mathbb{R}^p$ (**continuity**). Then f is **convex** if and only if*

$$f(\tilde{x}) \geq f(\tilde{y}) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top (\tilde{x} - \tilde{y}) \quad \text{for all } \tilde{x}, \tilde{y} \in \mathbb{R}^p : \quad (1)$$

every tangent plane (Definition 0.13) lies on or under the graph. If the inequality is strict whenever $\tilde{x} \neq \tilde{y}$, then f is strictly convex (Definition 0.9).

i Proof

Proof. Only if. Let $\tilde{x}, \tilde{y} \in \mathbb{R}^p$, let $\tilde{d} \stackrel{\text{def}}{=} \tilde{x} - \tilde{y}$, and let $g(t) \stackrel{\text{def}}{=} f(\tilde{y} + t\tilde{d})$. The inner function $t \mapsto \tilde{y} + t\tilde{d}$ has derivative $\tilde{d}^\top$ (**vector-valued derivative**), so the vector chain rule (**vector chain rule**) gives

$$\begin{aligned}
g'(0) &= \tilde{d}^\top \frac{\partial}{\partial \tilde{x}} f(\tilde{y}) && \text{(vector chain rule)} \\
&= \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top \tilde{d}. && \text{(a } 1 \times 1 \text{ matrix equals its transpose)}
\end{aligned}$$

For $t \in (0, 1]$, $\tilde{y} + t\tilde{d} = t\tilde{x} + (1 - t)\tilde{y}$, so

$$\begin{aligned}
g(t) - g(0) &= f(t\tilde{x} + (1 - t)\tilde{y}) - f(\tilde{y}) && \text{(definition of } g) \\
&\leq tf(\tilde{x}) + (1 - t)f(\tilde{y}) - f(\tilde{y}) && (f \text{ is convex}) \\
&= t(f(\tilde{x}) - f(\tilde{y})), && \text{(collect terms)}
\end{aligned}$$

and dividing by $t > 0$, $\frac{g(t) - g(0)}{t} \leq f(\tilde{x}) - f(\tilde{y})$. As $t \rightarrow 0$ from the right, the left side tends to $g'(0)$ (g is differentiable at 0, so the one-sided limit equals the two-sided derivative), and a limit of numbers that are all at most $f(\tilde{x}) - f(\tilde{y})$ is at most $f(\tilde{x}) - f(\tilde{y})$. So $\left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top (\tilde{x} - \tilde{y}) \leq f(\tilde{x}) - f(\tilde{y})$, which rearranges to Equation 1.

If. Let $\tilde{x}, \tilde{y} \in \mathbb{R}^p$, $t \in [0, 1]$, $\tilde{z} \stackrel{\text{def}}{=} t\tilde{x} + (1 - t)\tilde{y}$, and $\tilde{g} \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{z})$. Applying Equation 1 at

the point $\tilde{z}$, once toward $\tilde{x}$ and once toward $\tilde{y}$,

$$f(\tilde{x}) \geq f(\tilde{z}) + g^\top(\tilde{x} - \tilde{z}), \quad f(\tilde{y}) \geq f(\tilde{z}) + g^\top(\tilde{y} - \tilde{z}).$$

Multiplying the first by $t \geq 0$ and the second by $1 - t \geq 0$ and adding,

$$\begin{aligned} tf(\tilde{x}) + (1 - t)f(\tilde{y}) &\geq t(f(\tilde{z}) + g^\top(\tilde{x} - \tilde{z})) + (1 - t)(f(\tilde{z}) + g^\top(\tilde{y} - \tilde{z})) && \text{(multiply by } t, 1 - t \geq 0 \text{ and add)} \\ &= (t + (1 - t))f(\tilde{z}) + t g^\top(\tilde{x} - \tilde{z}) + (1 - t) g^\top(\tilde{y} - \tilde{z}) && \text{(distribute } t \text{ and } 1 - t) \\ &= f(\tilde{z}) + t g^\top(\tilde{x} - \tilde{z}) + (1 - t) g^\top(\tilde{y} - \tilde{z}) && (t + (1 - t) = 1) \\ &= f(\tilde{z}) + g^\top(t(\tilde{x} - \tilde{z}) + (1 - t)(\tilde{y} - \tilde{z})) && \text{(factor out } g^\top; \text{ matrix products distributive)} \\ &= f(\tilde{z}) + g^\top(t\tilde{x} + (1 - t)\tilde{y} - \tilde{z}) && \text{(collect terms)} \\ &= f(\tilde{z}). && \text{(definition of } \tilde{z}) \end{aligned}$$

That inequality is the defining inequality of a **convex function**. If Equation 1 is strict for distinct points, and $\tilde{x} \neq \tilde{y}$ and $t \in (0, 1)$, then:

- $\tilde{z}$ differs from both $\tilde{x}$ and $\tilde{y}$, since $\tilde{x} - \tilde{z} = (1 - t)(\tilde{x} - \tilde{y}) \neq 0$ and $\tilde{y} - \tilde{z} = t(\tilde{y} - \tilde{x}) \neq 0$;
- so both inequalities at $\tilde{z}$ are strict;
- both multipliers t and $1 - t$ are positive, so the first step of the display is strict, and f is strictly convex.

□

Example 0.26 (Tangent lines under a parabola). For $f(x) = x^2$, the right side of Equation 1 is $y^2 + 2y(x - y) = 2xy - y^2$, and

$$\begin{aligned} f(x) - (2xy - y^2) &= x^2 - 2xy + y^2 \\ &= (x - y)^2 \\ &\geq 0, \end{aligned}$$

with equality only at $x = y$, confirming that x^2 is strictly convex. For $g(x) = -x^2$ the inequality fails: at $y = 0$ the tangent line is $s = 0$, and $g(1) = -1 < 0$.

Theorem 0.11 (Second-derivative test for convexity). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first and second partial derivatives that are continuous on $\mathbb{R}^p$ (**continuity**).*

1. If $\mathbf{H}_f(\tilde{x})$ is **positive semidefinite** for every $\tilde{x}$, then f is **convex**.
2. If $\mathbf{H}_f(\tilde{x})$ is **positive definite** for every $\tilde{x}$, then f is strictly convex (Definition 0.9).

i Proof

Proof. Let $\tilde{x}, \tilde{y} \in \mathbb{R}^p$. By Theorem 0.5 with $\tilde{z} = \tilde{y}$ and $\tilde{h} = \tilde{x} - \tilde{y}$, there is a $\tau \in (0, 1)$ such that, with $\tilde{w} \stackrel{\text{def}}{=} \tilde{y} + \tau(\tilde{x} - \tilde{y})$,

$$\begin{aligned} f(\tilde{x}) &= f(\tilde{y}) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top (\tilde{x} - \tilde{y}) + \frac{1}{2} (\tilde{x} - \tilde{y})^\top \mathbf{H}_f(\tilde{w}) (\tilde{x} - \tilde{y}) \quad (\text{Theorem 0.5}) \\ &\geq f(\tilde{y}) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{y}) \right)^\top (\tilde{x} - \tilde{y}). \end{aligned} \quad (\mathbf{H}_f(\tilde{w}) \text{ is positive semidefinite})$$

This inequality is Equation 1, so f is convex by Theorem 0.10. Under the hypothesis of part 2 and with $\tilde{x} \neq \tilde{y}$, the quadratic-form term is positive, so the inequality is strict, and f is strictly convex by the last sentence of Theorem 0.10. $\square$

Example 0.27 (Using the Hessian to show convexity).

- For a symmetric positive semidefinite $\mathbf{A}$, $f(\tilde{x}) = \tilde{x}^\top \mathbf{A} \tilde{x}$ has Hessian $2\mathbf{A}$ (Hessian of a quadratic form), which is positive semidefinite because $\tilde{h}^\top (2\mathbf{A}) \tilde{h} = 2 \tilde{h}^\top \mathbf{A} \tilde{h} \geq 0$; so f is convex.
- $f(\tilde{x}) = x_1^2 + x_2^2 - 2x_1$ has Hessian $2\mathbf{I}_2$ (Example 0.13), which is positive definite, so f is strictly convex.
- $f(\tilde{x}) = e^{2x_1+x_2} - x_1$ has Hessian $e^u \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$ with $u = 2x_1 + x_2$ (Example), and

$$\begin{aligned} \tilde{h}^\top \mathbf{H}_f(\tilde{x}) \tilde{h} &= e^u (4h_1^2 + 4h_1h_2 + h_2^2) \\ &= e^u (2h_1 + h_2)^2 \\ &\geq 0, \end{aligned}$$

so f is convex. The Hessian is not positive definite ($\tilde{h} = (1, -2)^\top$ gives 0), so part 2 does not apply, and in fact f is not strictly convex: along $\tilde{x} = t(1, -2)^\top$,

$$\begin{aligned} u &= 2t - 2t \\ &= 0 \end{aligned}$$

and $f = 1 - t$ is linear in t , so the convexity inequality holds with equality there.

Example 0.28 (Strictly convex with a singular Hessian). Part 2 is sufficient but not necessary. $f(x) = x^4$ has

$$\begin{aligned} f''(0) &= 12 \cdot 0^2 \\ &= 0, \end{aligned}$$

so $\mathbf{H}_f(0) = [0]$ is **singular** and not positive definite. Yet f is strictly convex, by the strict form of Equation 1: for $x \neq y$,

$$\begin{aligned} x^4 - y^4 - 4y^3(x - y) &= (x - y)(x^3 + x^2y + xy^2 + y^3) - 4y^3(x - y) && \text{(factor } x^4 - y^4\text{)} \\ &= (x - y)(x^3 + x^2y + xy^2 - 3y^3) && \text{(collect the } y^3 \text{ terms)} \\ &= (x - y)^2(x^2 + 2xy + 3y^2) && ((x - y)(x^2 + 2xy + 3y^2) = x^3 + x^2y + xy^2 - 3y^3) \\ &= (x - y)^2((x + y)^2 + 2y^2), && \text{(complete the square)} \end{aligned}$$

which is positive: $(x - y)^2 > 0$, and $(x + y)^2 + 2y^2 = 0$ would need $y = 0$ and then

$$\begin{aligned} x &= 0 \\ &= y. \end{aligned}$$

Corollary 0.1 (Stationary points of a convex function are global minimizers). *If f is as in Theorem 0.10 and convex, then every stationary point $\tilde{x}^*$ of f (Definition 0.3) is a global minimizer.*

i Proof

Proof. For every $\tilde{x}$,

$$\begin{aligned} f(\tilde{x}) &\geq f(\tilde{x}^*) + \left(\frac{\partial}{\partial \tilde{x}} f(\tilde{x}^*) \right)^\top (\tilde{x} - \tilde{x}^*) && \text{(Equation 1 with } \tilde{y} = \tilde{x}^*) \\ &= f(\tilde{x}^*) + \tilde{0}^\top (\tilde{x} - \tilde{x}^*) && (\tilde{x}^* \text{ is stationary)} \\ &= f(\tilde{x}^*). && (\tilde{0} \cdot \tilde{v} = 0) \end{aligned}$$

□

Example 0.29 (A convex function, and a non-convex one).

- $f(\tilde{x}) = x_1^2 + x_2^2 - 2x_1$ is convex by Theorem 0.11, since its Hessian is $2\mathbf{I}_2$ (Example 0.13), so its stationary point $(1, 0)^\top$ (Example 0.5) is a global minimizer, as completing the square showed in Example 0.6.
- Without convexity, a stationary point need not be a minimizer at all: $x_1^2 - x_2^2$ has the stationary point $\tilde{0}$, which is not a minimizer (Example 0.6).
- The corollary says nothing about whether a minimizer exists: $e^{2x_1+x_2} - x_1$ is convex (Example 0.27) but has no stationary point, and so no local minimizer (Example 0.6).

Minimizing a quadratic under linear constraints

Source

This section is adapted from the Lagrange-multiplier part of Zhou [3], used under the MIT License (see the license text in [Linear Algebra](#)). The source solves the stationarity equations of the Lagrangian and reports the solution; here the solution is also proved to be the unique minimizer, directly from the positive definiteness of the matrices involved. The source assumes only that $\mathbf{S}$ is positive definite; the added hypothesis that the columns of $\mathbf{A}$ are linearly independent makes $\mathbf{A}^\top \mathbf{S}^{-1} \mathbf{A}$ invertible. The source's example with general coefficients a_1, a_2, b is specialized here to 3, 4, 5. These parts of the source are not part of this section:

- the Lagrange multiplier method for constraints that are not linear, or objectives that are not quadratic, including its multinomial maximum likelihood example
- the eigenvalues of the block matrix of the stationarity equations, and the convexity of the Lagrangian in $\tilde{x}$ and its concavity in the multipliers
- its homework examples

Definition 0.14 (Feasible set). In a problem that looks for a best point among some allowed points of $\mathbb{R}^p$, the set $F \subseteq \mathbb{R}^p$ of allowed points is the **feasible set**, [11] and its points are feasible.

Definition 0.15 (Constraints). When the feasible set F (Definition 0.14) is the set of points that satisfy some equations or inequalities, those equations or inequalities are the **constraints**.

Example 0.30 (A line as a feasible set). With the single constraint $3x_1 + 4x_2 = 5$, the feasible set is the line $\{\tilde{x} \in \mathbb{R}^2 : 3x_1 + 4x_2 = 5\}$. $(1, 0.5)^\top$ is feasible, since

$$\begin{aligned} 3 \cdot 1 + 4 \cdot 0.5 &= 3 + 2 \\ &= 5; \end{aligned}$$

$\mathbf{0}$ is not feasible, since $3 \cdot 0 + 4 \cdot 0 = 0 \neq 5$.

Definition 0.16 (Objective function, objective). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ and let $F \subseteq \mathbb{R}^p$ be a feasible set (Definition 0.14). In the problem of minimizing f over F , f is the **objective function** (or **objective**).

Definition 0.17 (Minimizer over a feasible set). With f and F as in Definition 0.16, a **minimizer of f over F** is a feasible $\tilde{x}^*$ with $f(\tilde{x}^*) \leq f(\tilde{x})$ for every feasible $\tilde{x}$: a **global minimizer** of f with its domain restricted to F .

Definition 0.18 (Minimization problem). With f and F as in Definition 0.16, the **minimization problem** [12] is to find a minimizer of f over F (Definition 0.17). When F is given by constraints (Definition 0.15), the problem is to minimize f subject to them.

Example 0.31 (Minimizing over a line). In the problem of minimizing $x_1^2 + x_2^2$ subject to $3x_1 + 4x_2 = 5$, the objective function is $f(x) = x_1^2 + x_2^2$, and the feasible set is the line of Example 0.30. The feasible point $(1, 0.5)^\top$ has

$$\begin{aligned} f(1, 0.5) &= 1 + 0.25 \\ &= 1.25. \end{aligned}$$

$\mathbf{0}$ minimizes f over all of $\mathbb{R}^2$, with $f(\mathbf{0}) = 0$, but it is not a minimizer over the feasible set, because it is not feasible.

Definition 0.19 (Optimization algorithm). Let:

- $f : \mathbb{R}^p \rightarrow \mathbb{R}$: an **objective function**;
- $F \subseteq \mathbb{R}^p$: a **feasible set**.

An **optimization algorithm** [12] is a step-by-step procedure that **maps** the pair (f, F) to a feasible point $\tilde{x}^\dagger$:

$$(f, F) \mapsto \tilde{x}^\dagger \in F. \quad (2)$$

In words: given (f, F) , the procedure returns a feasible point $\tilde{x}^\dagger$ intended to make $f(\tilde{x}^\dagger)$ small.

Example 0.32 (Exhaustive search over a finite feasible set). Let $f(\tilde{x}) = x_1^2 + x_2^2$, with the **feasible set** the three points $\{(1, 0.5)^\top, (3, -1)^\top, (-1, 2)^\top\}$ of the line of Example 0.30. The procedure “evaluate f at every feasible point; return the point with the smallest **value**” is an **optimization algorithm**. Computing f at each point gives

$$\begin{aligned} f(1, 0.5) &= 1 + 0.25 = 1.25, \\ f(3, -1) &= 9 + 1 = 10, \\ f(-1, 2) &= 1 + 4 = 5, \end{aligned}$$

so it returns $(1, 0.5)^\top$, a **minimizer of f over this feasible set**.

Here the objective (Definition 0.16) is a quadratic $\frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x}$ with $\mathbf{S}$ positive definite, and the constraints (Definition 0.15) are m linear equations $\mathbf{A}^\top \tilde{x} = \tilde{b}$, where $\mathbf{A}$ is a $p \times m$ matrix and $\tilde{b} \in \mathbb{R}^m$. The only stationary point of the objective alone (Definition 0.3) is $\mathbf{0}_{p \times 1}$ (its gradient is $\mathbf{S} \tilde{x}$, and $\mathbf{S}$ is invertible), which may not satisfy the constraints. A Lagrange multiplier adds one unknown per constraint, and the minimizer over the feasible set (Definition 0.14) is then found among the stationary points of a new function of the extended set of unknowns.

Definition 0.20 (Lagrangian). For the problem of minimizing $f(\tilde{x})$ over $\tilde{x} \in \mathbb{R}^p$ subject to $\mathbf{A}^\top \tilde{x} = \tilde{b}$, with $\mathbf{A}$ a $p \times m$ matrix and $\tilde{b} \in \mathbb{R}^m$, the **Lagrangian** [13] is the function of $\tilde{x} \in \mathbb{R}^p$ and $\tilde{\lambda} \in \mathbb{R}^m$

$$L(\tilde{x}, \tilde{\lambda}) \stackrel{\text{def}}{=} f(\tilde{x}) + \tilde{\lambda}^\top (\mathbf{A}^\top \tilde{x} - \tilde{b}).$$

Definition 0.21 (Lagrange multipliers). In the Lagrangian $L(\tilde{x}, \tilde{\lambda})$ (Definition 0.20), the entries of $\tilde{\lambda}$ are the **Lagrange multipliers** [13], one per constraint.

Example 0.33 (The gradients of the Lagrangian of a quadratic). Let $f(\tilde{x}) = \frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x}$ with $\mathbf{S}$ symmetric and constant. By the [transpose of a product](#), $\tilde{\lambda}^\top \mathbf{A}^\top = (\mathbf{A} \tilde{\lambda})^\top$, so

$$\begin{aligned} L(\tilde{x}, \tilde{\lambda}) &= \frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x} + \tilde{\lambda}^\top \mathbf{A}^\top \tilde{x} - \tilde{\lambda}^\top \tilde{b} \quad (\text{distribute}) \\ &= \frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x} + (\mathbf{A} \tilde{\lambda})^\top \tilde{x} - \tilde{\lambda}^\top \tilde{b}, \quad (\text{transpose of a product}) \end{aligned}$$

and, since $\tilde{\lambda}^\top (\mathbf{A}^\top \tilde{x} - \tilde{b})$ is 1×1 and so equals its transpose,

$$L(\tilde{x}, \tilde{\lambda}) = \frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x} + (\mathbf{A}^\top \tilde{x} - \tilde{b})^\top \tilde{\lambda}.$$

The partial derivatives of a sum are the sums of the partial derivatives, entry by entry. Using the first of these forms, [derivative of a quadratic form](#) for the first term, [derivative of a linear combination](#) for the second ($\mathbf{A} \tilde{\lambda}$ is constant in $\tilde{x}$), and [constant with respect to a vector](#) for the third,

$$\begin{aligned} \frac{\partial}{\partial \tilde{x}} L(\tilde{x}, \tilde{\lambda}) &= \frac{1}{2} \frac{\partial}{\partial \tilde{x}} (\tilde{x}^\top \mathbf{S} \tilde{x}) + \frac{\partial}{\partial \tilde{x}} ((\mathbf{A} \tilde{\lambda})^\top \tilde{x}) - \frac{\partial}{\partial \tilde{x}} (\tilde{\lambda}^\top \tilde{b}) \quad (\text{sum and constant multiple}) \\ &= \frac{1}{2} (2 \mathbf{S} \tilde{x}) + \mathbf{A} \tilde{\lambda} - \mathbf{0}_{p \times 1} \quad (\text{the three derivatives}) \\ &= \mathbf{S} \tilde{x} + \mathbf{A} \tilde{\lambda}. \quad (\text{simplify}) \end{aligned}$$

Using the second form, in which $f(\tilde{x})$ and $\mathbf{A}^\top \tilde{x} - \tilde{b}$ are constant in $\tilde{\lambda}$, and [derivative of a linear combination](#) again,

$$\begin{aligned}
\frac{\partial}{\partial \tilde{\lambda}} L(\tilde{x}, \tilde{\lambda}) &= \frac{\partial}{\partial \tilde{\lambda}} f(\tilde{x}) + \frac{\partial}{\partial \tilde{\lambda}} ((\mathbf{A}^\top \tilde{x} - \tilde{b})^\top \tilde{\lambda}) \quad (\text{sum}) \\
&= \mathbf{0}_{m \times 1} + (\mathbf{A}^\top \tilde{x} - \tilde{b}) \quad (\text{the two derivatives}) \\
&= \mathbf{A}^\top \tilde{x} - \tilde{b}. \quad (\text{simplify})
\end{aligned}$$

So $(\tilde{x}, \tilde{\lambda})$ makes the gradients equal $\mathbf{0}_{p \times 1}$ and $\mathbf{0}_{m \times 1}$ exactly when

$$\mathbf{S} \tilde{x} + \mathbf{A} \tilde{\lambda} = \mathbf{0}_{p \times 1} \quad \text{and} \quad \mathbf{A}^\top \tilde{x} = \tilde{b}, \quad \text{that is,} \quad \underbrace{\begin{bmatrix} \mathbf{S} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0}_{m \times m} \end{bmatrix}}_{(p+m) \times (p+m)} \underbrace{\begin{bmatrix} \tilde{x} \\ \tilde{\lambda} \end{bmatrix}}_{(p+m) \times 1} = \underbrace{\begin{bmatrix} \mathbf{0}_{p \times 1} \\ \tilde{b} \end{bmatrix}}_{(p+m) \times 1}, \quad (3)$$

the second equation being the constraint itself.

Theorem 0.12 (Minimizing a positive definite quadratic under linear constraints). *Let $\mathbf{S}$ be a $p \times p$ [positive definite](#) matrix, $\mathbf{A}$ a $p \times m$ matrix with [linearly independent](#) columns, and $\tilde{b} \in \mathbb{R}^m$. Then the $m \times m$ matrix $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{A}^\top \mathbf{S}^{-1} \mathbf{A}$ is invertible, and $f(\tilde{x}) = \frac{1}{2} \tilde{x}^\top \mathbf{S} \tilde{x}$ has exactly one minimizer over the $\tilde{x} \in \mathbb{R}^p$ with $\mathbf{A}^\top \tilde{x} = \tilde{b}$, namely*

$$\tilde{x}^* = \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b}.$$

With $\tilde{\lambda}^* \stackrel{\text{def}}{=} -\mathbf{M}^{-1} \tilde{b}$, the pair $(\tilde{x}^*, \tilde{\lambda}^*)$ solves Equation 3, and the minimum value is $f(\tilde{x}^*) = \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \tilde{b}$.

i Proof

Proof. The steps marked “transpose of a product” use the [transpose of a product](#), and $\mathbf{S}$ is symmetric because it is [positive definite](#).

M is invertible. $\mathbf{S}^{-1}$ exists and is positive definite ([positive definite inverse](#)), so $\mathbf{M} = \mathbf{A}^\top \mathbf{S}^{-1} \mathbf{A}$ is positive definite ([operations that preserve definiteness](#), part 1), and so invertible (positive definite inverse again). $\mathbf{S}^{-1}$ and $\mathbf{M}^{-1}$ are symmetric ([inverse of a symmetric matrix](#)), because $\mathbf{S}$ and $\mathbf{M}$ are.

$\tilde{x}^*$ satisfies the constraint, and the pair solves Equation 3.

$$\begin{aligned}
\mathbf{A}^\top \tilde{x}^* &= \mathbf{A}^\top \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\text{substitute } \tilde{x}^*) \\
&= \mathbf{M} \mathbf{M}^{-1} \tilde{b} && (\text{definition of } \mathbf{M}) \\
&= \tilde{b}, && (\mathbf{M} \mathbf{M}^{-1} = \mathbf{I}_m) \\
\mathbf{S} \tilde{x}^* + \mathbf{A} \tilde{\lambda}^* &= \mathbf{S} \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} - \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\text{substitute } \tilde{x}^* \text{ and } \tilde{\lambda}^*) \\
&= \mathbf{A} \mathbf{M}^{-1} \tilde{b} - \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\mathbf{S} \mathbf{S}^{-1} = \mathbf{I}_p) \\
&= \mathbf{0}_{p \times 1}. && (\text{subtract})
\end{aligned}$$

$\tilde{x}^*$ is the unique minimizer. Let $\tilde{x}$ satisfy $\mathbf{A}^\top \tilde{x} = \tilde{b}$, and let $\tilde{d} \stackrel{\text{def}}{=} \tilde{x} - \tilde{x}^*$, so

$$\begin{aligned}\mathbf{A}^\top \tilde{d} &= \mathbf{A}^\top \tilde{x} - \mathbf{A}^\top \tilde{x}^* \\ &= \tilde{b} - \tilde{b} \\ &= \mathbf{0}_{m \times 1}\end{aligned}$$

(matrix products distribute; both points satisfy the constraint). First, the cross term vanishes:

$$\begin{aligned}\tilde{d}^\top \mathbf{S} \tilde{x}^* &= \tilde{d}^\top \mathbf{S} \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} \quad (\text{substitute } \tilde{x}^*) \\ &= \tilde{d}^\top \mathbf{A} \mathbf{M}^{-1} \tilde{b} \quad (\mathbf{S} \mathbf{S}^{-1} = \mathbf{I}_p) \\ &= (\mathbf{A}^\top \tilde{d})^\top \mathbf{M}^{-1} \tilde{b} \quad (\text{transpose of a product; } (\mathbf{A}^\top)^\top = \mathbf{A}) \\ &= \mathbf{0}_{m \times 1}^\top \mathbf{M}^{-1} \tilde{b} \quad (\mathbf{A}^\top \tilde{d} = \mathbf{0}_{m \times 1}) \\ &= 0. \quad (\text{a product with a zero factor is zero})\end{aligned}$$

Then

$$\begin{aligned}f(\tilde{x}) &= \frac{1}{2} (\tilde{x}^* + \tilde{d})^\top \mathbf{S} (\tilde{x}^* + \tilde{d}) && (\tilde{x} = \tilde{x}^* + \tilde{d}) \\ &= \frac{1}{2} (\tilde{x}^{*\top} + \tilde{d}^\top) \mathbf{S} (\tilde{x}^* + \tilde{d}) && (\text{the transpose of a sum is the sum of the transposes}) \\ &= \frac{1}{2} (\tilde{x}^{*\top} \mathbf{S} \tilde{x}^* + \tilde{x}^{*\top} \mathbf{S} \tilde{d} + \tilde{d}^\top \mathbf{S} \tilde{x}^* + \tilde{d}^\top \mathbf{S} \tilde{d}) && (\text{distribute}) \\ &= \frac{1}{2} (\tilde{x}^{*\top} \mathbf{S} \tilde{x}^* + \tilde{d}^\top \mathbf{S}^\top \tilde{x}^* + \tilde{d}^\top \mathbf{S} \tilde{x}^* + \tilde{d}^\top \mathbf{S} \tilde{d}) && (1 \times 1: \tilde{x}^{*\top} \mathbf{S} \tilde{d} = (\tilde{x}^{*\top} \mathbf{S} \tilde{d})^\top; \text{transpose of a product}) \\ &= \frac{1}{2} (\tilde{x}^{*\top} \mathbf{S} \tilde{x}^* + 2 \tilde{d}^\top \mathbf{S} \tilde{x}^* + \tilde{d}^\top \mathbf{S} \tilde{d}) && (\mathbf{S}^\top = \mathbf{S}; \text{combine}) \\ &= \frac{1}{2} \tilde{x}^{*\top} \mathbf{S} \tilde{x}^* + \tilde{d}^\top \mathbf{S} \tilde{x}^* + \frac{1}{2} \tilde{d}^\top \mathbf{S} \tilde{d} && (\text{distribute } \frac{1}{2}) \\ &= f(\tilde{x}^*) + \tilde{d}^\top \mathbf{S} \tilde{x}^* && (\text{the cross term is 0}) \\ &\geq f(\tilde{x}^*), && (\mathbf{S} \text{ is positive definite})\end{aligned}$$

with equality only when $\tilde{d} = \mathbf{0}_{p \times 1}$, that is, $\tilde{x} = \tilde{x}^*$.

The minimum value.

$$\begin{aligned}f(\tilde{x}^*) &= \frac{1}{2} (\mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b})^\top \mathbf{S} \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\text{substitute } \tilde{x}^*) \\ &= \frac{1}{2} \tilde{b}^\top (\mathbf{M}^{-1})^\top \mathbf{A}^\top (\mathbf{S}^{-1})^\top \mathbf{S} \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\text{transpose of a product, applied repeatedly}) \\ &= \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \mathbf{A}^\top \mathbf{S}^{-1} \mathbf{S} \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\mathbf{S}^{-1} \text{ and } \mathbf{M}^{-1} \text{ are symmetric}) \\ &= \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \mathbf{A}^\top \mathbf{S}^{-1} \mathbf{A} \mathbf{M}^{-1} \tilde{b} && (\mathbf{S}^{-1} \mathbf{S} = \mathbf{I}_p) \\ &= \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \mathbf{M} \mathbf{M}^{-1} \tilde{b} && (\text{definition of } \mathbf{M}) \\ &= \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \tilde{b}. && (\mathbf{M}^{-1} \mathbf{M} = \mathbf{I}_m)\end{aligned}$$

□

Example 0.34 (The closest point on a line to the origin). Minimize $x_1^2 + x_2^2$ subject to $3x_1 + 4x_2 = 5$. In the form of Theorem 0.12, $x_1^2 + x_2^2 = \frac{1}{2} \tilde{x}^\top (2\mathbf{I}_2) \tilde{x}$, so $\mathbf{S} = 2\mathbf{I}_2$, $\mathbf{A} = (3, 4)^\top$ ($p = 2$, $m = 1$) and $b = [5]$. The hypotheses hold: $2\mathbf{I}_2$ is positive definite, and the single column $(3, 4)^\top$ is nonzero, so it is linearly independent. Also $\mathbf{S}^{-1} = \frac{1}{2}\mathbf{I}_2$. Then, from the formulas of Theorem 0.12,

$$\begin{aligned}
\mathbf{M} &= \mathbf{A}^\top \frac{1}{2} \mathbf{I}_2 \mathbf{A} \\
&= \frac{1}{2} (3^2 + 4^2) \\
&= \frac{25}{2}, \\
\tilde{x}^* &= \frac{1}{2} \mathbf{I}_2 \begin{bmatrix} 3 \\ 4 \end{bmatrix} \left(\frac{2}{25} \right) (5) \\
&= \frac{5}{25} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\
&= \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix}, \\
\tilde{\lambda}^* &= -\left(\frac{2}{25} \right) (5) \\
&= -0.4, \\
f(\tilde{x}^*) &= \frac{1}{2} (5) \left(\frac{2}{25} \right) (5) \\
&= 1.
\end{aligned}$$

Checks:

- the constraint:

$$\begin{aligned}
3(0.6) + 4(0.8) &= 1.8 + 3.2 \\
&= 5;
\end{aligned}$$

- the first block of Equation 3:

$$\begin{aligned}
\mathbf{S} \tilde{x}^* + \mathbf{A} \tilde{\lambda}^* &= (1.2, 1.6)^\top + (3, 4)^\top (-0.4) \\
&= (1.2, 1.6)^\top - (1.2, 1.6)^\top \\
&= \mathbf{0}_{2 \times 1};
\end{aligned}$$

- the value:

$$\begin{aligned}
0.6^2 + 0.8^2 &= 0.36 + 0.64 \\
&= 1.
\end{aligned}$$

Geometrically, $\tilde{x}^*$ is the point of the line nearest the origin, at distance 1.

Example 0.35 (The hypotheses are needed).

- If $\mathbf{S}$ is not positive definite there may be no minimizer: with $\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and the constraint $x_1 = 1$ ($\mathbf{A} = (1, 0)^\top$, $\tilde{b} = [1]$), $f(1, x_2) = \frac{1}{2}(1 - x_2^2)$ decreases without bound as x_2 grows.
- If the columns of $\mathbf{A}$ are dependent, $\mathbf{M}$ is not invertible: with $\mathbf{S} = \mathbf{I}_2$ and $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$,

$$\begin{aligned}\mathbf{M} &= \mathbf{A}^\top \mathbf{A} \\ &= \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix},\end{aligned}$$

whose columns are equal. So $\mathbf{M}(1, -1)^\top = \mathbf{0}_{2 \times 1}$, and $\mathbf{M}$ is not invertible (a square matrix is invertible exactly when its null space is zero). The two constraints both read $x_1 + x_2 = b_i$, so they have no solution at all when $b_1 \neq b_2$. When $b_1 = b_2$ a unique minimizer still exists,

$$\begin{aligned}x_1 &= x_2 \\ &= b_1/2,\end{aligned}$$

but the formula of Theorem 0.12 cannot produce it.

Definition 0.22 (Sensitivity). Suppose a minimization problem (Definition 0.18) depends on a vector $\tilde{b} \in \mathbb{R}^m$, and let $f^*(\tilde{b})$ be its minimum value, as a function of $\tilde{b}$. When the gradient $\frac{\partial}{\partial \tilde{b}} f^*(\tilde{b})$ exists (gradient), it is the **sensitivity** of the minimum value to $\tilde{b}$: by partial derivative, its entry i is the limit, as a change in b_i alone shrinks to 0, of the resulting change in the minimum value divided by the change in b_i .

Example 0.36 (The sensitivity of a one-point problem). Minimize x^2 subject to $x = b$, for a number b . The feasible set (Definition 0.14) is $\{b\}$, so the minimum value is $f^*(b) = b^2$, and its sensitivity to b is $\frac{d}{db} f^*(b) = 2b$. At $b = 3$ the sensitivity is 6: moving b from 3 to 3.1 raises the minimum value from 9 to 9.61, an increase of 0.61, close to $6 \times 0.1 = 0.6$.

Corollary 0.2 (The multiplier is the sensitivity of the minimum value). In Theorem 0.12, write the minimum value as a function of $\tilde{b}$, $f^*(\tilde{b}) = \frac{1}{2} \tilde{b}^\top \mathbf{M}^{-1} \tilde{b}$. Its sensitivity to $\tilde{b}$ (Definition 0.22) is

$$\begin{aligned}\frac{\partial}{\partial \tilde{b}} f^*(\tilde{b}) &= \mathbf{M}^{-1} \tilde{b} \\ &= -\tilde{\lambda}^*.\end{aligned}$$

i Proof

Proof. $\mathbf{M}^{-1}$ is symmetric and does not depend on $\tilde{b}$, so [derivative of a quadratic form](#) applies with $\mathbf{M}^{-1}$ in place of $\mathbf{S}$:

$$\begin{aligned}\frac{\partial}{\partial \tilde{b}} f^*(\tilde{b}) &= \frac{1}{2} \frac{\partial}{\partial \tilde{b}} (\tilde{b}^\top \mathbf{M}^{-1} \tilde{b}) \quad (\text{constant multiple}) \\ &= \frac{1}{2} (2\mathbf{M}^{-1} \tilde{b}) \quad (\text{derivative of a quadratic form}) \\ &= \mathbf{M}^{-1} \tilde{b} \quad (\text{simplify}) \\ &= -\tilde{\lambda}^*. \quad (\text{definition of } \tilde{\lambda}^*)\end{aligned}$$

□

Example 0.37 (Moving the line). In Example [0.34](#) with a general right side b in place of 5, $\mathbf{M}^{-1} = \frac{2}{25}$ and

$$\begin{aligned}f^*(b) &= \frac{1}{2} \left(\frac{2}{25} \right) b^2 \\ &= \frac{b^2}{25},\end{aligned}$$

so $\frac{d}{db} f^*(b) = \frac{2b}{25}$. At $b = 5$ this derivative is

$$\begin{aligned}\frac{10}{25} &= 0.4 \\ &= -\tilde{\lambda}^*:\end{aligned}$$

moving the line $3x_1 + 4x_2 = b$ outward by a small step ε in b raises the minimum value by about 0.4ε .

Newton's method and gradient descent

i Source

This section is adapted from the Newton-Raphson and gradient descent parts of Zhou [3], used under the MIT License (see the license text in [Linear Algebra](#)). The source motivates the Newton step by minimizing a quadratic approximation; here that step is proved to be the exact minimizer of the quadratic model. In the source's zigzag example the starting value of the objective is $\frac{1}{2}(b^2 + b)$, not $\frac{1}{2}(1 + b^2)$ as the source writes; the

ratio of successive values is unaffected. In the same example the source's formula for $\tilde{x}^{(1)}$ has $\tilde{x}^{(1)}$ on its right side where $\tilde{x}^{(0)}$ is meant. These parts of the source are not part of this section:

- the quadratic convergence theorem for Newton's method in general, and its remedies for instability (a positive definite substitute for the Hessian, and shortening the Newton step by a search along its direction)
- its remark that Hessians are expensive to compute
- the convergence-rate theorem for strongly convex functions
- gradient descent with momentum, and its optimal parameters
- Nesterov acceleration, and stochastic gradient descent
- its code and animations

Definition 0.23 (Iterative method). An **iterative method** [14] starts from a point $\tilde{x}^{(0)}$ and computes $\tilde{x}^{(t+1)}$ from $\tilde{x}^{(t)}$ by a fixed rule, for $t = 0, 1, 2, \dots$

Definition 0.24 (Iterate). In an iterative method (Definition 0.23), the points $\tilde{x}^{(t)}$ are its **iterates**.

Definition 0.25 (Error of an iterate). For a point $\tilde{x}^*$ that an iterative method (Definition 0.23) is meant to reach, $\|\tilde{x}^{(t)} - \tilde{x}^*\|$ is the **error** of the iterate $\tilde{x}^{(t)}$ (Definition 0.24).

Definition 0.26 (Convergence of a sequence of vectors). A **sequence** $\tilde{x}^{(0)}, \tilde{x}^{(1)}, \dots$ in $\mathbb{R}^p$ **converges** to $\tilde{x}^* \in \mathbb{R}^p$ if the sequence of distances $\|\tilde{x}^{(t)} - \tilde{x}^*\|$ **converges** to 0. For $p = 1$, $\|\tilde{x}^{(t)} - \tilde{x}^*\| = |\tilde{x}^{(t)} - \tilde{x}^*|$, and this is the convergence of a sequence of numbers to $\tilde{x}^*$.

Example 0.38 (Halving a vector). The iterative method $\tilde{x}^{(t+1)} = \frac{1}{2} \tilde{x}^{(t)}$ from $\tilde{x}^{(0)} = (4, 2)^\top$ has iterates $(4, 2)^\top, (2, 1)^\top, (1, 0.5)^\top, \dots$, that is, $\tilde{x}^{(t)} = 2^{-t} (4, 2)^\top$. With $\tilde{x}^* = \mathbf{0}$, the error of $\tilde{x}^{(t)}$ is

$$\begin{aligned} \|2^{-t} (4, 2)^\top - \mathbf{0}\| &= 2^{-t} \|(4, 2)^\top\| \quad (\text{norm of a multiple}) \\ &= 2^{-t} \sqrt{20}, \quad (4^2 + 2^2 = 20) \end{aligned}$$

which converges to 0, so the iterates converge to $\mathbf{0}$.

Definition 0.27 (Newton's method for minimization). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first and second partial derivatives, and let $\tilde{x}^{(0)}$ be a starting point. **Newton's method** [15]

computes, for $t = 0, 1, 2, \dots$, as long as $\mathbf{H}_f(\tilde{x}^{(t)})$ is invertible,

$$\tilde{x}^{(t+1)} \stackrel{\text{def}}{=} \tilde{x}^{(t)} - [\mathbf{H}_f(\tilde{x}^{(t)})]^{-1} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}).$$

Definition 0.28 (Newton step). In Newton's method (Definition 0.27), the vector $-[\mathbf{H}_f(\tilde{x}^{(t)})]^{-1} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)})$ is the **Newton step** at $\tilde{x}^{(t)}$.

Example 0.39 (Newton's method for a cubic). Let $f(x) = \frac{1}{3}x^3 - 4x$, so $f'(x) = x^2 - 4$ and $f''(x) = 2x$. For $x^{(t)} \neq 0$, one Newton step gives

$$\begin{aligned} x^{(t+1)} &= x^{(t)} - \frac{(x^{(t)})^2 - 4}{2x^{(t)}} && \text{(Newton's method with } p = 1) \\ &= \frac{2(x^{(t)})^2 - (x^{(t)})^2 + 4}{2x^{(t)}} && \text{(common denominator)} \\ &= \frac{1}{2} \left(x^{(t)} + \frac{4}{x^{(t)}} \right). && \text{(simplify)} \end{aligned}$$

From $x^{(0)} = 2.5$:

- $$\begin{aligned} x^{(1)} &= \frac{1}{2}(2.5 + 1.6) \\ &= 2.05; \end{aligned}$$
 - $$\begin{aligned} x^{(2)} &= \frac{1}{2}(2.05 + 1.951219\dots) \\ &= 2.000610 \end{aligned}$$
- (to six decimals);
- $x^{(3)} = 2.0000001$ (to seven decimals).

The iterates (Definition 0.24) approach 2, the local minimizer of f (a stationary point, since $f'(2) = 0$, with $f''(2) = 4 > 0$; Theorem 0.6).

Definition 0.29 (Quadratic model, second-order Taylor polynomial). Let f be as in Definition 0.27 and let $\tilde{x}$ be a point, with $g \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{x})$ and $\mathbf{H} \stackrel{\text{def}}{=} \mathbf{H}_f(\tilde{x})$. The **quadratic model** of f at $\tilde{x}$ is the function of a step $\tilde{\delta} \in \mathbb{R}^p$

$$q(\tilde{\delta}) \stackrel{\text{def}}{=} f(\tilde{x}) + g \cdot \tilde{\delta} + \frac{1}{2} \tilde{\delta}^\top \mathbf{H} \tilde{\delta},$$

also called the **second-order Taylor polynomial** of f at $\tilde{x}$.

Example 0.40 (The model at $x = 2.5$). For $f(x) = \frac{1}{3}x^3 - 4x$ at $x = 2.5$ (Example 0.39),

$$\begin{aligned} g &= 2.5^2 - 4 \\ &= 2.25 \end{aligned}$$

and

$$\begin{aligned} H &= [2 \cdot 2.5] \\ &= [5], \end{aligned}$$

so $q(\delta) = f(2.5) + 2.25\delta + 2.5\delta^2$. For $f(\tilde{x}) = \frac{1}{2}(x_1^2 + 3x_2^2)$, whose Hessian is $\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$ at every point, the model at any $\tilde{x}$ equals $f(\tilde{x} + \tilde{\delta})$ exactly (Example 0.9 with $\mathbf{S} = \begin{bmatrix} 1/2 & 0 \\ 0 & 3/2 \end{bmatrix}$).

i Remark (The quadratic model and Taylor's theorem)

Remark 0.3 (The quadratic model and Taylor's theorem). When the second partial derivatives of f are continuous, Theorem 0.5 says that $f(\tilde{x} + \tilde{\delta})$ equals the expression for $q(\tilde{\delta})$ in Definition 0.29 with the Hessian evaluated at a point between $\tilde{x}$ and $\tilde{x} + \tilde{\delta}$.

Theorem 0.13 (The Newton step minimizes the quadratic model). *Let q be the quadratic model of f at $\tilde{x}$ (Definition 0.29), with $\mathbf{H} = \mathbf{H}_f(\tilde{x})$ positive definite. Its unique minimizer is the Newton step $\tilde{\delta}^* = -\mathbf{H}^{-1}g$.*

i Proof

Proof. $\mathbf{H}$ is symmetric, by the definition of positive definite, and invertible (positive definite inverse), so $\tilde{\delta}^*$ exists and $\mathbf{H}\tilde{\delta}^* = -g$. For any $\tilde{d} \in \mathbb{R}^p$, first

$$\begin{aligned} \tilde{d}^\top \mathbf{H}\tilde{\delta}^* &= -\tilde{d} \cdot \tilde{g} \quad (\mathbf{H}\tilde{\delta}^* = -g) \\ &= -g \cdot \tilde{d}, \quad (\text{a } 1 \times 1 \text{ matrix equals its transpose}) \end{aligned}$$

and $\tilde{\delta}^{*\top} \mathbf{H}\tilde{d} = \tilde{d}^\top \mathbf{H}\tilde{\delta}^*$ in the same way, using $\mathbf{H}^\top = \mathbf{H}$. Then

$$\begin{aligned}
q(\delta^* + d) &= f(x) + g \cdot (\delta^* + d) + \frac{1}{2} (\delta^* + d)^\top \mathbf{H} (\delta^* + d) && \text{(definition of } q) \\
&= f(x) + g \cdot \delta^* + g \cdot d + \frac{1}{2} (\delta^{*\top} \mathbf{H} \delta^* + \delta^{*\top} \mathbf{H} d + d^\top \mathbf{H} \delta^* + d^\top \mathbf{H} d) && \text{(distribute)} \\
&= f(x) + g \cdot \delta^* + g \cdot d + \frac{1}{2} \delta^{*\top} \mathbf{H} \delta^* - g \cdot d + \frac{1}{2} d^\top \mathbf{H} d && \text{(both cross terms equal } -g \cdot d) \\
&= f(x) + g \cdot \delta^* + \frac{1}{2} \delta^{*\top} \mathbf{H} \delta^* + \frac{1}{2} d^\top \mathbf{H} d && \text{(cancel } g \cdot d) \\
&= q(\delta^*) + \frac{1}{2} d^\top \mathbf{H} d, && \text{(definition of } q)
\end{aligned}$$

which is greater than $q(\delta^*)$ for every $d \neq 0_{p \times 1}$, since $\mathbf{H}$ is positive definite. $\square$

Example 0.41 (The Newton step at $x = 2.5$). For the model $q(\delta) = f(2.5) + 2.25\delta + 2.5\delta^2$ of Example 0.40, $H = [5]$ is positive definite, so by Theorem 0.13 its minimizer is

$$\begin{aligned}
\delta^* &= -H^{-1}g \\
&= -2.25/5 \\
&= -0.45.
\end{aligned}$$

Setting the derivative $2.25 + 5\delta$ to 0 gives the same δ^* , and

$$\begin{aligned}
2.5 - 0.45 &= 2.05 \\
&= x^{(1)}
\end{aligned}$$

(Example 0.39).

Definition 0.30 (Quadratic convergence). A [sequence](#) $x^{(0)}, x^{(1)}, \dots$ in $\mathbb{R}^p$ that converges (Definition 0.26) to x^* **converges quadratically** [16] if there is a constant C with

$$\|x^{(t+1)} - x^*\| \leq C \|x^{(t)} - x^*\|^2 \quad \text{for every } t.$$

Example 0.42 (Halving the error is not quadratic convergence). $x^{(t)} = 2^{-t}$ converges to 0, but it does not converge quadratically: the bound would need $2^{-(t+1)} \leq C (2^{-t})^2 = C 2^{-2t}$, that is, $2^{t-1} \leq C$, which fails once $t > 1 + \log_2 C$. Each step only halves the error.

Example 0.43 (The error squares at each step). For the iteration $x^{(t+1)} = \frac{1}{2}(x^{(t)} + 4/x^{(t)})$ of Example 0.39,

$$\begin{aligned}
x^{(t+1)} - 2 &= \frac{1}{2} \left(x^{(t)} + \frac{4}{x^{(t)}} \right) - 2 && \text{(the iteration)} \\
&= \frac{(x^{(t)})^2 + 4 - 4x^{(t)}}{2x^{(t)}} && \text{(common denominator)} \\
&= \frac{(x^{(t)} - 2)^2}{2x^{(t)}}. && \text{(the numerator is a perfect square)}
\end{aligned}$$

So near 2 the error is roughly squared and divided by 4 at each step: the errors are 0.05, then 0.00061, then 0.000000093, and $0.05^2/(2 \cdot 2.05) \approx 0.00061$. The iterates converge quadratically (Definition 0.30): $C = \frac{1}{2}$ works for every $x^{(t)} \geq 1$, since then $\frac{1}{2x^{(t)}} \leq \frac{1}{2}$, and every iterate from 2.5 is at least 2, since the identity makes $x^{(t+1)} - 2 \geq 0$ whenever $x^{(t)} > 0$. Zhou [3] states such a bound for Newton's method in general; it holds from starting points close enough to a minimizer where the Hessian is positive definite, when near that minimizer the Hessian changes by at most a constant times the distance moved, which these notes do not prove.

Example 0.44 (Newton's method can find a maximum). From $x^{(0)} = -2.5$ the same iteration gives -2.05 , -2.000610 , -2.0000001 , approaching -2 . But -2 is not a minimizer of f : it is a stationary point of $-f$, whose second derivative there is $-f''(-2) = 4 > 0$, so it is a strict local minimizer of $-f$ (Theorem 0.6), that is, f is strictly larger at -2 than at all nearby points. Newton's method looks only for a stationary point, and here the Hessian $f''(x) = 2x$ is negative along the way, so the quadratic model has no minimizer (with $H < 0$, $q(\delta) = f + g\delta + \frac{1}{2}H\delta^2$ decreases without bound) and Theorem 0.13 does not apply: the positive definite hypothesis there is needed.

Theorem 0.14 (Newton's method minimizes a quadratic in one step). *Let $\mathbf{S}$ be a $p \times p$ positive definite matrix, $\underline{c} \in \mathbb{R}^p$, and $f(\underline{x}) = \frac{1}{2} \underline{x}^\top \mathbf{S} \underline{x} - \underline{c} \cdot \underline{x}$. From any $\underline{x}^{(0)}$, Newton's method (Definition 0.27) gives $\underline{x}^{(1)} = \mathbf{S}^{-1} \underline{c}$, the unique minimizer of f .*

i Proof

Proof. Partial derivatives of a sum or of a constant multiple are the sum or multiple of the partial derivatives, so by [derivative of a quadratic form](#) for $\frac{1}{2} \underline{x}^\top \mathbf{S} \underline{x}$ and [derivative of a linear combination](#) for $\underline{c} \cdot \underline{x}$,

$$\begin{aligned}
\frac{\partial}{\partial \underline{x}} f(\underline{x}) &= \frac{1}{2} (2\mathbf{S}\underline{x}) - \underline{c} \\
&= \mathbf{S}\underline{x} - \underline{c};
\end{aligned}$$

by [Hessian of a quadratic form](#) with $\frac{1}{2}\mathbf{S}$ in place of $\mathbf{S}$, and because the gradient $-\underline{c}$ of the linear term is constant ([constant with respect to a vector](#)), $\mathbf{H}_f(\underline{x}) = \mathbf{S}$, which is invertible

(positive definite inverse). So

$$\begin{aligned}
\tilde{x}^{(1)} &= \tilde{x}^{(0)} - \mathbf{S}^{-1}(\mathbf{S}\tilde{x}^{(0)} - \tilde{c}) && \text{(definition of Newton's method)} \\
&= \tilde{x}^{(0)} - \mathbf{S}^{-1}\mathbf{S}\tilde{x}^{(0)} + \mathbf{S}^{-1}\tilde{c} && \text{(distribute)} \\
&= \tilde{x}^{(0)} - \tilde{x}^{(0)} + \mathbf{S}^{-1}\tilde{c} && (\mathbf{S}^{-1}\mathbf{S} = \mathbf{I}_p) \\
&= \mathbf{S}^{-1}\tilde{c}. && \text{(cancel)}
\end{aligned}$$

The gradient at $\mathbf{S}^{-1}\tilde{c}$ is $\mathbf{S}\mathbf{S}^{-1}\tilde{c} - \tilde{c} = \mathbf{0}_{p \times 1}$, and f is strictly convex (Theorem 0.11, part 2), so $\mathbf{S}^{-1}\tilde{c}$ is a global minimizer (Corollary 0.1). It is the only one: a strictly convex function has at most one global minimizer, since the midpoint of two distinct global minimizers, which have equal values, would have a strictly smaller value (Definition 0.9 with $t = \frac{1}{2}$). $\square$

Example 0.45 (One step to the minimizer). Let $f(\tilde{x}) = \frac{1}{2}(x_1^2 + 3x_2^2)$, so $\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$ and $\tilde{c} = \mathbf{0}_{2 \times 1}$. From $\tilde{x}^{(0)} = (3, 1)^\top$ the gradient is $\mathbf{S}\tilde{x}^{(0)} = (3, 3)^\top$, and $\mathbf{S}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}$, so the Newton step is $-\mathbf{S}^{-1}(3, 3)^\top = -(3, 1)^\top$ and

$$\begin{aligned}
\tilde{x}^{(1)} &= (3, 1)^\top - (3, 1)^\top \\
&= \mathbf{0}_{2 \times 1},
\end{aligned}$$

the minimizer.

Definition 0.31 (Gradient descent). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first partial derivatives, let $\tilde{x}^{(0)}$ be a starting point, and let $s^{(0)}, s^{(1)}, \dots$ be positive **step lengths**. **Gradient descent** [17] computes, for $t = 0, 1, 2, \dots$,

$$\tilde{x}^{(t+1)} \stackrel{\text{def}}{=} \tilde{x}^{(t)} - s^{(t)} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}).$$

Example 0.46 (Fixed steps on a parabola). For $f(x) = x^2$, $f'(x) = 2x$, so gradient descent with fixed step length s is

$$\begin{aligned}
x^{(t+1)} &= x^{(t)} - 2s x^{(t)} \\
&= (1 - 2s) x^{(t)},
\end{aligned}$$

and $x^{(t)} = (1 - 2s)^t x^{(0)}$.

- With $s = 0.25$, $x^{(t)} = 0.5^t x^{(0)}$, which tends to the minimizer 0.
- With $s = 0.5$, the first step lands on 0, since $1 - 2s = 0$.

- With $s = 1.5$, $x^{(t)} = (-2)^t x^{(0)}$, which moves away from 0 for any $x^{(0)} \neq 0$: too long a step can make gradient descent fail.

i Remark (Gradient descent as Newton's method with a simpler Hessian)

Remark 0.4 (Gradient descent as Newton's method with a simpler Hessian). Gradient descent (Definition 0.31) is Newton's method (Definition 0.27) with the Hessian replaced by $\frac{1}{s^{(t)}} \mathbf{I}_p$, so it needs no second derivatives.

Definition 0.32 (Gradient ascent). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first [partial derivatives](#), let $x^{(0)}$ be a starting point, and let $s^{(0)}, s^{(1)}, \dots$ be positive [step lengths](#). **Gradient ascent** [17] repeats, for $t = 0, 1, 2, \dots$, the update

$$\tilde{x}^{(t+1)} \stackrel{\text{def}}{=} \tilde{x}^{(t)} + s^{(t)} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}).$$

Example 0.47 (Gradient ascent is gradient descent on the negative function). Gradient ascent on f takes the same steps as [gradient descent](#) on $-f$, since $\frac{\partial}{\partial x}(-f) = -\frac{\partial}{\partial x} f$. In one variable, for $f(x) = -x^2$ the [derivative](#) is $f'(x) = -2x$, so gradient ascent with fixed [step length](#) s is $x^{(t+1)} = (1 - 2s)x^{(t)}$, the iteration of Example 0.46, which tends to the maximizer 0 when $0 < s < 1$.

Definition 0.33 (Line search). In gradient descent (Definition 0.31), a **line search** [18] is a rule that chooses the step length $s^{(t)}$ by examining the values of f along the line $\{\tilde{x}^{(t)} - s \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}) : s > 0\}$.

Definition 0.34 (Exact line search). **Exact line search** is the line search (Definition 0.33) that chooses $s^{(t)}$ to minimize $f(\tilde{x}^{(t)} - s \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}))$ over $s > 0$.

Example 0.48 (Exact line search on a parabola). For $f(x) = x^2$ and $x^{(t)} = 2$, the derivative is $f'(2) = 4$, so the line is $\{2 - 4s : s > 0\}$ and the values along it are

$$\begin{aligned} \phi(s) &= f(2 - 4s) \quad (\text{the value at step length } s) \\ &= (2 - 4s)^2 \quad (f(x) = x^2) \end{aligned}$$

$\phi(s) \geq 0$ for every s , and

$$\begin{aligned}\phi(0.5) &= (2 - 2)^2 \\ &= 0,\end{aligned}$$

so exact line search chooses $s^{(t)} = 0.5$, which lands on the minimizer $x^{(t+1)} = 0$ in one step.

Definition 0.35 (Descent direction). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have a gradient $\tilde{g} \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{x})$ at $\tilde{x}$. A direction $\tilde{d}$ ([direction](#)) is a **descent direction** for f at $\tilde{x}$ if $\tilde{g} \cdot \tilde{d} < 0$. When $\tilde{g} \neq \mathbf{0}_{p \times 1}$, $-\tilde{g}$ is a descent direction, since $\tilde{g} \cdot (-\tilde{g}) = -\|\tilde{g}\|^2 < 0$.

Example 0.49 (Descent directions for $x_1^2 + x_2^2$ at $(1, 0)^\top$). Let $f(\tilde{x}) = x_1^2 + x_2^2$ and $\tilde{x} = (1, 0)^\top$, so

$$\begin{aligned}\tilde{g} &= (2x_1, 2x_2)^\top \\ &= (2, 0)^\top.\end{aligned}$$

- $\tilde{d} = (-1, 1)^\top$ is a descent direction:

$$\begin{aligned}\tilde{g} \cdot \tilde{d} &= -2 + 0 \\ &= -2 \\ &< 0.\end{aligned}$$

- $\tilde{d} = (0, 1)^\top$ is not: $\tilde{g} \cdot \tilde{d} = 0$.
- $-\tilde{g} = (-2, 0)^\top$ is: $\tilde{g} \cdot (-\tilde{g}) = -4 < 0$.

Theorem 0.15 (A short enough step downhill decreases the function). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first partial derivatives that are continuous on $\mathbb{R}^p$ ([continuity](#)), and let $\tilde{x} \in \mathbb{R}^p$ be a point with $\tilde{g} \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}) \neq \mathbf{0}_{p \times 1}$. Then there is an $\bar{s} > 0$ with $f(\tilde{x} - s\tilde{g}) < f(\tilde{x})$ for every $s \in (0, \bar{s})$: a short enough step in the descent direction $-\tilde{g}$ ([Definition 0.35](#)) decreases f .

Proof

Proof. Let $h(s) \stackrel{\text{def}}{=} f(\tilde{x} - sg)$. The inner function $s \mapsto \tilde{x} - sg$ has derivative $-\tilde{g}^\top$ ([vector-valued derivative](#)), so by the vector chain rule ([vector chain rule](#))

$$\begin{aligned}
h'(s) &= -\tilde{g}^\top \frac{\partial}{\partial x} f(x - sg), \quad (\text{chain rule}) \\
h'(0) &= -\tilde{g}^\top \frac{\partial}{\partial x} f(x) \quad (\text{set } s = 0) \\
&= -\tilde{g} \cdot \tilde{g} \quad (\text{definition of } \tilde{g}) \\
&= -\|\tilde{g}\|^2 \quad (\text{squared length}) \\
&< 0. \quad (\tilde{g} \neq 0_{p \times 1})
\end{aligned}$$

Since $h'(0) = \lim_{s \rightarrow 0} (h(s) - h(0))/s$ is negative, taking $\varepsilon = |h'(0)|$ in the definition of the limit gives an $\bar{s} > 0$ such that $(h(s) - h(0))/s < h'(0) + |h'(0)| = 0$ for $0 < s < \bar{s}$. Multiplying by $s > 0$, $h(s) < h(0)$, that is, $f(x - sg) < f(x)$. $\square$

Example 0.50 (How short is short enough). In Example 0.46, $f(x) = x^2$ at $x \neq 0$ has $g = 2x$, and $f(x - 2sx) = (1 - 2s)^2 x^2 < x^2$ exactly when $|1 - 2s| < 1$, that is, when $0 < s < 1$. So Theorem 0.15 holds with $\bar{s} = 1$, and the step $s = 1.5$ of that example was too long.

Example 0.51 (Gradient descent zigzags). Let $f(x) = \frac{1}{2}(x_1^2 + b x_2^2)$ with $b > 0$, so $\frac{\partial}{\partial x} f(x) = (x_1, b x_2)^\top$ and one gradient step of length s from $\tilde{x}$ gives $((1 - s)x_1, (1 - bs)x_2)^\top$. For exact line search (Definition 0.34), let $\phi(s) \stackrel{\text{def}}{=} f$ at that point:

$$\begin{aligned}
\phi(s) &= \frac{1}{2}((1 - s)^2 x_1^2 + b(1 - bs)^2 x_2^2), \quad (\text{definition of } f) \\
\phi'(s) &= -(1 - s)x_1^2 - b^2(1 - bs)x_2^2 \quad (\text{chain rule, term by term}) \\
&= -x_1^2 + s x_1^2 - b^2 x_2^2 + b^3 s x_2^2 \quad (\text{expand the products}) \\
&= s(x_1^2 + b^3 x_2^2) - (x_1^2 + b^2 x_2^2). \quad (\text{collect the terms in } s)
\end{aligned}$$

For $\tilde{x} \neq 0_{2 \times 1}$, $x_1^2 + b^3 x_2^2 > 0$, so ϕ' is increasing in s with a single zero, and ϕ is minimized at $s^* = \frac{x_1^2 + b^2 x_2^2}{x_1^2 + b^3 x_2^2}$. From $\tilde{x}^{(0)} = (b, 1)^\top$:

$$\begin{aligned}
s^{(0)} &= \frac{b^2 + b^2}{b^2 + b^3} \quad (\text{substitute } x_1 = b, x_2 = 1) \\
&= \frac{2}{1 + b}, \quad (\text{cancel } b^2)
\end{aligned}$$

and, with $r \stackrel{\text{def}}{=} \frac{b-1}{b+1}$,

$$\begin{aligned}
1 - s^{(0)} &= \frac{1 + b - 2}{1 + b} \\
&= r, & (\text{common denominator}) \\
1 - b s^{(0)} &= \frac{1 + b - 2b}{1 + b} \\
&= -r, & (\text{common denominator}) \\
\tilde{x}^{(1)} &= (r b, -r \cdot 1)^\top \\
&= r (b, -1)^\top. & (\text{the gradient step from } (b, 1)^\top)
\end{aligned}$$

For $b \neq 1$ the same holds at every step, by [induction](#) (when $b = 1$, $r = 0$ and $\tilde{x}^{(1)}$ is already the minimizer): if $\tilde{x}^{(t)} = (c b, \pm c)^\top$ for a number $c \neq 0$, then s^* , which depends only on $x_1^2 = c^2 b^2$ and $x_2^2 = c^2$, is again $\frac{2}{1+b}$ (the factor c^2 cancels), and the step multiplies the first entry by r and the second by $-r$. So

$$\begin{aligned}
\tilde{x}^{(t)} &= (b r^t, (-r)^t)^\top, \\
f(\tilde{x}^{(t)}) &= \frac{1}{2}(b^2 + b) r^{2t} \\
&= r^{2t} f(\tilde{x}^{(0)}).
\end{aligned}$$

For $b \neq 1$, exactly one entry changes sign at every step: the second when $b > 1$ (then $r > 0$), and the first when $b < 1$ (then $r < 0$). So the iterates zigzag toward the minimizer $0_{2 \times 1}$.

- With $b = 3$, $r = \frac{1}{2}$: $\tilde{x}^{(1)} = (1.5, -0.5)^\top$ and $\tilde{x}^{(2)} = (0.75, 0.25)^\top$, and f falls from 6 to 1.5 to 0.375, a quarter at each step.
- With $b = 1$, $r = 0$ and one step reaches the minimizer.
- With $b = 0.01$, $r^2 = (0.99/1.01)^2 \approx 0.96$, so f shrinks by only about 4% per step. Newton's method reaches the minimizer of any such f in one step (Theorem [0.14](#)).

Definition 0.36 (Armijo condition). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first partial derivatives, let $\tilde{x} \in \mathbb{R}^p$, and let $g \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{x})$. A step length $s > 0$ satisfies the **Armijo condition** [19] with constant $c \in (0, 1)$ at $\tilde{x}$ if

$$f(\tilde{x} - s g) \leq f(\tilde{x}) - c s \|g\|^2.$$

Example 0.52 (The Armijo condition on a parabola). Let $f(x) = x^2$ and $x = 2$, so $g = f'(2) = 4$ and $\|g\|^2 = 16$. With $c = 1/4$, the condition on a step length $s > 0$ is

$$\begin{aligned}
(2 - 4s)^2 &\leq 4 - 4s & (f(2 - 4s) \leq f(2) - c s \|g\|^2) \\
4 - 16s + 16s^2 &\leq 4 - 4s & (\text{expand the square}) \\
16s^2 &\leq 12s & (\text{add } 16s - 4 \text{ to both sides}) \\
s &\leq 3/4. & (\text{divide by } 16s > 0)
\end{aligned}$$

So $s = 1/2$ satisfies the condition: $f(0) = 0 \leq 4 - 2 = 2$. And $s = 1$ does not: $f(-2) = 4 > 4 - 4 = 0$.

Definition 0.37 (Backtracking line search). Let $s_0 > 0$ be an initial step length, $\beta \in (0, 1)$ a shrink factor, and $c \in (0, 1)$. **Backtracking line search** [19] is the [line search](#) that chooses $s^{(t)}$ to be the largest of $s_0, \beta s_0, \beta^2 s_0, \dots$ that satisfies the [Armijo condition](#) with constant c at $\tilde{x}^{(t)}$.

Example 0.53 (Backtracking on a parabola). For $f(x) = x^2$ at $x = 2$ with $c = 1/4$, the Armijo condition holds exactly when $s \leq 3/4$ (Example 0.52). With $s_0 = 1$ and $\beta = 1/2$, the trial step lengths are $1, 1/2, 1/4, \dots$. The first, $s = 1$, exceeds $3/4$ and is rejected. The second, $s = 1/2$, is at most $3/4$ and is accepted. So backtracking chooses $s^{(0)} = 1/2$, after two evaluations of f .

Quasi-Newton methods

Source and license

This section is adapted from the quasi-Newton lecture of Zhou [20], used under the MIT License. The license text is:

Copyright (c) 2025 ucla-biostat-257

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A

PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

From the source, this section takes the order of ideas: the cost of Newton's method, the secant condition, and the BFGS and limited-memory BFGS updates. The source states the updates as solutions of nearest-matrix problems without proof; here the BFGS update is written out. The Davidon-Fletcher-Powell and rank-1 updates, the nearest-matrix derivation, the history, and the Julia code are not part of this section.

Definition 0.38 (Secant condition). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first [partial derivatives](#), and let $\tilde{x}, \tilde{x}^+ \in \mathbb{R}^p$. Write $\tilde{s} \stackrel{\text{def}}{=} \tilde{x}^+ - \tilde{x}$ for the step and $\tilde{y} \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^+) - \frac{\partial}{\partial \tilde{x}} f(\tilde{x})$ for the change in the [gradient](#). A [symmetric matrix](#) $\mathbf{B}$ satisfies the **secant condition** [21, Ch. 6] for $(\tilde{s}, \tilde{y})$ when

$$\mathbf{B}_{\tilde{s}} = \tilde{y}.$$

Example 0.54 (The Hessian of a quadratic function satisfies the secant condition). Let $f(\tilde{x}) = \frac{1}{2} \tilde{x}^\top \mathbf{Q} \tilde{x} - \tilde{b} \cdot \tilde{x}$ with $\mathbf{Q}$ [symmetric](#). Then $\frac{\partial}{\partial \tilde{x}} f(\tilde{x}) = \mathbf{Q} \tilde{x} - \tilde{b}$, so for any $\tilde{x}$ and $\tilde{x}^+$

$$\begin{aligned} \tilde{y} &= (\mathbf{Q} \tilde{x}^+ - \tilde{b}) - (\mathbf{Q} \tilde{x} - \tilde{b}) && \text{(gradient at each point)} \\ &= \mathbf{Q}(\tilde{x}^+ - \tilde{x}) && \text{(collect terms)} \\ &= \mathbf{Q} \tilde{s}. && \text{(definition of } \tilde{s}) \end{aligned}$$

So the Hessian $\mathbf{Q}$ of the [quadratic function](#) satisfies the secant condition for every step. For a general f , $\tilde{y} \approx \mathbf{H}_f \tilde{s}$ holds for a short step, with $\mathbf{H}_f$ the Hessian of f near the step, which is why a [matrix](#) that satisfies the secant condition is a reasonable stand-in for the Hessian.

Definition 0.39 (Quasi-Newton method). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ have first [partial derivatives](#), let $\tilde{x}^{(0)}$ be a starting point, and let $\mathbf{H}^{(0)}$ be a [positive definite matrix](#). A **quasi-Newton method** [21, Ch. 6] produces, for $t = 0, 1, 2, \dots$,

$$\tilde{x}^{(t+1)} \stackrel{\text{def}}{=} \tilde{x}^{(t)} - \alpha^{(t)} \mathbf{H}^{(t)} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}),$$

where $\alpha^{(t)} > 0$ is a [step length](#) chosen by a [line search](#), and $\mathbf{H}^{(t+1)}$ is a [symmetric](#) positive definite matrix whose [inverse](#) satisfies the [secant condition](#), that is, $\mathbf{H}^{(t+1)} \tilde{y}^{(t)} = \tilde{s}^{(t)}$, with

$$\tilde{s}^{(t)} = \tilde{x}^{(t+1)} - \tilde{x}^{(t)} \text{ and } \tilde{y}^{(t)} = \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t+1)}) - \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)}).$$

Example 0.55 (One quasi-Newton step on a quadratic function). Take $f(\tilde{x}) = \frac{1}{2} \tilde{x}^\top \mathbf{Q} \tilde{x} - \tilde{b} \cdot \tilde{x}$ with $\mathbf{Q} = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}$ and $\tilde{b} = (1, 2)^\top$ (Example 0.54), $\tilde{x}^{(0)} = \mathbf{0}$ and $\mathbf{H}^{(0)} = \mathbf{I}_2$. For a quadratic function, the step length that minimizes f along a direction $\tilde{d}$ is $\alpha = -\frac{\partial}{\partial \tilde{x}} f(\tilde{x}) \cdot \tilde{d} / \tilde{d}^\top \mathbf{Q} \tilde{d}$.

```
Q <- matrix(c(4, 1, 1, 3), nrow = 2)
b <- c(1, 2)
f <- function(x) 0.5 * sum(x * (Q %*% x)) - sum(b * x)
grad <- function(x) drop(Q %*% x - b)
x0 <- c(0, 0)
H0 <- diag(2)
d <- -drop(H0 %*% grad(x0))
step <- -sum(grad(x0) * d) / drop(t(d) %*% Q %*% d)
x1 <- x0 + step * d
s <- x1 - x0
y <- grad(x1) - grad(x0)
```

The step length is 0.25, which moves the iterate to $x^{(1)} = (0.25, 0.5)$ and lowers f from 0 to -0.625. The difference of the gradients, $\tilde{y}^{(0)} = (1.5, 1.75)$, is $\mathbf{Q} \tilde{s}^{(0)}$ (the secant condition for a quadratic function).

i Remark (Why a quasi-Newton method is used)

Remark 0.5 (Why a quasi-Newton method is used). Because $\mathbf{H}^{(t)}$ is positive definite, $-\mathbf{H}^{(t)} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)})$ is a descent direction, so, when the gradient is not 0, a line search finds a step that decreases f . Newton's method needs the Hessian and the solution of a system of equations at every iteration, which costs about p^3 operations for a dense $p \times p$ Hessian, and requires the formulas for its entries. A quasi-Newton method needs only the gradient and an update of $\mathbf{H}^{(t)}$ costing about p^2 operations. Near a minimizer the method speeds up superlinearly: faster than gradient descent, slower than Newton's method [21, Ch. 6]. Replacing $\mathbf{H}^{(t)}$ by $\mathbf{I}_p$ in the step formula gives gradient descent.

Definition 0.40 (BFGS update (Broyden-Fletcher-Goldfarb-Shanno)). Let $\mathbf{H}$, an approximation of the inverse Hessian, be a $p \times p$ positive definite matrix, and let $\tilde{s}, \tilde{y} \in \mathbb{R}^p$ satisfy $\tilde{y} \cdot \tilde{s} > 0$. With $\rho \stackrel{\text{def}}{=} 1/\tilde{y} \cdot \tilde{s}$, the BFGS update [21, Ch. 6] of $\mathbf{H}$ is

$$\mathbf{H}^+ \stackrel{\text{def}}{=} (\mathbf{I}_p - \rho \tilde{s} \tilde{y}^\top) \mathbf{H} (\mathbf{I}_p - \rho \tilde{y} \tilde{s}^\top) + \rho \tilde{s} \tilde{s}^\top.$$

Example 0.56 (A BFGS update from the quasi-Newton step). Continue Example 0.55 with $\mathbf{H} = \mathbf{I}_2$, $\tilde{s} = \tilde{s}^{(0)}$, and $\tilde{y} = \tilde{y}^{(0)}$.

```
rho <- 1 / sum(y * s)
left <- diag(2) - rho * s %*% t(y)
H1 <- left %*% H0 %*% t(left) + rho * s %*% t(s)
H1
#>           [,1]      [,2]
#> [1,]  0.6625 -0.425
#> [2,] -0.4250  0.650
drop(H1 %*% y)
#> [1] 0.25 0.50
```

Here $\tilde{y} \cdot \tilde{s} = 1.25 > 0$, so $\rho = 0.8$. The updated **matrix** satisfies $\mathbf{H}^+ \tilde{y} = (0.25, 0.5)$, which is $\tilde{s} = (0.25, 0.5)$, and its smallest **eigenvalue** is 0.231204, so $\mathbf{H}^+$ is **positive definite**.

i Remark (The curvature condition $\tilde{y} \cdot \tilde{s} > 0$)

Remark 0.6 (The curvature condition $\tilde{y} \cdot \tilde{s} > 0$). If the **strict inequality** of Theorem 0.10 holds for distinct points, then $\tilde{y} \cdot \tilde{s} > 0$ for $\tilde{s} = \tilde{x}^+ - \tilde{x} \neq 0$ and $\tilde{y} = \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^+) - \frac{\partial}{\partial \tilde{x}} f(\tilde{x})$. Add the versions at $(\tilde{x}^+, \tilde{x})$ and at $(\tilde{x}, \tilde{x}^+)$:

$$f(\tilde{x}^+) + f(\tilde{x}) > f(\tilde{x}) + f(\tilde{x}^+) + \frac{\partial}{\partial \tilde{x}} f(\tilde{x}) \cdot \tilde{s} - \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^+) \cdot \tilde{s},$$

so $0 > -\tilde{y} \cdot \tilde{s}$. For a general f , a **line search** that enforces a curvature condition on the **step length** (the Wolfe conditions [21, Ch. 3]) guarantees $\tilde{y} \cdot \tilde{s} > 0$. Such a step can be found along a **descent direction** when f cannot keep decreasing forever.

Definition 0.41 (Limited-memory BFGS (L-BFGS)). **Limited-memory BFGS (L-BFGS)** [21, Ch. 7] is the **quasi-Newton method** that never forms $\mathbf{H}^{(t)}$. It stores only the m most recent pairs $(\tilde{s}^{(j)}, \tilde{y}^{(j)})$, with m small, such as 5 or 20, and obtains $\mathbf{H}^{(t)} \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(t)})$ implicitly, through a few vector operations, as if the **BFGS update** for those m pairs were applied to a simple starting **matrix**, such as a multiple of $\mathbf{I}_p$.

Example 0.57 (L-BFGS applies the update without forming the matrix). Continue Example 0.56 with $m = 1$ stored pair $(\tilde{s}^{(0)}, \tilde{y}^{(0)})$ and starting **matrix** $\mathbf{I}_2$. Below, $\mathbf{H}^+ \frac{\partial}{\partial \tilde{x}} f(\tilde{x}^{(1)})$ is formed from vectors alone, starting from H0, then compared with the **matrix product**.

```

g1 <- grad(x1)
alpha <- rho * sum(s * g1)
q <- g1 - alpha * y
r <- drop(H0 %*% q)
beta <- rho * sum(y * r)
r <- r + s * (alpha - beta)
rbind(from_vectors = r, from_matrix = drop(H1 %*% g1))
#>           [,1]    [,2]
#> from_vectors 0.4375 -0.375
#> from_matrix  0.4375 -0.375

```

The two rows agree to within $5.6\text{e-}17$, although the first never forms the matrix. BFGS stores a dense $p \times p$ matrix and spends about p^2 operations per iteration; L-BFGS stores $2m$ vectors with p entries and spends about mp operations per iteration, so it remains usable when p is in the millions.

Further reading

- Boyd and Vandenberghe [22] is the standard reference on convex optimization, covering convex sets, convex functions, convex optimization problems, duality, unconstrained minimization, and interior-point methods.
- Nocedal and Wright [21] is a comprehensive reference on continuous numerical optimization, covering line search methods, trust-region methods, conjugate gradient methods, quasi-Newton methods, and nonlinear equations.
- Hua Zhou’s [course page for “UCLA Biostat 257 - Statistical Computing” \(2025 Spring\)](#) has lectures on optimization, from root finding to constrained problems. Its source is [on GitHub](#) under the MIT License.
- Hua Zhou’s [lecture notes for “UCLA Biostat 216 - Mathematical Methods for Biostatistics” \(2024 Fall\)](#)
- [Neural Networks](#) is a YouTube playlist by Grant Sanderson (3Blue1Brown); its chapters on gradient descent and backpropagation calculus visually illustrate how gradients of multivariate cost functions are computed and used for optimization.

References

- [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_\(stationary_points\)](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
- [8] Wikipedia contributors, “Convex set — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Convex_set
- [9] Wikipedia contributors, “Epigraph (mathematics) — Wikipedia, the free encyclopedia.” 2026. Available: [https://en.wikipedia.org/wiki/Epigraph_\(mathematics\)](https://en.wikipedia.org/wiki/Epigraph_(mathematics))
- [10] Wikipedia contributors, “Tangent space — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Tangent_space
- [11] Wikipedia contributors, “Feasible region — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Feasible_region
- [12] Wikipedia contributors, “Mathematical optimization — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Mathematical_optimization
- [13] Wikipedia contributors, “Lagrange multiplier — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Lagrange_multiplier
- [14] Wikipedia contributors, “Iterative method — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Iterative_method
- [15] Wikipedia contributors, “Newton’s method in optimization — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization
- [16] Wikipedia contributors, “Rate of convergence — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Rate_of_convergence
- [17] Wikipedia contributors, “Gradient descent — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Gradient_descent
- [18] Wikipedia contributors, “Line search — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Line_search

- [19] Wikipedia contributors, “Backtracking line search — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Backtracking_line_search
- [20] H. Zhou, “Quasi-Newton methods.” Lecture notes for Biostat 257, Statistical Computing, University of California, Los Angeles, 2025. Accessed: Oct. 26, 2026. [Online]. Available: <https://ucla-biostat-257.github.io/2025spring/slides/24-quasinewton/quasinewton.html>
- [21] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed. in Springer series in operations research and financial engineering. Springer, 2006. doi: [10.1007/978-0-387-40065-5](https://doi.org/10.1007/978-0-387-40065-5).
- [22] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004. doi: [10.1017/CBO9780511804441](https://doi.org/10.1017/CBO9780511804441).