

Multivariable Calculus

Aayush Bajaj

July 10, 2026

Contents

1	partial derivatives and the gradient	2
1.1	definitions	2
1.2	steepest ascent, with proof	2
2	jacobian, hessian, chain rule	3
2.1	the derivative is a matrix	3
2.2	second derivatives	3
2.3	taylor to second order	3
3	optimisation	3
3.1	unconstrained: critical points and the hessian test	3
3.2	constrained: lagrange multipliers	4
3.3	worked example	4
4	multiple integrals	5
4.1	fubini	5
4.2	change of variables	5
4.3	worked example: the gaussian via polar coordinates	5
5	vector calculus	6
5.1	div, curl, and the two kinds of derivative of a field	6
5.2	line and surface integrals	6
5.3	the three big theorems, and why they are one	6
6	numerics: verifying the lagrange example	7
7	see also	7

calculus in $\mathbb{R}^n$: functions of several variables, the surfaces they define, and the fields that flow over them. the programme is the same as **one variable** — linearise locally, integrate globally — but the derivative becomes a matrix, the chain rule becomes matrix multiplication, and the fundamental theorem splits into three named theorems (green, stokes, gauss) that are secretly one (Courant, Richard, 1996). this page is also the mathematical spine of machine learning: gradients, Hessians, Jacobians and constrained optima are chapter 5 of (Deisenroth, Marc Peter and Faisal, A. Aldo and Ong, Cheng Soon, 2020).

1 partial derivatives and the gradient

1.1 definitions

for $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the **partial derivative** freezes all coordinates but one:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h \mathbf{e}_i) - f(\mathbf{x})}{h}, \quad (1)$$

and the **gradient** collects them, $\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^\top$. but the honest definition of differentiability is the one-variable idea verbatim — existence of a best linear approximation:

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + o(\|\mathbf{h}\|). \quad (2)$$

partials existing does *not* imply differentiability (partials only probe axis directions); continuous partials do.

for differentiable f , the **directional derivative** along a unit vector $\mathbf{v}$ is

$$D_{\mathbf{v}}f(\mathbf{x}) = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \nabla f(\mathbf{x})^\top \mathbf{v}. \quad (3)$$

*the standard counterexample
 $f = xy/(x^2 + y^2)$ with $f(0) = 0$
 has both partials at the origin
 yet is not even continuous
 there.*

1.2 steepest ascent, with proof

claim. among unit vectors, $D_{\mathbf{v}}f$ is maximised by $\mathbf{v} = \nabla f / \|\nabla f\|$, with maximum value $\|\nabla f\|$; and ∇f is orthogonal to the level set through $\mathbf{x}$.

proof. cauchy–schwarz gives $D_{\mathbf{v}}f = \nabla f^\top \mathbf{v} \leq \|\nabla f\| \|\mathbf{v}\| = \|\nabla f\|$, with equality iff $\mathbf{v}$ is a positive multiple of ∇f — so the gradient direction wins, and the opposite direction is steepest descent. for orthogonality: let $\gamma(t)$ be any differentiable curve inside the level set $\{f = c\}$ with $\gamma(0) = \mathbf{x}$. then $f(\gamma(t)) \equiv c$, and the chain rule gives $\nabla f^\top \gamma'(0) = 0$ — the gradient kills every tangent direction. ■

this pair of facts is the whole geometry of gradient descent: move perpendicular to the contour lines, at a rate set by their crowding.

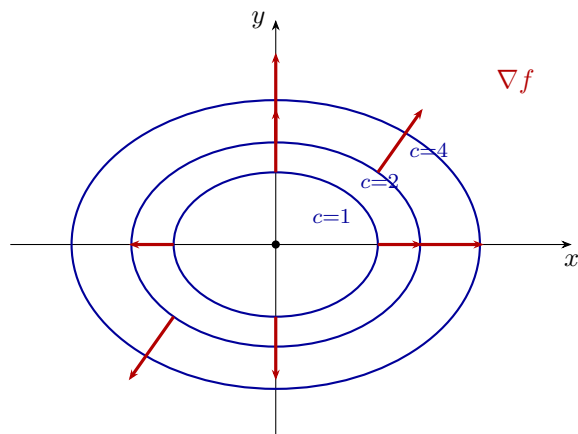


Figure 1: level curves of $f(x, y) = x^2 + 2y^2$ with gradient vectors: always orthogonal to the contours, always pointing uphill, longer where the contours crowd.

2 jacobian, hessian, chain rule

2.1 the derivative is a matrix

for $\mathbf{f}: \mathbb{R}^n \rightarrow \mathbb{R}^m$, the derivative at $\mathbf{x}$ is the $m \times n$ **jacobian**

$$J_{\mathbf{f}}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}, \quad \mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + J_{\mathbf{f}}(\mathbf{x}) \mathbf{h} + o(\|\mathbf{h}\|). \quad (4)$$

a scalar f has $J_f = \nabla f^\top$ (a row); a curve $\gamma: \mathbb{R} \rightarrow \mathbb{R}^m$ has $J_\gamma = \gamma'$ (a column). the **chain rule** is now literally matrix multiplication:

$$J_{\mathbf{g} \circ \mathbf{f}}(\mathbf{x}) = J_{\mathbf{g}}(\mathbf{f}(\mathbf{x})) J_{\mathbf{f}}(\mathbf{x}), \quad (5)$$

composition of linear approximations. every backpropagation pass in a neural network is this equation evaluated right-to-left.

reverse-mode autodiff is just a bracketing choice: multiply the jacobian chain starting from the scalar-loss end, so every intermediate product stays a row vector.

2.2 second derivatives

for scalar f , the **hessian** collects second partials, $H(\mathbf{x})_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. by **schwarz's theorem**, if the second partials are continuous then $\partial_i \partial_j f = \partial_j \partial_i f$ — the hessian is symmetric, which is what lets the spectral theorem act on it in the second-derivative test below.

2.3 taylor to second order

for f twice continuously differentiable,

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top H(\mathbf{x}) \mathbf{h} + o(\|\mathbf{h}\|^2). \quad (6)$$

the local model of any smooth function is a quadratic: a plane tilted by ∇f , curved by H . newton's method in optimisation is nothing but minimising this quadratic exactly at each step.

3 optimisation

3.1 unconstrained: critical points and the hessian test

at an interior extremum of differentiable f , $\nabla f = \mathbf{0}$ (fermat, coordinate by coordinate). classification at a critical point $\mathbf{x}^*$ reads off the taylor quadratic $\frac{1}{2} \mathbf{h}^\top H \mathbf{h}$:

- H **positive definite** (all eigenvalues > 0): strict local minimum.
- H **negative definite**: strict local maximum.
- H **indefinite** (eigenvalues of both signs): saddle point.
- H **singular**: the test is silent — higher-order terms decide (e.g. $f(x, y) = x^4 + y^4$ vs $x^4 - y^4$ at the origin).

in two variables the eigenvalue conditions compress to the familiar discriminant: with $D = f_{xx}f_{yy} - f_{xy}^2$, $D > 0$ with $f_{xx} > 0$ is a min, $D > 0$ with $f_{xx} < 0$ a max, $D < 0$ a saddle.

3.2 constrained: lagrange multipliers

problem. extremise $f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$, with $\nabla g \neq \mathbf{0}$ on the constraint set.

derivation. let $\mathbf{x}^*$ be a constrained extremum and $\gamma(t)$ any differentiable curve lying in the constraint surface with $\gamma(0) = \mathbf{x}^*$. two observations:

- $t \mapsto f(\gamma(t))$ has an extremum at 0, so $\nabla f(\mathbf{x}^*)^\top \gamma'(0) = 0$: ∇f is orthogonal to every tangent direction of the surface.
- the tangent space of $\{g = 0\}$ at $\mathbf{x}^*$ is exactly the orthogonal complement of $\nabla g(\mathbf{x}^*)$ (differentiate $g(\gamma(t)) \equiv 0$; the implicit function theorem guarantees enough curves to fill the whole complement).

so $\nabla f(\mathbf{x}^*)$ lies in the one-dimensional space spanned by $\nabla g(\mathbf{x}^*)$:

$$\boxed{\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*)} \quad (7)$$

for some multiplier $\lambda \in \mathbb{R}$. geometrically: at the optimum, the level curve of f is *tangent* to the constraint — if they crossed transversally, sliding along the constraint would still improve f .

with several constraints $g_1, \dots, g_k$ the same argument gives ∇f as a linear combination of the ∇g_i — one multiplier each.

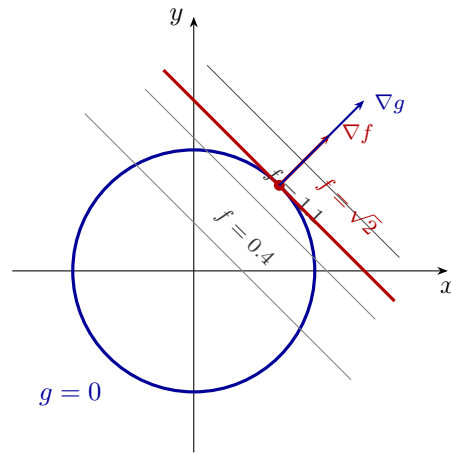


Figure 2: lagrange tangency: level lines of $f(x, y) = x + y$ sweep across the constraint circle $g = x^2 + y^2 - 1 = 0$; the optimum sits where a level line touches the circle and the two gradients align.

3.3 worked example

maximise $f(x, y) = xy$ on the circle $g(x, y) = x^2 + y^2 - 1 = 0$.

- **stationarity:** $\nabla f = \lambda \nabla g$ gives $y = 2\lambda x$ and $x = 2\lambda y$.
- **eliminate:** substituting, $y = 4\lambda^2 y$. if $y = 0$ then $x = 0$, off the constraint; so $\lambda = \pm \frac{1}{2}$ and $y = \pm x$.
- **constraint:** $2x^2 = 1$, so $x = \pm \frac{1}{\sqrt{2}}$.

- **classify:** $f = \frac{1}{2}$ at $\pm(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ (maxima, $\lambda = \frac{1}{2}$), $f = -\frac{1}{2}$ at $\pm(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ (minima). the constraint set is compact, so evt guarantees these are global.

the numeric check at the bottom of the page sweeps the circle and lands on the same point, value and multiplier.

4 multiple integrals

4.1 fubini

the double integral over a rectangle $R = [a, b] \times [c, d]$ is defined by riemann sums over grid cells, exactly as in one dimension. **fubini's theorem:** for f continuous on R ,

$$\iint_R f \, dA = \int_a^b \left(\int_c^d f(x, y) \, dy \right) dx = \int_c^d \left(\int_a^b f(x, y) \, dx \right) dy \quad (8)$$

— iterate in either order. over non-rectangular regions, describe the region by bounding curves and let the inner limits depend on the outer variable.

4.2 change of variables

if $T : U \rightarrow \mathbb{R}^n$ is injective and continuously differentiable with $\det J_T \neq 0$, then

$$\int_{T(U)} f(\mathbf{x}) \, d\mathbf{x} = \int_U f(T(\mathbf{u})) |\det J_T(\mathbf{u})| \, d\mathbf{u}. \quad (9)$$

the jacobian determinant is the local volume-scaling factor of the map — an infinitesimal $d\mathbf{u}$ -box lands as a parallelepiped of volume $|\det J_T| \, d\mathbf{u}$. in one dimension this is substitution; the absolute value appears because multiple integrals are unoriented.

4.3 worked example: the gaussian via polar coordinates

polar coordinates $T(r, \theta) = (r \cos \theta, r \sin \theta)$ have

$$J_T = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}, \quad \det J_T = r. \quad (10)$$

let $I = \int_{-\infty}^{\infty} e^{-x^2} dx$. square it and switch to polar:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r \, dr \, d\theta \\ &= 2\pi \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty} = 2\pi \cdot \frac{1}{2} = \pi, \end{aligned}$$

so $I = \sqrt{\pi}$. the factor r from the jacobian is precisely what makes the integrand elementary — the trick that has no one-variable counterpart, and the normalising constant of the entire theory of **probability**.

*triple integrals work identically:
spherical coordinates carry
jacobian $r^2 \sin \phi$.*

5 vector calculus

5.1 div, curl, and the two kinds of derivative of a field

a vector field $\mathbf{F} = (P, Q, R) : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ has two first-order derivatives that matter:

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}, \quad \operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F} = \begin{pmatrix} \frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z} \\ \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} \\ \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \end{pmatrix}. \quad (11)$$

divergence measures net outflow per unit volume (source density); curl measures circulation per unit area (local rotation). two identities anchor the algebra: $\operatorname{curl}(\nabla f) = \mathbf{0}$ and $\operatorname{div}(\operatorname{curl} \mathbf{F}) = 0$.

5.2 line and surface integrals

- **line integral of a field** along an oriented curve C parametrised by $\gamma(t)$, $t \in [a, b]$:

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\gamma(t)) \cdot \gamma'(t) dt \quad (12)$$

— work done by the field along the path. for gradient fields, $\int_C \nabla f \cdot d\mathbf{r} = f(\gamma(b)) - f(\gamma(a))$: path-independent, the ftc for curves.

- **surface integral (flux)** through an oriented surface S with unit normal $\mathbf{n}$:

$$\iint_S \mathbf{F} \cdot \mathbf{n} dS \quad (13)$$

— net flow of the field through the membrane.

5.3 the three big theorems, and why they are one

theorem	statement	dimension
green (plane)	$\oint_{\partial D} P dx + Q dy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$	curve = boundary of region
stokes (surface)	$\oint_{\partial S} \mathbf{F} \cdot d\mathbf{r} = \iint_S (\nabla \times \mathbf{F}) \cdot \mathbf{n} dS$	curve = boundary of surface
divergence (gauss)	$\iiint_V \mathbf{F} \cdot \mathbf{n} dS = \iiint_V (\nabla \cdot \mathbf{F}) dV$	surface = boundary of solid

(orientations: boundaries traversed so the region stays on the left; outward normals for gauss.) each says the same thing the one-variable ftc says: **accumulating a derivative over a region equals evaluating the original object on the boundary**. green is stokes for flat surfaces; the modern formulation absorbs all three (and the ftc itself) into the generalised stokes theorem

$$\int_{\partial\Omega} \omega = \int_{\Omega} d\omega \quad (14)$$

for differential forms ω — one theorem, every dimension (Courant, Richard, 1996). physics runs on the special cases: maxwell's equations pair gauss with stokes, and every conservation law is a divergence theorem in disguise.

6 numerics: verifying the lagrange example

sweep the constraint circle densely and compare against the analytic answer $x = y = \frac{1}{\sqrt{2}}$, $f = \frac{1}{2}$, $\lambda = \frac{1}{2}$:

```
1 import numpy as np
2
3 # maximise f(x,y) = xy on the unit circle x^2 + y^2 = 1.
4 # lagrange says: grad f = lambda grad g => (y, x) = lambda (2x, 2y)
5 # solving gives x = y = 1/sqrt(2), f = 1/2 (and the symmetric variants).
6
7 theta = np.linspace(0, 2 * np.pi, 2_000_001)    # dense sweep of the
8 ↪ constraint
9 f = np.cos(theta) * np.sin(theta)
10
11 i = f.argmax()
12 x, y = np.cos(theta[i]), np.sin(theta[i])
13 print(f"numeric argmax : x = {x:.6f}, y = {y:.6f}")
14 print(f"analytic      : x = y = 1/sqrt(2) = {1/np.sqrt(2):.6f}")
15 print(f"numeric max f  : {f[i]:.8f}    (analytic: 0.5)")
16
17 # check the gradients are parallel at the optimum: (y, x) vs (2x, 2y)
18 grad_f = np.array([y, x])
19 grad_g = np.array([2 * x, 2 * y])
20 lam = grad_f @ grad_g / (grad_g @ grad_g)
21 print(f"lambda        : {lam:.6f}    (analytic: 0.5)")
22 print(f"residual |grad f - lambda grad g| = {np.linalg.norm(grad_f - lam *
23 ↪ grad_g):.2e}")
```

```
numeric argmax : x = 0.707107, y = 0.707107
analytic      : x = y = 1/sqrt(2) = 0.707107
numeric max f  : 0.50000000    (analytic: 0.5)
lambda        : 0.500000    (analytic: 0.5)
residual |grad f - lambda grad g| = 1.57e-16
```

the argmax lands on the analytic point to six decimals, the least-squares multiplier equals the analytic λ , and the stationarity residual is at machine precision — the gradients really are parallel there.

7 see also

- [calculus](#) — the parent map
- [single-variable calculus](#) — where every theorem here comes from
- [linear algebra](#) — jacobians, hessians and definiteness live here
- [optimisation](#) — gradient descent, newton, and constraints at industrial scale
- [differential equations](#) — fields as dynamics

references

Courant, Richard (1996). *Differential and Integral Calculus*, Springer.

Deisenroth, Marc Peter and Faisal, A. Aldo and Ong, Cheng Soon (2020). *Mathematics for Machine Learning*, Cambridge University Press.