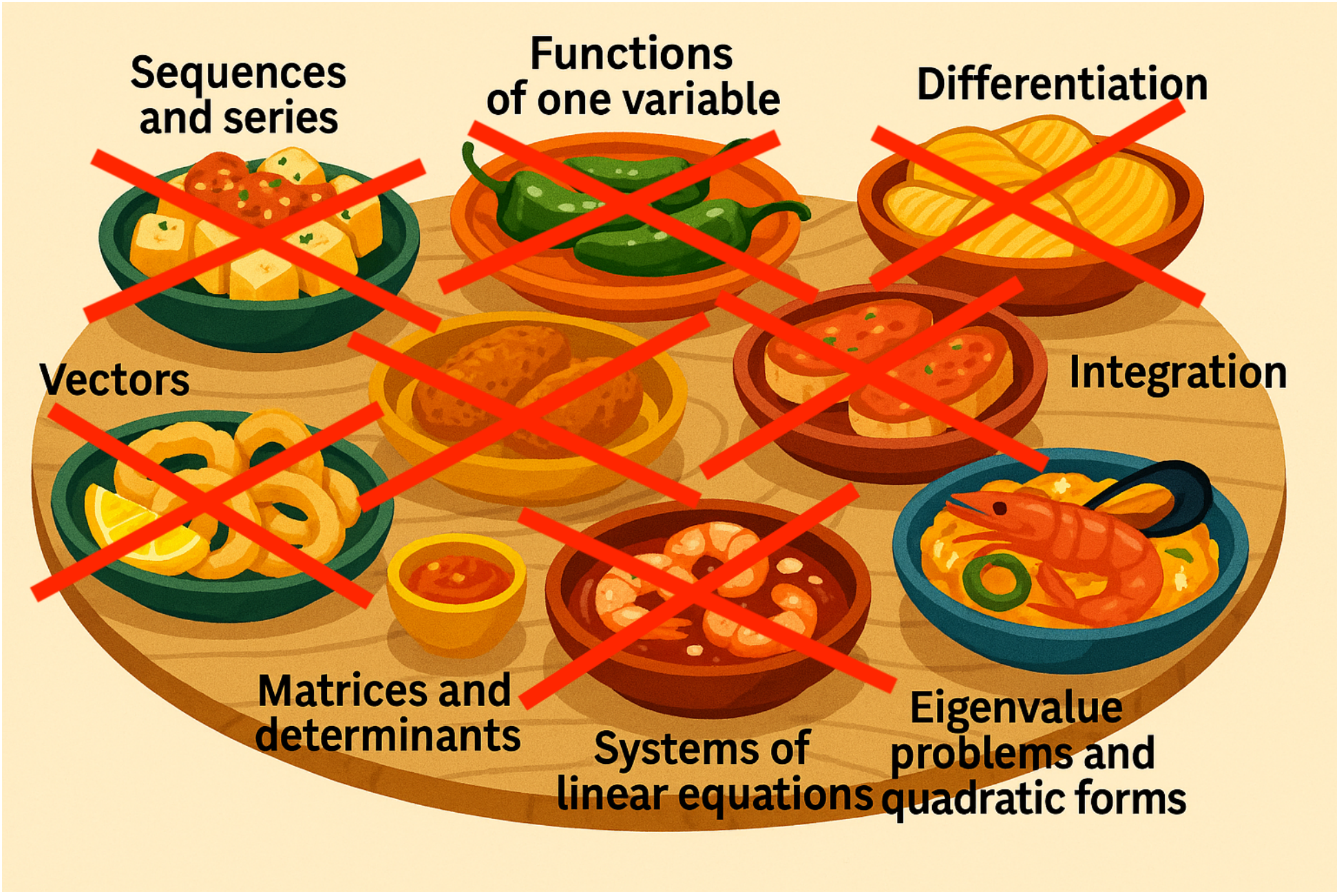
Lecture 5: Calculus and Linear Algebra

Chiara Amorino

Mathematics Brush-up

Course content (math tapas)



Spectral decomposition and applications

Spectral decomposition (symmetric case)

Theorem (Spectral Theorem)

If $A \in \mathbb{R}^{n \times n}$ is symmetric, then there exist an orthogonal $U \in \mathbb{R}^{n \times n}$ and a diagonal $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ such that

$$A = U \Lambda U^\top, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

The columns of U are orthonormal eigenvectors of A.

Application: Principal Component Analysis (PCA)

We observe n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

Step 1 (Centering): Arrange the data in a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.
Define the centering matrix $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and then the centered data is $H\boldsymbol{X}$.
(its rows are the centered data points $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$)

Step 2 (Sample covariance):

$$S = \tfrac{1}{n}(H\boldsymbol{X})^\top (H\boldsymbol{X}).$$

Step 3 (Spectral decomposition):

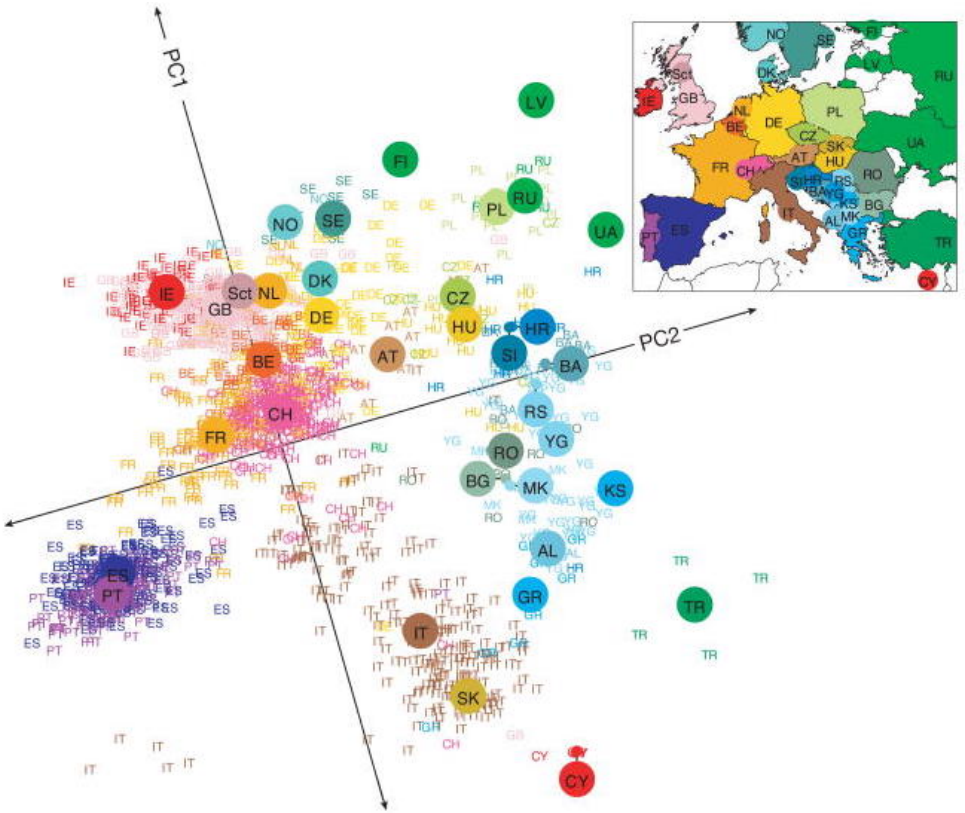
$$S = U\Lambda U^\top \quad \text{and define } \boldsymbol{Y} = H\boldsymbol{X}U \quad \text{(rotate the centered data)}.$$

The sample covariance of $\boldsymbol{Y}$ is $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} = U^\top S U = \Lambda$ (diagonal! and so $\boldsymbol{y}_i$ are uncorrelated).

The columns of U are principal directions and diagonal entries of Λ are the explained variances.

Sorting eigenvalues in decreasing order ranks components by variance captured.

PCA: why it matters



Dimensionality reduction: compress high-dimensional data into a few directions that preserve most variability.

Noise filtering: small eigenvalues $\rightarrow$ directions with little signal.

Visualization: projecting onto first 2 PCs often reveals hidden structure.

Nice example: genetic variation in Europeans — plotting individuals on the first two PCs of SNP data *recovers the geographic map of Europe*!

PCA & Unique football styles

"*Mille viae ducunt Barcinonem*" (A thousand roads lead to Barcelona)

"Searching for a Unique Style in Soccer" (Gyarmati, Kwak & Rodriguez, 2014), studies sequential pass patterns in a team's passing.

They defined "flow motifs" as statically significant pass subsequences (e.g., $A \to B \to C \to A$).

Applied motif analysis to compare styles across teams. The data are projected to the first two principal components.

Quadratic forms

Quadratic forms

A quadratic form is

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}, \qquad A \in \mathbb{R}^{n\times n}.$$

Why symmetric? If $a_{ij} \neq a_{ji}$, the asymmetric part cancels:

$$x^\top Ax = x^\top \tfrac{1}{2}(A + A^\top)x.$$

So every quadratic form has a unique symmetric representative.

Examples:

$$Q(x, y) = x^2 + 5xy + 4y^2$$

comes from $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ or equivalently $\begin{pmatrix} 1 & 2.5 \\ 2.5 & 4 \end{pmatrix}$.

Geometry: level sets $\{x : Q(x) = c\}$ are ellipses, hyperboloids, linear spaces.

Signs of quadratic forms

For symmetric A:

- Positive definite (PD): $x^\top Ax > 0$ for all $x \neq 0$.
- Negative definite (ND): $x^\top Ax < 0$ for all $x \neq 0$.
- Semidefinite: always ≥ 0 (PSD) or ≤ 0 (NSD).
- Indefinite: can take both signs.

Why it matters:

- In optimization, a PD Hessian $\Rightarrow$ strict local minimum.
- In economics, concave utility functions $\Rightarrow$ risk aversion.

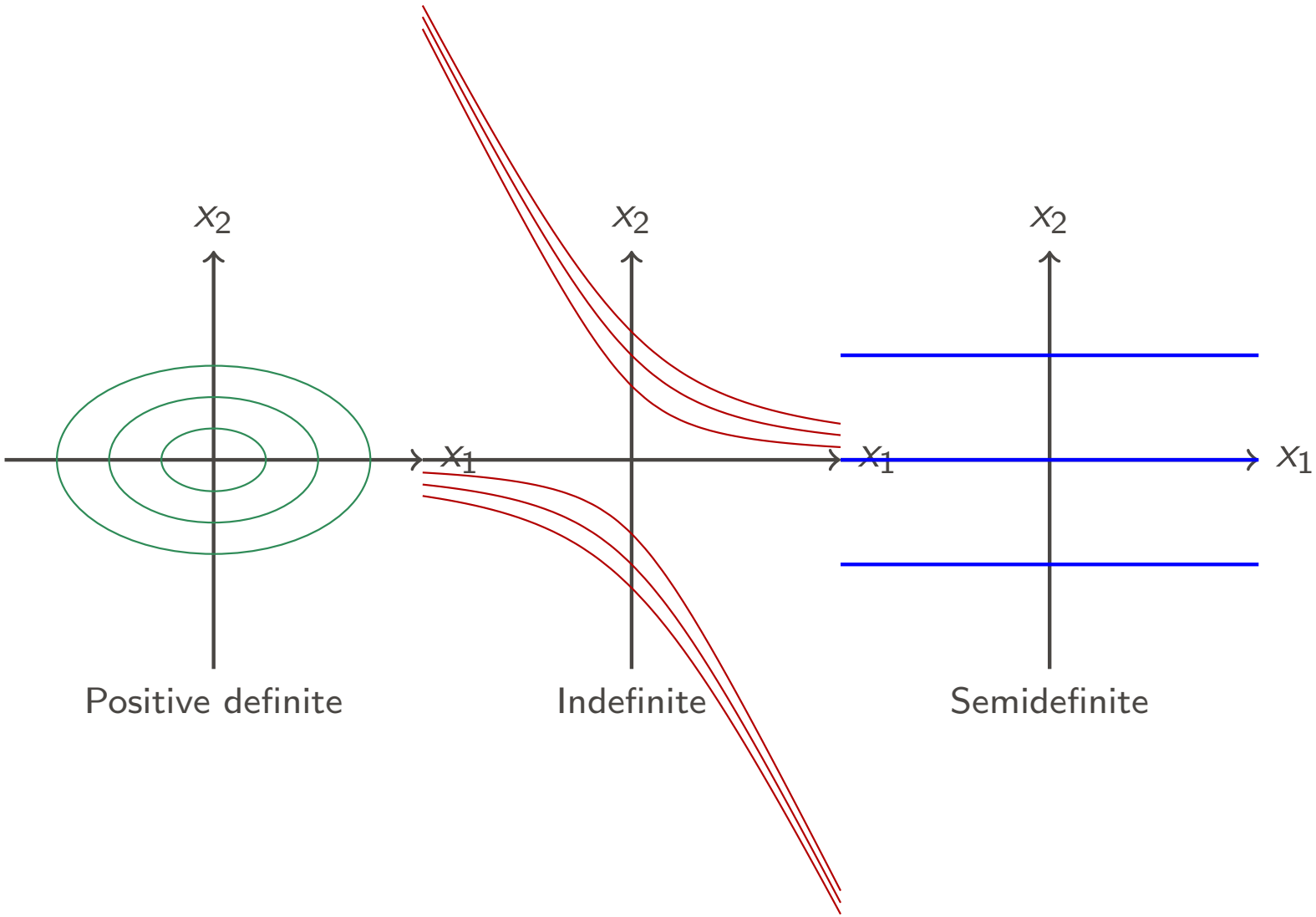
Example:

$$x^\top \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} x = (x-y)^2 \geq 0$$

so the matrix is PSD.

Geometry of quadratic forms

Level sets $\{x : Q(x) = c\}$ illustrate definiteness:



Ellipses $\rightarrow$ PD, Hyperbolas $\rightarrow$ Indefinite, Parallel lines $\rightarrow$ Semidefinite.

Eigenvalue and minor tests

Spectral test (symmetric A):

- PD $\iff$ all eigenvalues > 0; ND $\iff$ all < 0;
- PSD $\iff$ all ≥ 0; NSD $\iff$ all ≤ 0;
- Indefinite $\iff$ eigenvalues of different signs.

Sylvester's criterion (leading principal minors D_k):

- PD $\iff$ $D_k > 0$ for $k = 1, \ldots, n$;
- ND $\iff$ $(-1)^k D_k > 0$ for $k = 1, \ldots, n$.

Example: $A = \begin{pmatrix} -3 & 2 & 0 \\ 2 & -3 & 0 \\ 0 & 0 & -5 \end{pmatrix}$ has $D_1 = -3$, $D_2 = 5$, $D_3 = -25 \Rightarrow$ negative definite.

Examples of quadratic forms

1. $A = \begin{pmatrix} 1 & 4 & 6 \\ 4 & 2 & 1 \\ 6 & 1 & 6 \end{pmatrix} \Rightarrow$ indefinite (mixed signs of minors).
2. $A = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix} \Rightarrow$ positive definite.
3. $Q(x, y, z) = xy + yz \Rightarrow$ indefinite (matrix has both positive and negative eigenvalues).

Takeaway: sign of eigenvalues = curvature type.

Chapter 9: Functions of several variables

Many economic and data science models depend on several variables simultaneously.

Why this matters: In economics, technologies and preferences are multivariate (production, utility). In data science/ML, high-dimensional loss functions $L(\theta_1, \ldots, \theta_d)$ are optimized using gradients/Hessians.

Examples:

- **Economics:** Cobb–Douglas production $Y = K^{\alpha} L^{1-\alpha}$, utility $U(x, y)$.
- **Data science:** Loss functions $L(\theta)$ with $d \gg 1$ parameters.

Reading: Werner–Sotskov (Ch. 11); Simon–Blume (Chs. 14, 17).
Exercises: 11.11(a), 11.21, 11.22 (Werner–Sotskov).

Basic definitions and examples

What is a multivariable function?

A function of n variables is a map

$$f: \ D \subset \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} = (x_1, \ldots, x_n) \mapsto f(\mathbf{x}),$$

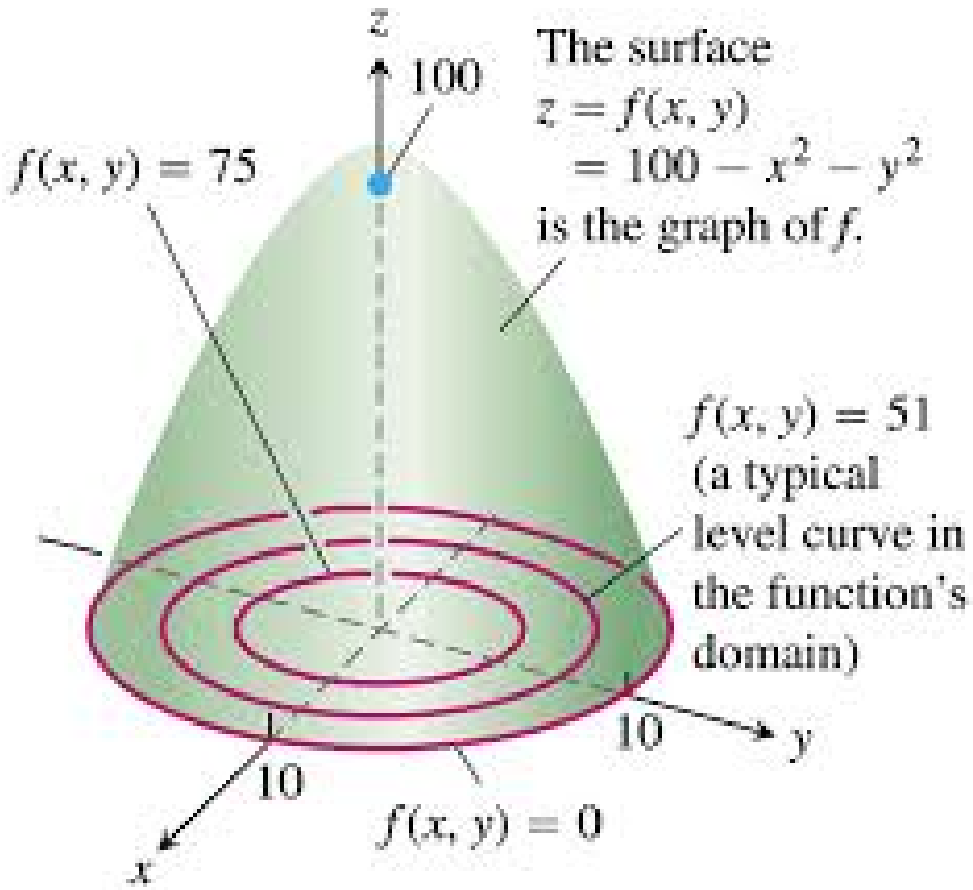
typically with D open. When discussing gradients/tangent planes we assume $f \in C^1$ on D.

- For $n = 2$: the graph $z = f(x, y)$ is a surface in $\mathbb{R}^3$.
- Level curves (contours): $\{(x, y) \in D : \ f(x, y) = c\}$.
- Sub/superlevel sets: $\{f \leq c\}, \{f \geq c\}$ — useful for (quasi-)convexity.
- For $n > 2$: use slices or projections to visualize.

Example: quadratic function

$$f(x, y) = 100 - x^2 - y^2$$

- Graph of f: paraboloid.
- Level curves: concentric circles.



Economic example: Cobb–Douglas

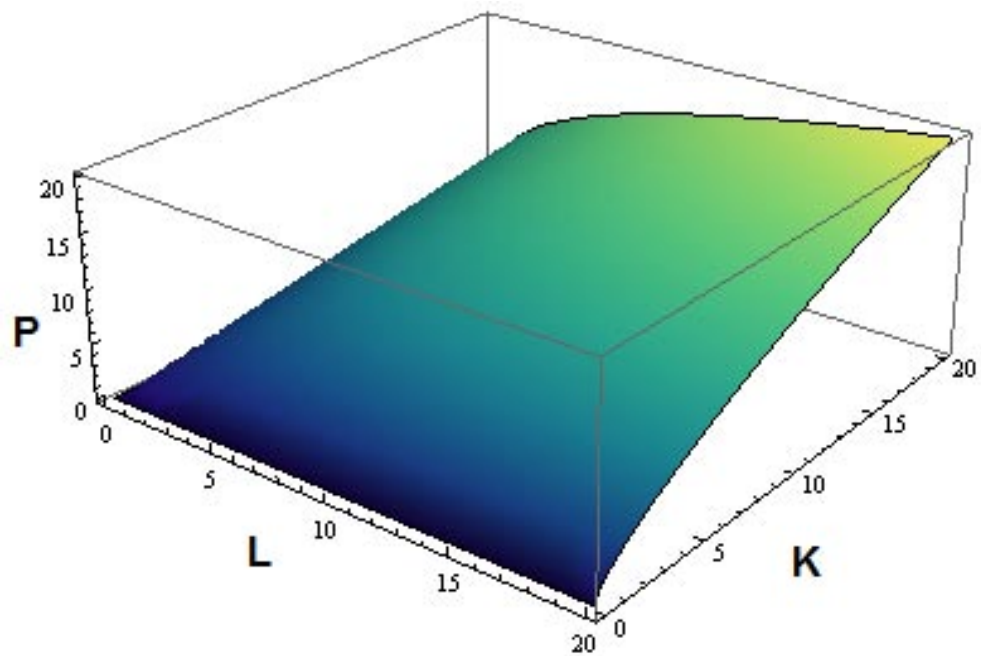
Cobb–Douglas production function

$$P(L, K) = b\, L^{\alpha} K^{\beta}.$$

P = output, L = labour, K = capital, b = total factor productivity, α, β = output elasticities.

Domain:

$$D = \{(L, K) \in \mathbb{R}^2 : \ L \geq 0, \ K \geq 0\}.$$



Multivariate Gaussian density

A random vector $\mathbf{X} \in \mathbb{R}^d$ is multivariate normal with mean $\mathbf{m}$ and $\Sigma \succ 0$ if

$$f(\mathbf{x}) = (2\pi)^{-d/2}(\det \Sigma)^{-1/2} \exp\Big(-\tfrac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m})\Big).$$

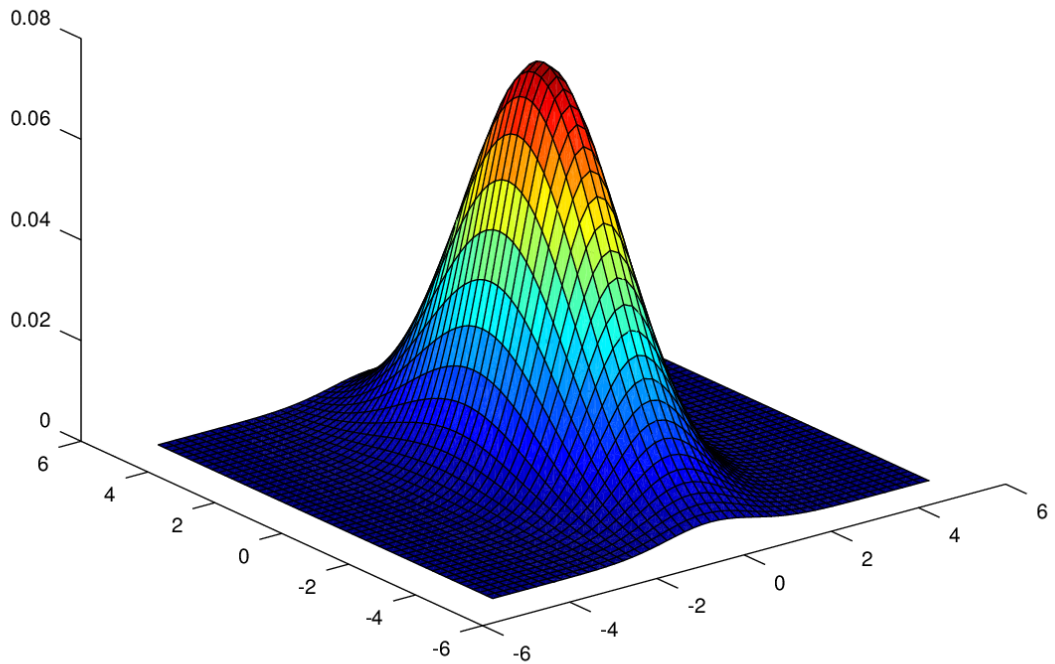
- Parameters: $\mathbf{m} = E(\mathbf{X})$, $\Sigma = \mathrm{Cov}(\mathbf{X}) \succ 0$.
- Depends on *Mahalanobis* distance

$$\|\mathbf{x}-\mathbf{m}\|_\Sigma = \sqrt{(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m})}.$$

- Level sets are ellipsoids $(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m}) = \text{const}$.

Multivariate Gaussian density

Example: $d = 2$, $\mathbf{m} = \mathbf{0}$, variances $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ (zero covariance).



Differentiation

Partial derivatives

Definition

For $f : D \subset \mathbb{R}^2 \to \mathbb{R}$, the partial derivatives at (x, y) are

$$f_x(x,y) = \lim_{h\to 0} \frac{f(x+h,y) - f(x,y)}{h}, \qquad f_y(x,y) = \lim_{h\to 0} \frac{f(x,y+h) - f(x,y)}{h}$$

when these limits exist. (We also use more standar notation $\frac{\partial f}{\partial x}(x,y)$, $\frac{\partial f}{\partial y}(x,y)$)

Equivalently, fix y and define $g(x) = f(x,y)$. Then $f_x(x,y) = g'(x)$.

Example (marginal costs): if

$$C(x,y) = 200 + 22x + 16y^{3/2},$$

then $C_x(x,y) = 22$ and $C_y(x,y) = 24\sqrt{y}$.

The gradient and linear approximation

For $\mathbf{x} \in \mathbb{R}^n$, the gradient is

$$\nabla f(\mathbf{x}) \;=\; \big(f_{x_1}(\mathbf{x}), \ldots, f_{x_n}(\mathbf{x})\big) \;\in\; \mathbb{R}^n.$$

Best linear approximation

If f is a C^1-function, then for $\boldsymbol{h} \in \mathbb{R}^n$,

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

So the gradient gives the best linear approximation of f near $\mathbf{x}$.

Geometry: ∇f is normal to level sets $f(\mathbf{x}) = c$

Why is the gradient the direction of steepest ascent?

Take $\boldsymbol{h} = t\boldsymbol{u}$ with $\|\boldsymbol{u}\| = 1$, $t > 0$ small. Then

$$f(\mathbf{x} + t\boldsymbol{u}) = f(\mathbf{x}) + t\,\langle \nabla f(\mathbf{x}), \boldsymbol{u} \rangle + o(t).$$

The instantaneous rate of change in direction $\boldsymbol{u}$ is

$$D_u f(\mathbf{x}) \;:=\; \lim_{t \to 0} \frac{f(\mathbf{x} + t\boldsymbol{u}) - f(\mathbf{x})}{t} \;=\; \langle \nabla f(\mathbf{x}), u \rangle.$$

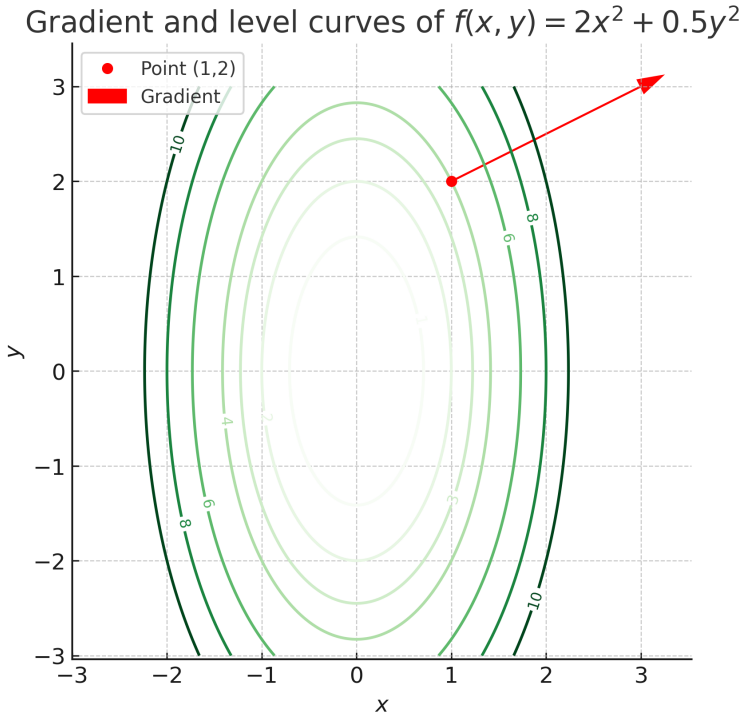
By Cauchy-Schwarz,

$$|D_u f(\mathbf{x})| \;\leq\; \|\nabla f(\mathbf{x})\|,$$

with equality if u points in the same direction as $\nabla f(\mathbf{x})$.

Conclusion: $\nabla f(\mathbf{x})$ points in the direction of steepest increase, $-\nabla f(\mathbf{x})$ in the direction of steepest decrease.

Example: gradient and level curves



Let $f(x, y) = 2x^2 + \frac{1}{2}y^2$.

$$\nabla f(x, y) = (4x, y).$$

At $(1, 2)$, $\nabla f = (4, 2)$.

Geometry: The gradient is perpendicular to the level curve

$$2x^2 + \frac{1}{2}y^2 = c$$

through $(1, 2)$.

Jacobian and matrix differentiation rules

Let $F : \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian at $\mathbf{x}$ is

$$\mathrm{J}F(\mathbf{x}) = \left[\frac{\partial F_i}{\partial x_j}(\mathbf{x})\right]_{i=1,\dots,m;\, j=1,\dots,n} \in \mathbb{R}^{m\times n}.$$

If $m = 1$, $F = f$, then $\mathrm{J}f(\mathbf{x}) = \nabla f(\mathbf{x})^\top$.

First-order approximation

$$F(\boldsymbol{x}+\boldsymbol{h}) \;=\; F(\boldsymbol{x}) + \mathrm{J}F(\boldsymbol{x})\boldsymbol{h} + o(\|\boldsymbol{h}\|).$$

Quick identities:

1. If $F(\mathbf{x}) = A\mathbf{x}$ with $A \in \mathbb{R}^{m\times n}$: $\mathrm{J}F = A$.
2. If $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$: $\nabla f = \mathbf{a}$.
3. If $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$: $\nabla f = (A + A^\top)\mathbf{x}$ ($= 2A\mathbf{x}$ if A symmetric).

Higher partial derivatives

Higher partials are iterated derivatives, e.g.

$$f_{xy}(x,y) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}(x,y)\right).$$

Young's Theorem:

If f_{xy} and f_{yx} are continuous, then $f_{xy} = f_{yx}$. We write $\frac{\partial^2 f}{\partial x \partial y}$.

Example: $f(x,y) = x^2 y + e^{xy}$ gives

$$f_{xy} = 2x + e^{xy}(1+xy) = f_{yx}.$$

Hessian $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{n\times n}$

This is a $n \times n$ matrix defined by $(\nabla^2 f(\mathbf{x}))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. (symmetric matrix!)

If $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x}$ with A symmetric, then $\nabla^2 f(\boldsymbol{x}) = 2A$.

Second order Taylor expansion

If $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}^n$ is C^2 then

$$f(\boldsymbol{x}+\boldsymbol{h}) \;=\; f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{h} + \tfrac{1}{2}\boldsymbol{h}^\top(\nabla^2 f(\boldsymbol{x}))\boldsymbol{h} + o(\|\boldsymbol{h}\|^2)$$

Note that $\boldsymbol{h}^\top(\nabla^2 f(\boldsymbol{x}))\boldsymbol{h}$ is a quadratic form in $\boldsymbol{h}$.

If you prefer working with indices (which you should not) this becomes

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{x}) h_i + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) h_i h_j + o(\|\boldsymbol{h}\|^2).$$

Elements of optimization

Unconstrained optimization

Local optima. $\mathbf{x}_0$ is a local max (min) if $f(\mathbf{x}) \leq (\geq)\ f(\mathbf{x}_0)$ nearby. If f is differentiable and $\mathbf{x}_0$ is an interior local extremum,

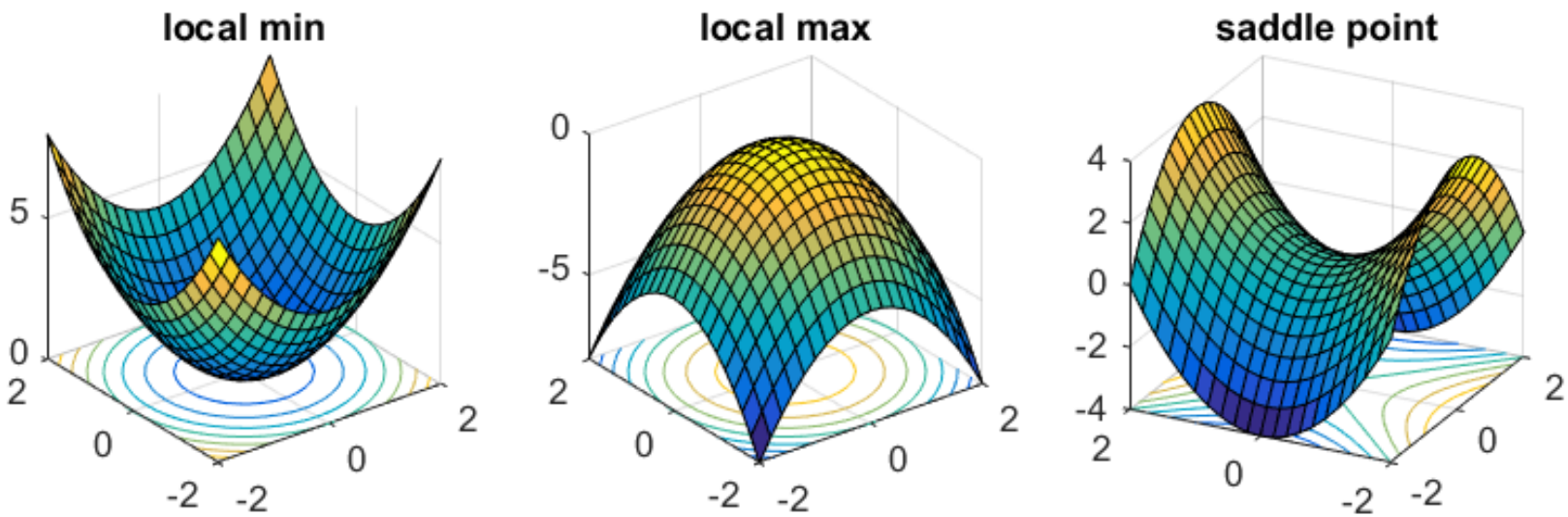
$$\nabla f(\mathbf{x}_0) = \mathbf{0} \quad \text{(first-order condition)}.$$

If $f \in C^2$, Taylor expansion at a stationary point $\mathbf{x}_0$ gives

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \tfrac{1}{2}\,\mathbf{h}^\top(\nabla^2 f(\boldsymbol{x}_0))\mathbf{h} + o(\|\mathbf{h}\|^2),$$

so the sign of the quadratic form $\mathbf{h}^\top(\nabla^2 f(\boldsymbol{x}_0))\mathbf{h}$ determines local behavior.

Unconstrained optimization (pictures)



Local optimality: second-order tests

Assume $f \in C^2$ and let $\nabla^2 f(\mathbf{x})$ be the Hessian matrix.

At a stationary point $\mathbf{x}_0$:

- $\nabla^2 f(\boldsymbol{x}_0)$ positive definite $\Rightarrow$ local minimum.
- $\nabla^2 f(\boldsymbol{x}_0)$ negative definite $\Rightarrow$ local maximum.
- $\nabla^2 f(\boldsymbol{x}_0)$ indefinite $\Rightarrow$ saddle.

$n = 2$ test: Let $D_2 = f_{xx} f_{yy} - f_{xy}^2$ at $\mathbf{x}_0$.

$$
\begin{aligned}
&D_2 > 0,\ \ f_{xx} > 0 \Rightarrow \text{local min},\\
&D_2 > 0,\ \ f_{xx} < 0 \Rightarrow \text{local max},\\
&D_2 < 0 \Rightarrow \text{saddle}, \quad D_2 = 0 :\ \ \text{inconclusive}.
\end{aligned}
$$

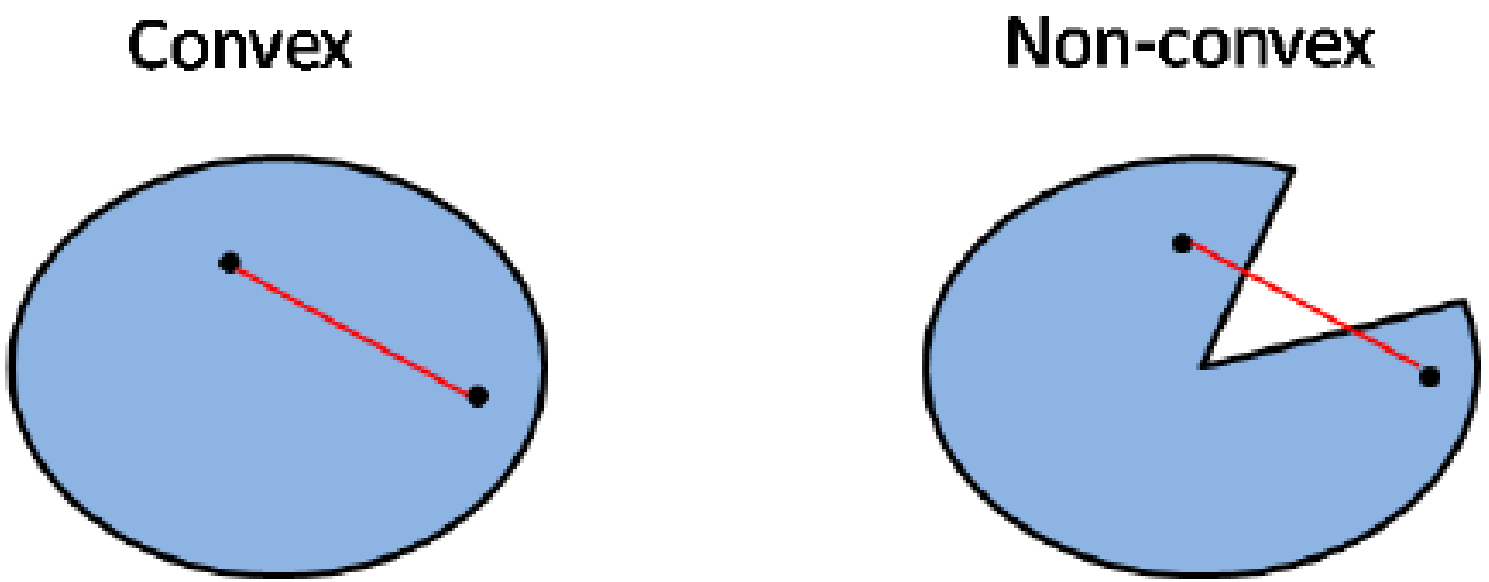
Examples

1. $f(x, y) = x^2 - y^2 - xy$. Then $\nabla f = (2x - y, -2y - x)$. The only stationary point is $(0, 0)$. The Hessian $\nabla^2 f(0, 0) = \begin{pmatrix} 2 & -1 \\ -1 & -2 \end{pmatrix}$ is indefinite $\Rightarrow$ $(0, 0)$ is a saddle.
2. $f(x, y) = x^2 + y^4$. Stationary point: $(0, 0)$. $\nabla^2 f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ is positive semidefinite; $f \geq 0$ so $(0, 0)$ is a global minimum.
3. $f(x, y) = x^3 + y^3$. Stationary point: $(0, 0)$. The Hessian at $(0, 0)$ is 0; the point is a saddle.

Convexity

Convex domains

A set $D \subset \mathbb{R}^n$ is convex if for any $\mathbf{x}, \mathbf{y} \in D$ and $t \in [0, 1]$, the point $(1 - t)\mathbf{x} + t\mathbf{y} \in D$.



Convexity, concavity, and global optimality

Definition

On a convex domain $D \subset \mathbb{R}^n$, f is convex if

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \;\leq\; \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in D,\ \theta \in [0,1].$$

Concave: reverse the inequality.

Curvature test (for C^2 functions)

$\nabla^2 f(\boldsymbol{x}) \succeq 0\ \forall \mathbf{x} \in D \iff f$ convex on D, $\quad \nabla^2 f(\boldsymbol{x}) \preceq 0 \iff f$ concave.
Strict definiteness $\Rightarrow$ strict convexity/concavity.

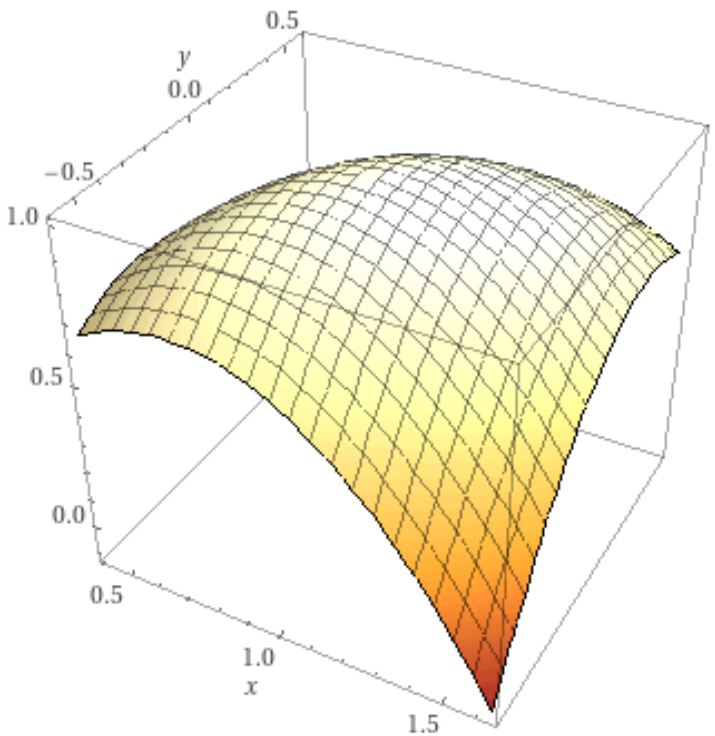
Important fact: If f is strictly convex (concave) then any local minimum (maximum) is a global minimum (maximum).

Example

Let $f(x, y) = 2x - y - x^2 + xy - y^2$. Then

$$\nabla f = (2 - 2x + y,\ -1 + x - 2y), \qquad \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}.$$

$\nabla^2 f(\boldsymbol{x})$ is negative definite $\Rightarrow$ f is strictly concave. The unique stationary point solves $\nabla f = \mathbf{0}$, giving $(x, y) = (0, 1)$, which is a global maximum.



Economic example: profit maximization

A firm sells products X/Y at 45/55 euros. Revenue $R(x, y) = 45x + 55y$. Cost

$$C(x, y) = 300 + x^2 + 1.5\, y^2 - 25x - 35y.$$

Profit $f(x, y) = R(x, y) - C(x, y)$. Then

$$f_x = -2x + 70, \qquad f_y = -3y + 90 \;\Rightarrow\; (x^*, y^*) = (35, 30).$$

Since $\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix}$ is negative definite everywhere, f is strictly concave and $(35, 30)$ is the global maximum. The maximal profit is $f(35, 30) = 2275$.

Least squares as orthogonal projection

Given data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and response $\boldsymbol{y} \in \mathbb{R}^n$, the least-squares estimator solves

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2, \qquad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}.$$

Geometric view (recall Lecture 2): $\hat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$, hence

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = \boldsymbol{0} \quad \Longleftrightarrow \quad (\boldsymbol{X}^\top\boldsymbol{X})\hat{\beta} = \boldsymbol{X}^\top\boldsymbol{y}.$$

Polynomial regression: A common use of least squares is fitting nonlinear trends by expanding the design matrix $\boldsymbol{X}$. For instance, with one predictor x, we can set

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{pmatrix},$$

so that the fitted model is $y \approx c_0 + c_1 x + \cdots + c_m x^m$.

Ridge regression: stabilizing high-variance fits

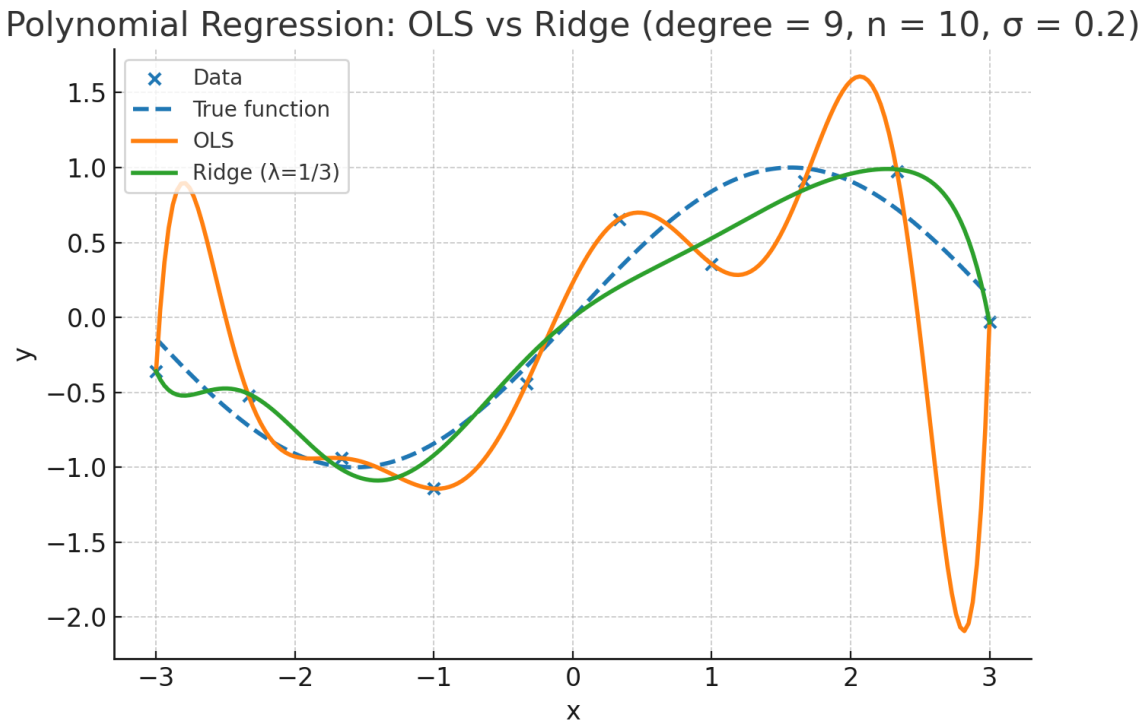
When $\boldsymbol{X}^\top \boldsymbol{X}$ is ill-conditioned or d is large, add ℓ_2 regularization:

$$\hat{\beta}_\lambda = \arg\min_\beta \left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \|\beta\|^2 \right) \quad \Longrightarrow \quad \hat{\beta}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Overfitting vs. ridge: degree-9 polynomial demo

$n = 10$ points from $y = \sin x + \varepsilon$ on $[-3, 3]$, degree 9 polynomial.

OLS (no penalty) vs. Ridge with $\lambda = \frac{1}{3}$.



OLS wiggles violently between sparse points; ridge damps the high-frequency coefficients and tracks the signal more smoothly.

Modern applied examples (multivariable)

- Portfolio risk (mean–variance):

$$f(\mathbf{w}) = \mathbf{w}^\top \Sigma\, \mathbf{w}, \quad g(\mathbf{w}) = \mu^\top \mathbf{w}, \quad \mathbf{w} \in \mathbb{R}^n,\ \sum_i w_i = 1,\ w_i \geq 0.$$

- Logistic regression (binary choice):

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big(\log\left(1 + e^{x_i^\top \beta}\right) - y_i\, x_i^\top \beta \Big) \quad \text{(convex in } \beta\text{)}.$$

- CES utility/production:

$$U(x) = \Big(\sum_{i=1}^{n} \alpha_i x_i^\rho \Big)^{1/\rho}, \quad P(L,K) = A(\theta)\, L^\alpha K^\beta.$$

Numerical optimization

Gradient descent

Goal: minimize $f(\boldsymbol{x})$.

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t).$$

Pieces you pick:

- **Step size** η_t: constant, diminishing, or via backtracking.
- **Stop** when $\|\nabla f(\boldsymbol{x}_t)\|$ small.

In practice: a good η_t schedule matter a lot.

GD on least squares (closed form vs iterations)

Least squares problem has a closed form solution. This still requires inverting a potentially large matrix $\boldsymbol{X}^\top \boldsymbol{X}$. GD gives an alternative way to find a solution.

$$f(\beta) = \frac{1}{n}\|\boldsymbol{X}\beta - \mathbf{y}\|^2, \qquad \nabla f(\beta) = \frac{2}{n}\boldsymbol{X}^\top(\boldsymbol{X}\beta - \mathbf{y}).$$

GD update:

$$\beta_{t+1} = \beta_t - \eta \frac{2}{n}\boldsymbol{X}^\top(\boldsymbol{X}\beta_t - \mathbf{y}).$$

Ridge:

$$f_\lambda(\beta) = \frac{1}{n}\|\boldsymbol{X}\beta - \mathbf{y}\|^2 + \lambda\|\beta\|_2^2, \quad \nabla f_\lambda = \frac{2}{n}\boldsymbol{X}^\top(\boldsymbol{X}\beta - \mathbf{y}) + 2\lambda\beta.$$

Closed form exists $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$, but GD scales better to huge n, p or streaming data.

When to use second order methods?

In general, we update $\boldsymbol{x}_t$ as

$$\boldsymbol{x}_{t+1} := \arg\min f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} \rangle + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_t)^\top K (\boldsymbol{x} - \boldsymbol{x}_t).$$

If $K = I_n$, we recover gradient descent.

If $K = \nabla \nabla^\top f(\boldsymbol{x}_t)$, we get the Newton method.

- **Newton:** $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - H^{-1}(\boldsymbol{x}_t) \nabla J(\boldsymbol{x}_t)$ (fast near solution, expensive to form/solve).
- **Quasi-Newton (e.g., BFGS, L-BFGS):** approximate H^{-1} from gradients only; great for medium scale convex problems.
- **Takeaway:** for huge data/models use (S)GD; for smaller smooth convex problems, quasi-Newton shines.

Constrained optimization

Constrained optimization: Lagrange and KKT (teaser)

We want to maximize/minimize $f(\boldsymbol{x})$ subject to constraints:

$$g_i(\boldsymbol{x}) = 0 \quad \text{(equalities)}, \qquad h_j(\boldsymbol{x}) \leq 0 \quad \text{(inequalities)}.$$

Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \mu_j h_j(\boldsymbol{x}).$$

KKT conditions (when applicable):

1. **Stationarity:** $\nabla_x \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0.$
2. **Primal feasibility:** $g_i(\boldsymbol{x}^*) = 0,\ h_j(\boldsymbol{x}^*) \leq 0.$
3. **Dual feasibility:** $\mu_j^* \geq 0.$
4. **Complementary slackness:** $\mu_j^*\, h_j(\boldsymbol{x}^*) = 0.$

Constrained optimization in 3D (example)

Problem: maximize

$$f(x,y,z) = xyz \quad \text{subject to} \quad g(x,y,z) = x+y+z-1 = 0, \quad x,y,z \geq 0.$$

Lagrangian:

$$\mathcal{L}(x,y,z,\lambda,\mu_1,\mu_2,\mu_3) = xyz + \lambda(1-x-y-z) - \mu_1 x - \mu_2 y - \mu_3 z.$$

Stationarity:

$$\begin{aligned}
\partial_x \mathcal{L} &= yz - \lambda - \mu_1 = 0, \\
\partial_y \mathcal{L} &= xz - \lambda - \mu_2 = 0, \\
\partial_z \mathcal{L} &= xy - \lambda - \mu_3 = 0.
\end{aligned}$$

Complementary slackness: $\mu_1 x = \mu_2 y = \mu_3 z = 0$.

Solution and geometry

At an interior solution $(x, y, z > 0)$, complementary slackness $\Rightarrow \mu_1 = \mu_2 = \mu_3 = 0$.

Simplified system:

$$\begin{aligned} yz &= \lambda, \\ xz &= \lambda, \\ xy &= \lambda, \\ x + y + z &= 1. \end{aligned}$$

Deduction: $yz = xz = xy \Rightarrow x = y = z$. Constraint $\Rightarrow x = y = z = \frac{1}{3}$.

Maximizer: $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, value $f^* = \frac{1}{27}$.

Geometric intuition:

- Constraint $x + y + z = 1$ is a plane.
- Level sets $xyz = c$ are curved surfaces.
- At optimum: $\nabla f = (yz, xz, xy)$ is parallel to $\nabla g = (1, 1, 1)$ — the plane just *touches* the surface.