Lecture 4: Calculus and Linear Algebra

Piotr Zwiernik

Mathematics Brush-up

Barcelona School of Economics

Barcelona School of Economics

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

Examples:

- **Economics:** Cobb–Douglas production $Y = K^\alpha L^{1-\alpha}$, or utility $U(x, y)$.
- **Data science:** Loss functions $L(\theta_1, \ldots, \theta_d)$ depending on many parameters.

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

What is a multivariable function?

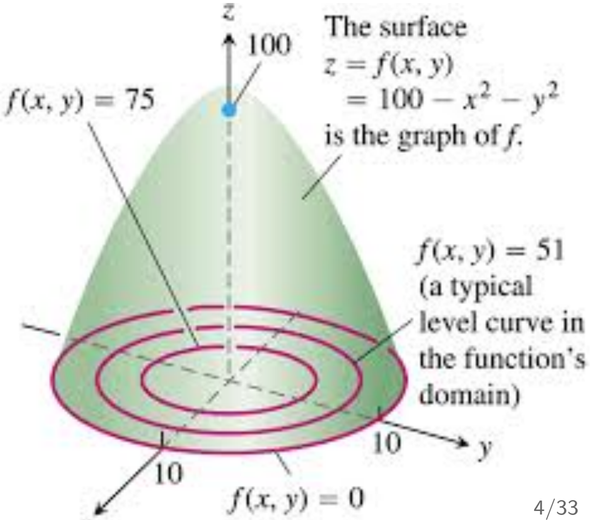
A function of n variables is a map

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

- For $n = 2$: graph $z = f(x, y)$ is a surface in $\mathbb{R}^3$.
- Level curves (contours):
$$\{(x, y) \in D : f(x, y) = c\}.$$
- For $n > 2$: use slices or projections to visualize.

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

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



Economic example: Cobb–Douglas

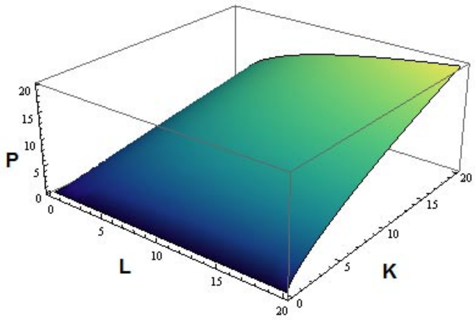
Cobb–Douglas production function

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

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

Domain:

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



Returns to scale in Cobb–Douglas

- α (resp. β) measures the % change in output after a 1% change in labour (resp. capital), ceteris paribus.
- If $\alpha + \beta = 1$, there are constant returns to scale: scaling (L, K) by $t > 0$ scales P by t.
- If $\alpha + \beta < 1$, decreasing returns; if $\alpha + \beta > 1$, increasing returns.

Multivariate Gaussian density

A random vector $\mathbf{X} \in \mathbb{R}^d$ is multivariate normal with mean $\mathbf{m}$ and positive definite covariance Σ if

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

where

- $\mathbf{m} = E(\mathbf{X})$ is the mean,
- $\Sigma = E\big((\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^\top\big)$ is the covariance matrix.

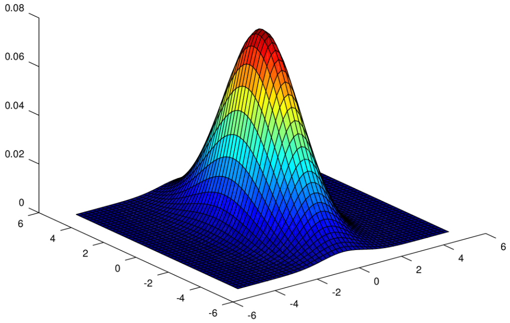
Note: f is strictly positive. It depends on $\mathbf{x}$ through the Mahalonobis distance

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

Thus, the level sets are the elipsoids $\{\mathbf{x} : \|\mathbf{x} - \boldsymbol{m}\| = \mathrm{const}\}$.

Multivariate Gaussian density

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



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 $C_y(x,y) = 24\sqrt{y}$.

Cobb–Douglas: marginal productivities

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

$$\frac{\partial P}{\partial L} = \alpha\, \frac{P}{L}, \qquad \frac{\partial P}{\partial K} = \beta\, \frac{P}{K}.$$

Interpretation: marginal productivity of labour/capital is proportional to average productivity per unit. Under suitable regularity, these proportionality laws lead back to the Cobb–Douglas form.

Tangent plane and linear approximation

Geometrically, $f_x(x_0, y_0)$ (resp. $f_y(x_0, y_0)$) is the slope of the tangent to the curve cut by the plane $y = y_0$ (resp. $x = x_0$) at (x_0, y_0, z_0), where $z_0 = f(x_0, y_0)$.

The tangent plane at (x_0, y_0, z_0) is

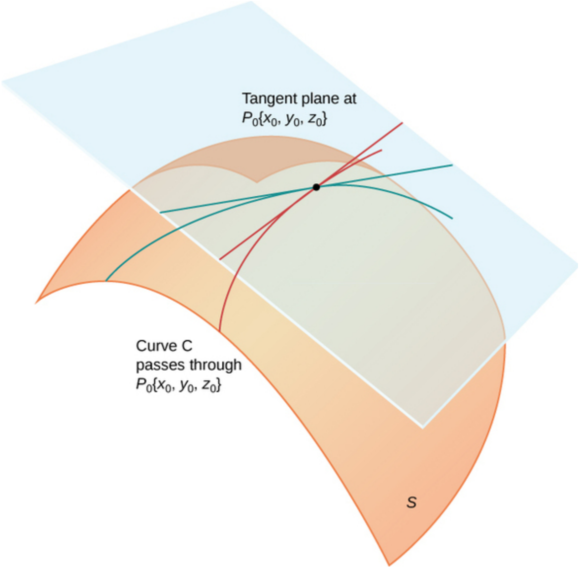
$$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Linear (first-order) approximation near (x_0, y_0):

$$f(x, y) \approx f(x_0, y_0) + f_x(x_0, y_0)\,\Delta x + f_y(x_0, y_0)\,\Delta y,$$

In differential notation: $df \approx f_x\,dx + f_y\,dy$.

Tangent plane (visual)



Higher partial derivatives

Higher derivatives are defined by iterating partials, e.g.

$$f_{xy}(x,y) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}(x,y)\right) = \frac{\partial^2 f}{\partial y\,\partial x}(x,y).$$

Young's theorem: If f_{xy} and f_{yx} are continuous near a point, then $f_{xy} = f_{yx}$ there.

Example: $f(x,y) = \sin(3x-y) \Rightarrow f_{xy} = f_{yx} = 3\sin(3x-y)$.

The gradient and linear approximation

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

$$\nabla f(\mathbf{x}) = \big(f_{x_1}(\mathbf{x}), \ldots, f_{x_n}(\mathbf{x})\big).$$

best linear approximation

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

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

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

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

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

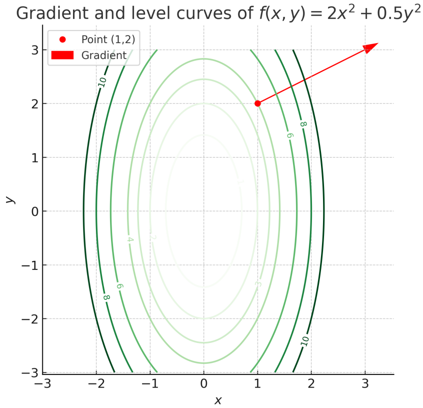
By Cauchy–Schwarz,

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

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 $(1, 2)$.

Note: ∇f is always normal to level sets. Why?

Jacobian and matrix differentiation rules

Let $F : \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian matrix of F at $\mathbf{x} \in \mathbb{R}^n$ 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 $m = 1$, then $F = f : \mathbb{R}^n \to \mathbb{R}$ and $\mathrm{J}f(\mathbf{x}) = \nabla f(\mathbf{x})^\top$.

Useful identities:

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

Unconstrained optimization

A point $\mathbf{x}_0$ is a local maximum (minimum) if there exists a ball $B_r(\mathbf{x}_0) \subset D$ such that

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \quad (\text{resp. } f(\mathbf{x}) \geq f(\mathbf{x}_0)) \quad \text{for all } \mathbf{x} \in B_r(\mathbf{x}_0).$$

If this holds on all of D, the optimum is global.

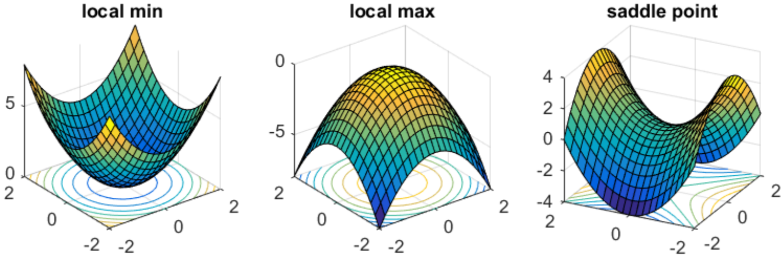
If $\mathbf{x}_0$ is an interior local extremum and f is differentiable, then the first-order condition holds:

$$\nabla f(\mathbf{x}_0) = \mathbf{0}.$$

Such points are stationary; a stationary point that is neither max nor min is a saddle.

Indeed: By Slide 14, if $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, an infinitesimal move can increase/decrease the value of f.

Unconstrained optimization (pictures)



Local optimality: second-order tests

Assume $f \in C^2$ and let $H_f(\mathbf{x}) = [f_{x_i x_j}(\mathbf{x})]_{i,j}$ be the (symmetric) Hessian.

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

- $H_f(\mathbf{x}_0)$ positive definite $\Rightarrow$ local minimum.
- $H_f(\mathbf{x}_0)$ negative definite $\Rightarrow$ local maximum.
- $H_f(\mathbf{x}_0)$ indefinite $\Rightarrow$ saddle.

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

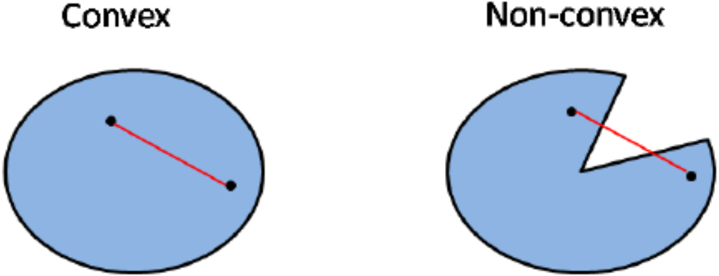
$$\begin{aligned}
&D_2 > 0,\ \ f_{xx} > 0 \Rightarrow \text{local min},\\
&D_2 > 0,\ \ f_{xx} < 0 \Rightarrow \text{local max},\\
&D_2 < 0 \Rightarrow \text{saddle}, \quad D_2 = 0:\ \ \text{inconclusive}.
\end{aligned}$$

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 $H_f = \begin{pmatrix} 2 & -1 \\ -1 & -2 \end{pmatrix}$ is indefinite $\Rightarrow$ $(0,0)$ is a saddle.
2. $f(x,y) = x^2 + y^4$. Stationary point: $(0,0)$. $H_f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ 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.

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 (Convexity/Concavity)

Let $D \subset \mathbb{R}^n$ be convex. A function $f : D \to \mathbb{R}$ is convex if for all $\mathbf{x}, \mathbf{y} \in D$ and $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \;\leq\; \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}),$$

i.e. the graph lies *below* every chord.

- f is concave if the inequality is reversed, i.e. the graph lies *above* every chord.

Curvature test (for C^2 functions)

1. $H_f(\mathbf{x}) \succeq 0$ on $D \;\Leftrightarrow\; f$ convex. $H_f(\mathbf{x}) \preceq 0$ on $D \;\Leftrightarrow\; f$ concave.
2. Strict definiteness implies strict convexity/concavity.

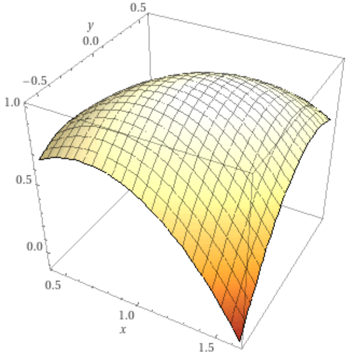
Key fact: If f is convex (concave) on D, any stationary point is a **global** minimum (maximum).

Example

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

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

H_f 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 $H_f = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix}$ 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 $X \in \mathbb{R}^{n\times d}$ and response $y \in \mathbb{R}^n$, the least–squares estimator solves

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2, \qquad \hat{y} = X\hat{\beta}.$$

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

$$X^\top(y - X\hat{\beta}) = 0 \quad \Longleftrightarrow \quad (X^\top X)\hat{\beta} = X^\top y \;\; (\text{if } X^\top X \text{ invertible}).$$

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

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

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

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

Spectral view: if $X^\top X = U\mathrm{diag}(s_1^2, \dots, s_d^2)U^\top$, then

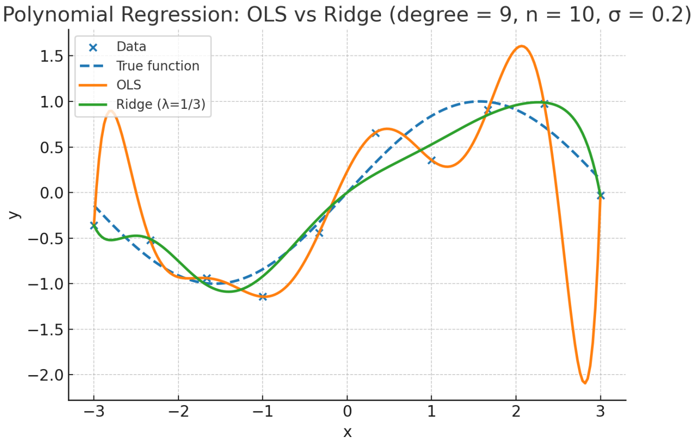
$$\hat{\beta}_\lambda = \sum_{j=1}^{d} \frac{s_j}{s_j^2 + \lambda}\, u_j \,\langle y,\, Xu_j\rangle,$$

so ridge shrinks directions with small s_j (low variance) the most $\left(\frac{s_j}{s_j^2+\lambda} < \frac{1}{s_j}\right)$, reducing variance and overfitting.

Overfitting vs. ridge: degree-9 polynomial demo

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

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



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:

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

Gradient descent

Goal: minimize $f(\theta)$.

$$\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t).$$

Pieces you pick:

- **Step size** η_t: constant, diminishing, or via backtracking.
- **Stop** when $\|\nabla f(\theta_t)\|$ small.

In practice: feature scaling and 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 $X^\top X$. GD gives an alternative way to find a solution.

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

GD update:

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

Ridge:

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

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

Constrained optimization: Lagrange and KKT (teaser)

Equality constraints $g_i(x) = 0$ and inequality constraints $h_j(x) \le 0$. The Lagrangian:

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

KKT conditions (when they apply):

- **Stationarity:** $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \mathbf{0}$.
- **Primal feasibility:** $g_i(x^*) = 0,\ h_j(x^*) \le 0$.
- **Dual feasibility:** $\mu_j^* \ge 0$.
- **Complementary slackness:** $\mu_j^* \, h_j(x^*) = 0$.

Example (budgeted utility max): maximize $U(x)$ s.t. $p^\top x \le B$. Then $\mathcal{L}(x, \mu) = U(x) + \mu\,(B - p^\top x)$ and at optimum $\nabla U(x^*) = \mu^* p$, $p^\top x^* \le B$, $\mu^* \ge 0$, $\mu^*(B - p^\top x^*) = 0$.

When to use second order methods?

In general, we update θ_t as

$$\theta_{t+1} \; := \; \arg\min f(\theta_t) + \langle \nabla f(\theta_t), \theta \rangle + \frac{1}{2}(\theta - \theta_t)^\top K (\theta - \theta_t).$$

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

If $K = \nabla\nabla^\top f(\theta_t)$, we get the Newton method.

- **Newton:** $\theta_{t+1} = \theta_t - H^{-1}(\theta_t)\nabla J(\theta_t)$ (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.