DegreeLive

Bachelor of Commerce (B2001) | Level 3 • Part B Major | Monash University

Introduction to machine learning

Course Code: ETC3250 • Module 2 Notes

Module 2: Regularised Linear Models, Splines & Generalised Additive Models

2.1 Beyond Ordinary Least Squares: High Dimensionality and Variance Inflation

The classical multiple linear regression model relates a continuous response Y to p predictors $X_1, \dots, X_p$ via:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \dots, n$$

Where:

- y_i is the observed scalar response for observation i.
- x_{ij} is the value of the j-th predictor for observation i.
- β_0 is the intercept and $\beta_1, \dots, \beta_p$ are the unknown slope coefficients.
- ε_i is a random error term with $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$.

Ordinary Least Squares (OLS) estimates $\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y$ by minimising the Residual Sum of Squares $\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$. By the Gauss–Markov theorem, OLS achieves the minimum variance among all linear *unbiased* estimators. However, unbiasedness does not guarantee minimal test MSE when p is large relative to n or when predictors exhibit strong collinearity:

- **High-Variance Instability ($p \approx n$):** As the number of predictors approaches the sample size, OLS fits noise directions aggressively, inflating $\mathrm{Var}(\hat{\beta}_{\mathrm{OLS}}) = \sigma_\varepsilon^2 (X^\top X)^{-1}$ and degrading out-of-sample forecasts.
- **Singularity ($p > n$):** In high-dimensional economic or text datasets (for example, predicting ASX earnings surprises from hundreds of financial ratios and NLP sentiment tokens), $X^\top X$ is rank-deficient and non-invertible, yielding infinitely many zero-residual solutions.
- **Multicollinearity:** When predictors are strongly correlated (such as multiple lagged ABS inflation sub-indices), $(X^\top X)^{-1}$ has enormous eigenvalues, producing wildly oscillating coefficient estimates with opposite signs.
- **Model Interpretability:** Retaining dozens of irrelevant features obscures the genuine drivers of the business problem.

Three classical remedies address these limitations: **subset selection** (discarding predictors), **shrinkage / regularisation** (constraining coefficient magnitudes toward zero), and **dimension reduction** (Module 4).

2.2 Subset Selection and Information Criteria

Subset selection methods restrict the linear model to a subset of $k \leq p$ predictors. Because comparing 2^p candidate models via training RSS always favours the largest model (RSS never increases when adding a feature), selection requires criteria that adjust for model complexity or estimate test error directly.

Best Subset vs. Stepwise Algorithms

- **Best Subset Selection:** Evaluates all $\binom{p}{k}$ models of size k for each $k = 0, 1, \dots, p$, recording the lowest-RSS model M_k at each size, then selects among $\mathrm{M}_0, \dots, \mathrm{M}_p$ using cross-validation or an information criterion. Optimal in-sample for each k, yet computationally infeasible once $p > 30$ because 2^p exceeds one billion.
- **Forward Stepwise Selection:** Starts from the null model M_0 (intercept only) and greedily adds one predictor at each step that produces the greatest reduction in RSS. Requires only $1 + p(p+1)/2$ model fits and remains feasible even when $p > n$ (stopping at $k < n$).
- **Backward Stepwise Selection:** Starts from the full model M_p and iteratively removes the least statistically useful predictor (smallest increase in RSS). Requires $n > p$ so the initial full model is uniquely estimable.

DegreeLive

Analytical Model Selection Criteria

Rather than running full k-fold CV across every subset size, analysts frequently adjust training RSS using theoretical penalties for effective parameter count d:

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right), \qquad \text{AIC} = n\ln\left(\frac{\text{RSS}}{n}\right) + 2d, \qquad \text{BIC} = n\ln\left(\frac{\text{RSS}}{n}\right) + d\ln(n)$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

Where:

- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares of the candidate model, and $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.
- d is the number of estimated slope predictors in the candidate model (for Gaussian linear models, $d + 1$ parameters excluding variance, though constant offsets do not alter ranking).
- $\hat{\sigma}^2$ is an unbiased estimate of the irreducible error variance σ_ε^2 typically obtained from the full model M_p.
- n is the number of training observations.

Notice the fundamental distinction between AIC and BIC: whenever $n \geq 8$, $\ln(n) > 2$ (for instance, $\ln(200) \approx 5.30 > 2$). Consequently, BIC imposes a substantially heavier penalty per additional predictor than AIC or C_p, favouring more parsimonious models. Under standard regularity conditions, BIC is *selection consistent* (recovering the true sparse model with probability approaching 1 as $n \to \infty$), whereas AIC / C_p behaves asymptotically equivalently to LOOCV and tends to minimise prediction error when the true data-generating process has many small non-zero effects.

Worked Example 2.1: Comparing Subsets Using C_p, AIC, BIC, and Adjusted R^2

An Australian property valuation team models suburban house sale prices across $n = 100$ transactions with $\text{TSS} = 5{,}000$ and full-model error variance estimate $\hat{\sigma}^2 = 10.0$. Three candidate nested models yield the following training RSS values: Model 1 ($d = 2$, $\text{RSS} = 1{,}120$), Model 2 ($d = 4$, $\text{RSS} = 980$), and Model 3 ($d = 8$, $\text{RSS} = 935$). Determine which model each criterion selects (note $\ln(100) \approx 4.605$).

Candidate Model	d	RSS	C_p (lower better)	AIC (lower better)	BIC (lower better)	Adjusted R^2 (higher better)
Model 1	2	1120	$(1120 + 40)/100 = 11.60$	$100\ln(11.2) + 4 = 245.59$	$100\ln(11.2) + 9.21 = 250.80$	$1 - (1120/97)/(5000/99) = 0.7714$
Model 2	4	980	$(980 + 80)/100 =$ **10.60**	$100\ln(9.8) + 8 =$ **236.24**	$100\ln(9.8) + 18.42 =$ **246.66**	$1 - (980/95)/(5000/99) = 0.7957$
Model 3	8	935	$(935 + 160)/100 = 10.95$	$100\ln(9.35) + 16 = 239.54$	$100\ln(9.35) + 36.84 = 260.38$	$1 - (935/91)/(5000/99) =$ **0.7966**

All three penalised prediction/parsimony criteria (C_p, AIC, BIC) reject the extra four predictors in Model 3 and select **Model 2** ($d = 4$). Adjusted R^2 improves slightly to 0.7966 for Model 3 because its degrees-of-freedom penalty is weaker than C_p and BIC.

2.3 Ridge Regression (L_2 Regularisation)

Subset selection makes discrete keep-or-drop decisions that can suffer from high selection variance: small perturbations in the training sample may cause completely different variables to enter. **Shrinkage methods** avoid this discontinuity by fitting all p predictors while continuously shrinking estimated coefficients toward zero.

Ridge Regression (Hoerl and Kennard, 1970) augments the OLS objective with a quadratic L_2 penalty on the slope coefficients (never penalising the intercept β_0). Assuming predictors are centred and standardised (so $\hat{\beta}_0 = \bar{y}$), Ridge minimises:

$$\mathrm{J}_{\text{ridge}}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \|y - \beta_0 \mathbf{1} - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Where:

- $\lambda \geq 0$ is the tuning (regularisation) hyperparameter controlling shrinkage intensity. When $\lambda = 0$, Ridge reduces exactly to OLS; as $\lambda \to \infty$, all slope estimates shrink continuously toward zero, approaching the null intercept-only model.
- $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the squared Euclidean (L_2) norm of the slope vector.

Matrix Solution, Scale Sensitivity, and Effective Degrees of Freedom

Differentiating the centred Ridge objective with respect to β and setting the gradient to zero yields a closed-form linear system:

$$\hat{\beta}_{\text{ridge}}(\lambda) = \left(X^\top X + \lambda I_p\right)^{-1} X^\top y$$

Where:

- $X \in \mathrm{R}^{n \times p}$ is the standardised predictor matrix (each column has mean 0 and unit sample standard deviation, or unit L_2 norm).
- I_p is the $p \times p$ identity matrix.

Adding $\lambda > 0$ to every diagonal entry guarantees that $X^\top X + \lambda I_p$ is strictly positive definite and invertible even when $p > n$ or predictors are perfectly collinear. Crucially, unlike OLS, Ridge estimates are **not scale-equivariant**: multiplying predictor X_j by 10 alters $\hat{\beta}_j^2$ by a factor of 100 under the same λ. Consequently, standardising every predictor to unit variance inside the cross-validation pipeline is mandatory before applying Ridge.

From a Bayesian standpoint, $\hat{\beta}_{\text{ridge}}$ coincides exactly with the posterior mode (and posterior mean) when slope coefficients share an independent zero-mean Gaussian prior $\beta_j \sim \mathrm{N}(0, \tau^2)$ and noise is $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, with $\lambda = \sigma_\varepsilon^2 / \tau^2$. Furthermore, the fitted vector $\hat{y}_{\text{ridge}} = H_\lambda y$ relies on the Ridge hat matrix $H_\lambda = X(X^\top X + \lambda I_p)^{-1} X^\top$. Using the singular values $d_1, \ldots, d_p$ of X, the **effective degrees of freedom** of Ridge regression is:

$$\mathrm{df}(\lambda) = \mathrm{tr}(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Where d_j are the singular values of the centred design matrix X. When $\lambda = 0$, $\mathrm{df}(0) = p$; as λ increases, $\mathrm{df}(\lambda)$ declines monotonically toward 0, directly trading a modest rise in squared bias for substantial variance reduction.

Singular Value Decomposition View of Ridge Shrinkage

Decomposing the centred design matrix via Singular Value Decomposition $X = UDV^\top$ (where $U^\top U = I_p$, $V^\top V = V V^\top = I_p$, and $D = \mathrm{diag}(d_1, \ldots, d_p)$ contains ordered singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$) reveals precisely which directions Ridge shrinks most aggressively:

DegreeLive

$$X\hat{\beta}_{\text{ridge}} = UD(D^2 + \lambda I_p)^{-1}DU^\top y = \sum_{j=1}^{p} u_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) u_j^\top y$$

Where u_j is the j-th left singular vector (corresponding to the j-th principal component direction of X, with sample variance proportional to d_j^2). Notice that the shrinkage factor $\frac{d_j^2}{d_j^2+\lambda}$ is close to 1 along high-variance principal directions (large d_j^2) and shrinks sharply toward 0 along low-variance, multicollinear directions (small d_j^2). Ridge therefore preserves dominant signal axes while suppressing unstable, near-singular directions.

Worked Example 2.2A: Effective Degrees of Freedom Along Singular Values

Suppose a standardised $n \times 3$ macroeconomic design matrix has singular values $d_1 = 8.0$, $d_2 = 3.0$, and $d_3 = 1.0$ (so $d_1^2 = 64, d_2^2 = 9, d_3^2 = 1$). Compute the directional shrinkage factors and total effective degrees of freedom $\text{df}(\lambda)$ at $\lambda = 4$:

- Direction 1 (dominant signal): $\frac{64}{64+4} = \frac{64}{68} \approx 0.941$ (shrunk by only 5.9%).
- Direction 2 (moderate signal): $\frac{9}{9+4} = \frac{9}{13} \approx 0.692$ (shrunk by 30.8%).
- Direction 3 (collinear direction): $\frac{1}{1+4} = \frac{1}{5} = 0.200$ (shrunk by 80.0%).

$$\text{df}(4) = 0.941 + 0.692 + 0.200 = 1.833$$

Out of $p = 3$ nominal parameters, Ridge at $\lambda = 4$ uses only 1.833 effective degrees of freedom, concentrating regularisation heavily on the noisy third direction.

Worked Example 2.2: Exact Ridge Calculation and Variance Reduction

Suppose $X^\top X$ for two standardised predictors (Australian consumer confidence X_1 and ASX volatility index X_2) is nearly collinear, with:

$$X^\top X = \begin{pmatrix} 10 & 9 \\ 9 & 10 \end{pmatrix}, \qquad X^\top y = \begin{pmatrix} 21 \\ 17 \end{pmatrix}$$

Compute both the OLS solution ($\lambda = 0$) and the Ridge solution with $\lambda = 3$, and verify the shrinkage in coefficient norm.

1. **OLS Estimate ($\lambda = 0$):** The determinant of $X^\top X$ is $10(10) - 9(9) = 19$. Inverting gives:

$$\hat{\beta}_{\text{OLS}} = \frac{1}{19}\begin{pmatrix} 10 & -9 \\ -9 & 10 \end{pmatrix}\begin{pmatrix} 21 \\ 17 \end{pmatrix} = \frac{1}{19}\begin{pmatrix} 210 - 153 \\ -189 + 170 \end{pmatrix} = \begin{pmatrix} 57/19 \\ -19/19 \end{pmatrix} = \begin{pmatrix} 3.00 \\ -1.00 \end{pmatrix}$$

   Notice the large opposing signs induced by strong collinearity ($r = 0.90$), with squared norm $\|\hat{\beta}_{\text{OLS}}\|_2^2 = 3^2 + (-1)^2 = 10.0$.

2. **Ridge Estimate ($\lambda = 3$):** Adding λI_2 yields $X^\top X + 3I_2 = \begin{pmatrix} 13 & 9 \\ 9 & 13 \end{pmatrix}$, whose determinant is $13^2 - 9^2 = 169 - 81 = 88$ (over four times better conditioned). Inverting gives:

$$\hat{\beta}_{\text{ridge}} = \frac{1}{88}\begin{pmatrix} 13 & -9 \\ -9 & 13 \end{pmatrix}\begin{pmatrix} 21 \\ 17 \end{pmatrix} = \frac{1}{88}\begin{pmatrix} 273 - 153 \\ -189 + 221 \end{pmatrix} = \begin{pmatrix} 120/88 \\ 32/88 \end{pmatrix} \approx \begin{pmatrix} 1.364 \\ 0.364 \end{pmatrix}$$

   Now $\|\hat{\beta}_{\text{ridge}}\|_2^2 = 1.364^2 + 0.364^2 \approx 1.993$, dramatically stabilising the estimates and eliminating the spurious negative coefficient on X_2.

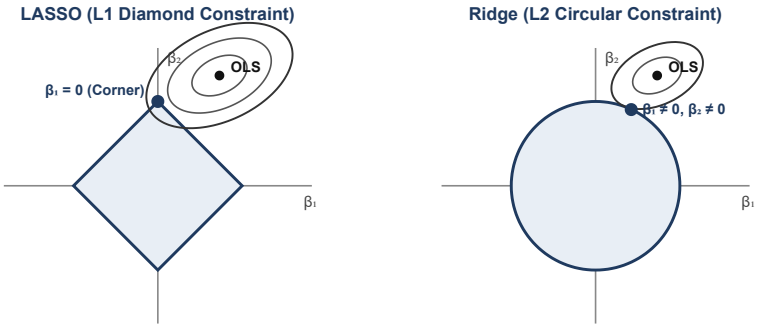
DegreeLive

2.4 The LASSO (L_1 Regularisation) and Sparsity Geometry

While Ridge regression shrinks collinear coefficients stably toward zero, it never sets any coefficient *exactly* to zero unless $\lambda = \infty$. In high-dimensional problems with hundreds of irrelevant variables, analysts often prefer sparse models that simultaneously regularise and perform automatic feature selection. The **Least Absolute Shrinkage and Selection Operator (LASSO)** (Tibshirani, 1996) replaces the L_2 penalty with an L_1 penalty:

$$\mathrm{J}_{\text{lasso}}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where $\lambda \geq 0$ controls the strength of the L_1 penalty $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. Equivalently, in constrained optimisation form, LASSO minimises RSS subject to $\sum_{j=1}^{p} |\beta_j| \leq s$, whereas Ridge minimises RSS subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$.



The geometric diagram above explains why LASSO induces sparsity while Ridge does not. The elliptical contours represent constant levels of RSS centred around $\hat{\beta}_{\text{OLS}}$. As the RSS ellipse expands outward to make first contact with the shaded constraint region, the smooth circular boundary of Ridge almost never hits an axis intercept, keeping both $\hat{\beta}_1$ and $\hat{\beta}_2$ non-zero. By contrast, the sharp diamond corners of the L_1 constraint sit directly on the coordinate axes, making corner tangency—where one or more coefficients equal zero identically—highly probable.

Soft-Thresholding Solution in Orthonormal Design

Although general LASSO estimation has no closed-form matrix inverse and is solved efficiently via coordinate descent, an exact closed form exists when predictors are orthonormal ($X^\top X = I_p$, using factor $1/2$ scaling on RSS: $\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$):

$$\hat{\beta}_j^{\text{lasso}} = \mathrm{S}_\lambda(\hat{\beta}_j^{\text{OLS}}) = \text{sign}(\hat{\beta}_j^{\text{OLS}}) \max(|\hat{\beta}_j^{\text{OLS}}| - \lambda,\ 0)$$

Where:

- $\hat{\beta}_j^{\text{OLS}}$ is the ordinary least squares estimate for orthonormal predictor j.
- $\mathrm{S}_\lambda(\cdot)$ is the soft-thresholding operator, which shrinks every coefficient toward zero by a constant amount λ and sets any coefficient with $|\hat{\beta}_j^{\text{OLS}}| \leq \lambda$ identically to zero.

2.5 Elastic Net Regularisation and Tuning via Cross-Validation

While LASSO delivers valuable sparsity, it exhibits two weaknesses: (1) when $p > n$, LASSO can select at most n variables before saturating, and (2) when a group of predictors is highly correlated (such as multiple retail price indices), LASSO tends to pick one arbitrary member of the group and zero out the rest, producing unstable variable selection across samples. **Elastic Net** (Zou and Hastie, 2005) resolves both issues by taking a convex combination of L_1 and L_2 penalties:

Degree

$$J_{\text{enet}}(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\left[\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right]$$

Where:

- $\lambda \geq 0$ controls overall regularisation strength.
- $\alpha \in [0, 1]$ is the mixing parameter: $\alpha = 1$ recovers LASSO, $\alpha = 0$ recovers Ridge, and $0 < \alpha < 1$ induces both sparsity (via the L_1 corner geometry) and the **grouping effect** (via strict convexity of the L_2 term, assigning similar weights to strongly correlated features).

Choosing λ: Minimum CV vs. the One-Standard-Error Rule

Across a logarithmic grid of candidate values $\lambda_1 > \lambda_2 > \cdots > \lambda_M$, k-fold cross-validation estimates $\text{CV}(\lambda_m)$ and its fold-level standard error $\text{SE}(\lambda_m)$. Two principled selection rules exist:

- $\lambda_{\min}$**:** Selects the penalty value achieving the lowest cross-validated MSE.
- $\lambda_{1\text{SE}}$ **(One-Standard-Error Rule):** Selects the *largest* (most heavily regularised and most parsimonious) λ whose CV error lies within one standard error of the minimum: $\text{CV}(\lambda) \leq \text{CV}(\lambda_{\min}) + \text{SE}(\lambda_{\min})$. This guards against selecting an overly complex model due to random fold variation.

Method	Penalty Form	Sparsity (Exact Zeros)	Handling Correlated Groups	Works When $p > n$
OLS	None	No	Extreme variance / sign flips	No (non-unique)
Best Subset	$\|\beta\|_0 \leq k$	Yes	Selects arbitrarily	Only up to $k < n$
Ridge	$\lambda\|\beta\|_2^2$	No	Shares weight evenly (grouping)	Yes
LASSO	$\lambda\|\beta\|_1$	Yes	Selects one, drops peers	Yes (selects $\leq n$)
Elastic Net	$\lambda[\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2]$	Yes	Groups correlated predictors together	Yes (can select $> n$)

Worked Example 2.3: Orthonormal Ridge vs. LASSO Soft-Thresholding

In an orthonormal factor model predicting quarterly Australian equity returns, OLS estimates on four standardised factors are $\hat{\beta}^{\text{OLS}} = (1.80, -0.95, 0.35, -0.15)^\top$. Compare Ridge ($\lambda = 0.5$ under $\frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$, giving $\hat{\beta}_j^{\text{ridge}} = \hat{\beta}_j^{\text{OLS}}/(1+\lambda)$) against LASSO soft-thresholding at $\lambda = 0.5$.

1. **Ridge Shrinkage ($\lambda = 0.5$):** Dividing each coordinate by $1 + 0.5 = 1.5$:

$$\hat{\beta}_1^{\text{ridge}} = \frac{1.80}{1.5} = 1.200, \quad \hat{\beta}_2^{\text{ridge}} = \frac{-0.95}{1.5} = -0.633, \quad \hat{\beta}_3^{\text{ridge}} = \frac{0.35}{1.5} = 0.233, \quad \hat{\beta}_4^{\text{ridge}} = \frac{-0.15}{1.5} = -0.100$$

   Every factor remains non-zero, shrinking proportionally by one-third.

2. **LASSO Soft-Thresholding ($\lambda = 0.5$):** Applying $\text{sign}(\hat{\beta}_j^{\text{OLS}})\max(|\hat{\beta}_j^{\text{OLS}}| - 0.5, 0)$:
   - $j = 1$: $+\max(1.80 - 0.50, 0) = +1.30$
   - $j = 2$: $-\max(0.95 - 0.50, 0) = -0.45$
   - $j = 3$: $+\max(0.35 - 0.50, 0) = 0.00$ (eliminated)
   - $j = 4$: $-\max(0.15 - 0.50, 0) = 0.00$ (eliminated)

   LASSO yields the sparse vector $\hat{\beta}^{\text{lasso}} = (1.30, -0.45, 0, 0)^\top$, automatically discarding the two weakest factors while preserving more magnitude on the strongest signal β_1 than proportional Ridge shrinkage.

DegreeLive

2.6 Moving Beyond Linearity: Basis Expansions and Splines

Real-world economic relationships—such as the age-earnings profile in HILDA labour data or the response of electricity demand to daily temperature—are rarely linear across the entire domain. Rather than abandoning the linear estimation toolkit, **basis expansion** replaces a raw predictor X with a set of fixed nonlinear transformations $b_1(X), b_2(X), \dots, b_M(X)$ and fits a linear model in the expanded feature space:

$$y_i = \beta_0 + \sum_{m=1}^{M} \beta_m b_m(x_i) + \varepsilon_i$$

Where $b_m(\cdot)$ are known basis functions chosen before fitting, allowing OLS or regularised linear regression to estimate $\beta_0, \dots, \beta_M$ effortlessly.

Global Polynomials vs. Piecewise Polynomials

Setting $b_m(X) = X^m$ produces global **polynomial regression** of degree d. However, high-degree global polynomials ($d \geq 4$) suffer from severe edge instability (Runge's phenomenon): an extreme observation in one tail forces the entire polynomial curve to oscillate wildly in distant regions. **Step functions** divide the range of X into bins $C_1, \dots, C_K$ via indicator functions $\mathrm{I}(X \in C_k)$, ensuring locality, but introduce artificial jump discontinuities at bin cutpoints.

Regression Splines and Truncated Power Basis

A **cubic regression spline** combines the flexibility of polynomials with strict locality by partitioning the domain of X at K distinct interior **knots** $\xi_1 < \xi_2 < \cdots < \xi_K$ and fitting separate cubic polynomials within each interval, constrained so that the curve, its first derivative (slope), and its second derivative (curvature) are all continuous across every knot (C^2 smoothness). Because third derivatives may jump at the knots, a cubic spline with K interior knots requires $K + 4$ basis functions (intercept plus $K + 3$ degrees of freedom) and can be represented cleanly using the **truncated power basis**:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{k=1}^{K} \theta_k (x_i - \xi_k)_+^3 + \varepsilon_i$$

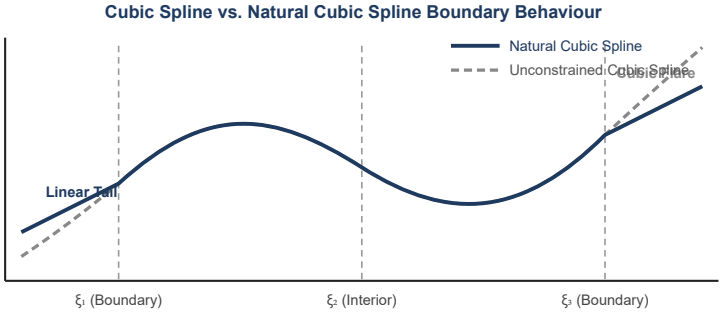
Where:

- $(x - \xi_k)_+ = \max(x - \xi_k, 0)$ is the positive-part operator equal to $x - \xi_k$ when $x > \xi_k$ and 0 when $x \leq \xi_k$.
- $\xi_1, \dots, \xi_K$ are the K interior knots (typically placed at sample quantiles of X, such as the 25th, 50th, and 75th percentiles).
- θ_k measures the change in the third derivative as X crosses knot ξ_k, while preserving continuity of $f, f', \text{and } f''$ everywhere.

Natural Cubic Splines

Even with knots, unconstrained cubic splines exhibit elevated variance near the outer boundaries $X < \xi_1$ and $X > \xi_K$, where data are sparse and cubic extrapolations diverge quickly. A **Natural Cubic Spline** imposes four additional boundary constraints: the function is forced to be strictly *linear* (zero second and third derivatives) beyond the two outermost boundary knots ξ_1 and ξ_K. Enforcing linearity in both boundary tails eliminates four parameters, leaving a natural cubic spline with K knots with exactly K total degrees of freedom (including the intercept), dramatically narrowing confidence intervals at the edges of the data.

DegreeLive



Worked Example 2.4: Evaluating a Truncated Power Cubic Spline

Suppose a fitted cubic regression spline with a single interior knot at $\xi_1 = 5$ models daily electricity peak load $\hat{y}$ (in GW) as a function of afternoon temperature deviation x:

$$\hat{f}(x) = 10 + 0.4x + 0.05x^2 - 0.002x^3 + 0.03(x-5)_+^3$$

Compute the predicted load and verify C^2 smoothness at $x = 3, x = 5$, and $x = 8$.

1. **At $x = 3$ (below knot $\xi_1 = 5$):** Since $3 - 5 = -2 < 0$, the truncated term $(3-5)_+^3 = 0$. Thus:

$$\hat{f}(3) = 10 + 0.4(3) + 0.05(9) - 0.002(27) + 0 = 10 + 1.20 + 0.45 - 0.054 = 11.596 \text{ GW}$$

2. **At $x = 5$ (exact knot):** Both left and right branches yield $(5-5)_+^3 = 0$:

$$\hat{f}(5) = 10 + 0.4(5) + 0.05(25) - 0.002(125) = 10 + 2.00 + 1.25 - 0.25 = 13.000 \text{ GW}$$

   Evaluating the derivatives of $0.03(x-5)_+^3$ gives first derivative $0.09(x-5)_+^2 = 0$ and second derivative $0.18(x-5)_+ = 0$ at $x = 5$, confirming continuity of $\hat{f}, \hat{f}'$, and $\hat{f}''$ across the knot.

3. **At $x = 8$ (above knot $\xi_1 = 5$):** Here $(8-5)_+^3 = 3^3 = 27$, activating the post-knot cubic acceleration:

$$\hat{f}(8) = 10 + 0.4(8) + 0.05(64) - 0.002(512) + 0.03(27) = 10 + 3.20 + 3.20 - 1.024 + 0.81 = 16.186 \text{ GW}$$

2.7 Smoothing Splines

Regression splines require the analyst to choose both the number and placement of knots. **Smoothing Splines** avoid manual knot selection by solving an infinite-dimensional variational optimisation problem that balances data fidelity against integrated curvature (roughness):

$$\min_{g \in C^2} \{ \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b [g''(t)]^2 \, dt \}$$

Where:

- $\sum_{i=1}^{n} (y_i - g(x_i))^2$ measures closeness to the observed training points.
- $\int_a^b [g''(t)]^2 dt$ measures total integrated squared second derivative (total curvature or wiggliness) across the domain $[a, b]$.

DegreeLive

- $\lambda \geq 0$ is the smoothing penalty hyperparameter. When $\lambda = 0$, g can interpolate every data point; as $\lambda \to \infty$, $g''(t) \to 0$ everywhere, forcing $\hat{g}(x)$ to become the global OLS straight line.

Remarkably, a classical theorem (Reinsch, 1967) proves that the unique minimiser of this infinite-dimensional objective is a **shrunk natural cubic spline** with knots placed at every single unique observed predictor value $x_1, \ldots, x_n$. Rather than estimating n unconstrained parameters, the penalty shrinks the spline coefficients toward linearity, yielding an effective degrees of freedom $\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda) \in [2, n]$, where $\hat{g} = S_\lambda y$ is the linear smoother matrix. Because smoothing splines are linear smoothers, the optimal λ (or target df_λ) can be chosen automatically in $O(n)$ time via the LOOCV leverage shortcut $\sum_{i=1}^{n}[(y_i - \hat{g}(x_i))/(1 - (S_\lambda)_{ii})]^2$.

2.8 Generalised Additive Models (GAMs) and Interpretability

Extending univariate splines to p predictors via unrestricted multivariate interactions rapidly runs into the curse of dimensionality. **Generalised Additive Models (GAMs)** (Hastie and Tibshirani, 1990) bridge rigid linear models and opaque black-box learners by modelling the expected response (or transformed mean via a link function $g(\cdot)$) as an additive sum of smooth, univariate nonlinear component functions:

$$g\left(\mathrm{E}[Y \mid X]\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Where:

- $g(\cdot)$ is a monotonic link function (the identity $g(\mu) = \mu$ for Gaussian regression, or the logit $g(\mu) = \ln(\mu/(1-\mu))$ for binary classification).
- β_0 is the global intercept, and each $f_j(X_j)$ is a smooth one-dimensional function (commonly a natural spline, penalised B-spline, or smoothing spline) constrained to zero sample mean $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ for identifiability.

The Backfitting Algorithm

When each component f_j is estimated via a non-parametric smoother such as a smoothing spline, GAMs are fitted using **backfitting**:

1. Initialise $\hat{\beta}_0 = \bar{y}$ and $\hat{f}_j(x_j) = 0$ for all $j = 1, \ldots, p$.
2. Cycle through $j = 1, 2, \ldots, p$, holding all other functions $\hat{f}_k$ $(k \neq j)$ fixed, and compute the partial residual vector:

$$r_i^{(j)} = y_i - \hat{\beta}_0 - \sum_{k \neq j} \hat{f}_k(x_{ik})$$

3. Fit a univariate smoother to $r^{(j)}$ against X_j to update $\hat{f}_j$, then centre $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_j(x_{ij})$.
4. Repeat steps 2–3 until all fitted functions change by less than a convergence tolerance.

XAI and Interpretability Advantages

Because each predictor enters additively, GAMs offer built-in **explainable artificial intelligence (XAI)** transparency: analysts can plot each partial effect curve $\hat{f}_j(x_j)$ with pointwise confidence bands while holding all other predictors constant, directly diagnosing nonlinear thresholds (for example, diminishing marginal returns to marketing spend or nonlinear credit risk surges above an 80% loan-to-value ratio). Moreover, individual predictors can combine seamlessly—linear terms for binary indicators, smooth splines for continuous covariates, and bivariate tensor-product surfaces $f_{jk}(X_j, X_k)$ when domain theory identifies critical two-way interactions—making GAMs a workhorse in regulated Australian banking, insurance pricing, and public policy analytics. Their principal limitation is the omission of unmodelled high-order interactions, which tree ensembles (Module 4) capture automatically.

DegreeLive

Coordinate Descent and Post-Selection Inference Caveats

In software such as R's `glmnet` package (and its `tidymodels` wrapper), LASSO and Elastic Net solutions are computed over an entire decreasing grid of λ values using **cyclic coordinate descent** combined with warm starts. Holding all coefficients β_k ($k \neq j$) fixed at their current estimates, minimising the penalised objective with respect to a single scalar β_j reduces to a closed-form soft-thresholding update on the partial residual $r_i^{(-j)} = y_i - \beta_0 - \sum_{k \neq j} \beta_k x_{ik}$. Repeating across $j = 1, \dots, p$ converges rapidly even when $p > 10{,}000$.

Analysts must exercise extreme caution regarding **post-selection inference**: fitting LASSO to select a sparse subset of predictors and then running standard OLS to obtain classical t-statistics and p-values on the surviving variables produces severely anti-conservative (artificially small) p-values. Because OLS ignores the data-driven search step that screened out weaker predictors, only predictors large enough to survive selection are tested, biasing standard errors downward. Valid inference after LASSO requires sample splitting (selecting variables on half the data and running OLS inference strictly on the independent second half), bootstrap debiasing, or selective inference adjustments.

2.9 Comparing Nonlinear Smoothers and Additive Workflows

Selecting among global basis expansions, local splines, and additive models depends on domain smoothness, sample size near boundaries, and interpretability requirements:

Nonlinear Method	Basis / Penalty Mechanism	Boundary Behaviour	Tuning Requirement	Primary Strength
Global Polynomial	$1, X, X^2, \dots, X^d$	Unstable oscillations for $d \geq 4$	Degree d via CV / AIC	Simple parametric equation
Step Function	Piecewise constant bin indicators	Flat constant within outer bins	Number & location of cutpoints	Natural for discrete policy brackets
Cubic Regression Spline	Truncated power or B-spline basis ($K + 4$ df)	High variance in outer tails	Knot count K & placement	Smooth C^2 local flexibility
Natural Cubic Spline	Cubic spline + linear tail constraints (K df)	Stable linear extrapolation	Degrees of freedom K	Reliable tail confidence bands
Smoothing Spline	Natural spline at all x_i + $\lambda \int [g''(t)]^2 dt$	Linear outside data range	Penalty λ via LOOCV / GCV	Avoids manual knot selection
GAM (e.g., `mgcv`)	Sum of smooth univariate terms + link $g(\cdot)$	Linear or penalised tail smooths	Smoothing penalties via REML / GCV	High XAI transparency across p features

Worked Example 2.5: Additive GAM Prediction for Victorian Electricity Demand

An energy analyst models daily peak Victorian electricity demand Y (in gigawatts, GW) using a Gaussian GAM combining a linear weekend indicator $I_{\text{wknd}} \in \{0, 1\}$ with two centred smooth natural spline terms for daily maximum temperature T (in $^\circ\text{C}$) and log wholesale gas price G:

$$\hat{\text{E}}[Y \mid T, G, I_{\text{wknd}}] = 7.20 + \hat{f}_1(T) + \hat{f}_2(G) - 1.15\, I_{\text{wknd}}$$

Where $\hat{f}_1(T)$ exhibits a U-shaped thermal profile (elevated winter heating load below 14°C and steep summer air-conditioning load above 28°C): specifically, $\hat{f}_1(12) = +0.85$, $\hat{f}_1(22) = -0.60$, and $\hat{f}_1(38) = +2.40$, while $\hat{f}_2(2.0) = -0.25$ and $\hat{f}_2(3.0) = -0.45$. Compute the predicted peak demand and marginal regime shifts:

1. **Mild Weekday ($T = 22^\circ\text{C}, G = 2.0, I_{\text{wknd}} = 0$):**

$$\hat{Y} = 7.20 + (-0.60) + (-0.25) - 1.15(0) = 6.35 \text{ GW}$$

2. **Heatwave Weekday ($T = 38°\text{C}, G = 2.0, I_{\text{wknd}} = 0$):**

$$\hat{Y} = 7.20 + 2.40 + (-0.25) - 0 = 9.35 \text{ GW}$$

Because of additivity, the $16°\text{C}$ temperature jump adds exactly $\hat{f}_1(38) - \hat{f}_1(22) = 2.40 - (-0.60) = +3.00$ GW regardless of gas price or day type.

3. **Heatwave Weekend with High Gas Price ($T = 38°\text{C}, G = 3.0, I_{\text{wknd}} = 1$):**

$$\hat{Y} = 7.20 + 2.40 + (-0.45) - 1.15(1) = 8.00 \text{ GW}$$

Notice how each term translates directly into an auditable operational breakdown for NEM grid dispatch planners.

DegreeLive