Penalized Likelihood and Regularization

Statistics · Module 10 — Model Selection Learning

Node 10.4

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.4.

Penalized likelihood is the analytical implementation of the bias–variance tradeoff (Node 10.1): instead of selecting a model from a discrete set, it continuously shrinks parameter estimates toward a structured target. Ridge regression (L^2), LASSO (L^1), and the elastic net ($L^1 + L^2$) are the canonical instances. The geometry of the penalty ball determines whether the estimator achieves sparsity. This node provides the foundation for high-dimensional methods (Node 10.5) and informs model averaging through regularization paths (Node 10.6).

Formal Definition

Let $\ell(\beta; X)$ be the log-likelihood for $\beta \in \mathbb{R}^p$. The **penalized likelihood estimator** is

$$\hat{\beta}^\lambda = \arg\max_{\beta \in \mathbb{R}^p} \left[\ell(\beta; X) - \lambda\, \Omega(\beta) \right],$$

where $\Omega(\beta) \geq 0$ is a penalty function and $\lambda \geq 0$ controls the penalty strength. Equivalently, for the Gaussian linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$:

$$\hat{\beta}^\lambda = \arg\min_{\beta} \left[\frac{1}{2n} \|Y - X\beta\|^2 + \lambda\, \Omega(\beta) \right].$$

Intuition

When p is large relative to n, the MLE overfits: it finds coefficients that explain the noise in the training data perfectly but predict poorly on new data. Regularization constrains the estimator by penalizing large or numerous coefficients. The geometry matters: the L^2 ball is smooth and round, shrinking all coefficients proportionally; the L^1 ball has corners at the axes, pushing small coefficients to exactly zero. This geometric difference is why LASSO performs variable selection and ridge does not.

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1), maximum likelihood theory (Node 3.2), and the Bayesian perspective (Node 6.1). **Enables:** High-dimensional theory (Node 10.5) uses LASSO as its central estimator; model averaging via stacking (Node 10.6) uses regularization to combine models.

Penalized likelihood unifies frequentist and Bayesian perspectives: the penalty is a prior (Node 6.1), and the penalized MLE is the posterior mode. The tuning parameter λ controls the bias–variance tradeoff continuously rather than discretely.

Mathematical Development

Ridge Regression (L^2 Penalty)

The **ridge estimator** is

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2.$$

Setting the gradient to zero: $-X^\top(Y - X\beta) + \lambda\beta = 0$, so

$$\hat{\beta}^{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Properties via the SVD. Let $X = UDV^\top$ with singular values $d_1 \geq \cdots \geq d_p$. Then

$$\hat{\beta}^{\text{ridge}} = V \operatorname{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) V^\top \hat{\beta}^{\text{OLS}}.$$

Each coefficient in the rotated basis is shrunk by the factor $d_j^2/(d_j^2 + \lambda) \in [0, 1)$. Components along small singular values (where the data provides little information) are shrunk the most.

Bias–variance decomposition for ridge:

$$\text{Bias}^2 = \lambda^2 \beta_0^\top (X^\top X + \lambda I)^{-2} \beta_0, \quad \text{Variance} = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2}.$$

As $\lambda \to 0$, bias vanishes and variance equals the OLS variance. As $\lambda \to \infty$, variance vanishes and bias approaches $\|\beta_0\|^2$.

Bayesian interpretation: Ridge is the posterior mean under $\beta \sim \mathcal{N}(0, (\sigma^2/\lambda)I)$.

LASSO (L^1 Penalty)

The **LASSO** estimator is

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

The L^1 penalty is not differentiable at $\beta_j = 0$. The **subdifferential** (KKT conditions) for coordinate j is:

$$-X_j^\top(Y - X\hat{\beta}) + \lambda s_j = 0, \quad s_j \in \begin{cases} \{\text{sign}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \hat{\beta}_j = 0. \end{cases}$$

This means $\hat{\beta}_j = 0$ whenever $|X_j^\top(Y - X\hat{\beta}_{-j})| \leq \lambda$: the penalty drives the coefficient to zero if the marginal correlation is too small to justify including the variable.

Soft-thresholding operator. For orthogonal X (i.e., $X^\top X = nI$), the LASSO solution decouples coordinate-wise:

$$\hat{\beta}_j^{\text{LASSO}} = S_{\lambda/n}(\hat{\beta}_j^{\text{OLS}}) = \text{sign}(\hat{\beta}_j^{\text{OLS}})\big(|\hat{\beta}_j^{\text{OLS}}| - \lambda/n\big)_+.$$

Derivation of soft-thresholding. For orthogonal X, the objective separates: $\min_{\beta_j} \frac{n}{2}(\hat{\beta}_j^{\text{OLS}} - \beta_j)^2 + \lambda|\beta_j|$. For $\beta_j > 0$: differentiate to get $-n(\hat{\beta}_j^{\text{OLS}} - \beta_j) + \lambda = 0$, so $\beta_j = \hat{\beta}_j^{\text{OLS}} - \lambda/n$, valid when $\hat{\beta}_j^{\text{OLS}} > \lambda/n$. By symmetry and the KKT conditions for $\beta_j = 0$, the full solution is the soft-thresholding operator.

Geometry of sparsity. The constraint set $\{\beta : \|\beta\|_1 \leq t\}$ is a cross-polytope (diamond in 2D). Its corners lie on the coordinate axes. The contours of the quadratic loss are ellipsoids. The point of tangency between an ellipsoid and a cross-polytope generically lies at a corner, setting one or more coordinates to zero. By contrast, the L^2 ball $\{\beta : \|\beta\|_2 \leq t\}$ is smooth, so the tangency generically occurs at an interior point with all coordinates nonzero.

Bayesian interpretation: LASSO is the posterior mode under independent Laplace priors $\beta_j \sim \text{Laplace}(0, \sigma^2/\lambda)$.

Elastic Net

The **elastic net** combines L^1 and L^2 penalties:

$$\hat{\beta}^{\text{EN}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|^2.$$

The L^2 term makes the objective strictly convex, ensuring a unique solution even when $p > n$. The **grouping effect**: for highly correlated predictors $X_j \approx X_k$, the elastic net assigns similar coefficients, whereas LASSO may select one and zero the other arbitrarily.

Equivalently, the elastic net can be written as a LASSO problem on an augmented dataset:

$$\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \quad \tilde{X} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2}\, I \end{pmatrix},$$

and solving the LASSO with penalty $\lambda_1/\sqrt{1+\lambda_2}$.

Selection of λ

The penalty parameter is not estimated from the likelihood. Standard approaches:

1. **Cross-validation** (Node 10.3): select λ minimizing K-fold CV error. For LASSO, compute the full regularization path via LARS (least angle regression) at cost comparable to a single OLS fit.
2. **Information criteria**: for LASSO, the effective degrees of freedom equals $|\hat{S}|$ (the number of nonzero coefficients), so AIC and BIC can be applied with $k = |\hat{S}|$.
3. **Theoretical bounds** (Node 10.5): set $\lambda = C\sigma\sqrt{\log p/n}$ for oracle-rate guarantees.

Worked Examples

Example 1: Ridge vs. OLS with Collinear Predictors

Data: $n = 20$, $p = 2$, X_1 and X_2 have correlation 0.99. True model: $Y = 3X_1 + 3X_2 + \varepsilon$, $\sigma = 1$.

The OLS estimates have high variance because $X^\top X$ is nearly singular. Suppose $X^\top X = 20\begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$.

Eigenvalues: $d_1^2 = 20(1.99) = 39.8$, $d_2^2 = 20(0.01) = 0.2$.

OLS variance: $\sigma^2\text{Tr}((X^\top X)^{-1}) = 1 \cdot (1/39.8 + 1/0.2) = 0.025 + 5.0 = 5.025$. The small eigenvalue inflates variance enormously.

Ridge with $\lambda = 1$: variance $= \sum_j d_j^2/(d_j^2+1)^2 = 39.8/40.8^2 + 0.2/1.2^2 = 0.024 + 0.139 = 0.163$. Bias^2 is small: the component along the large eigenvalue is barely shrunk, and the component along the small eigenvalue is shrunk significantly but the true signal along that direction is small. Ridge MSE $\approx 0.163 + \text{bias}^2 \ll 5.025$.

Example 2: LASSO Variable Selection

Data: $n = 100$, $p = 10$. True model: $Y = 2X_1 - X_3 + 0.5X_7 + \varepsilon$, $\sigma = 1$. Predictors are independent standard normal.

At $\lambda = 0.5$ (chosen by CV), the LASSO estimates: $\hat{\beta}_1 = 1.85$, $\hat{\beta}_3 = -0.82$, $\hat{\beta}_7 = 0.31$, and all others exactly zero. The LASSO correctly identifies the support $S = \{1, 3, 7\}$ and provides shrunken but useful estimates of the coefficients.

At $\lambda = 2.0$ (too large): only $\hat{\beta}_1 = 0.94$ survives; X_3 and X_7 are incorrectly zeroed. At $\lambda = 0.01$ (too small): all 10 predictors have nonzero coefficients, overfitting.

The CV error curve is U-shaped in λ, with minimum near $\lambda = 0.5$, confirming the bias–variance tradeoff.

Cross-Links

AI. Weight decay in neural networks is L^2 regularization applied to network weights. Dropout induces an implicit regularization effect similar to an adaptive L^2 penalty. Sparse attention mechanisms and network pruning exploit the L^1/sparsity principle.

Economics. Penalized regression underlies high-dimensional empirical applications including factor model selection, treatment effect estimation with many controls, and forecasting with large predictor sets. The post-LASSO estimator (OLS on the LASSO-selected variables) is standard in empirical work to remove shrinkage bias.

Linear Algebra. The geometry of L^p balls for $p \leq 1$ (non-convex), $p = 1$ (convex, corners), and $p = 2$ (smooth, round) determines the sparsity properties of the solution. The LASSO is a linear program (for fixed λ) and can be solved by coordinate descent or the LARS algorithm, both exploiting the polyhedral geometry.

Quantum Mechanics. Compressed sensing (the signal-processing analogue of LASSO) is used in quantum state tomography to reconstruct low-rank density matrices from a small number of measurements. The sparsity-inducing geometry of the nuclear norm (trace norm) plays the role of the L^1 norm for matrix-valued signals.

Structural Compression

Invariant

Penalized likelihood reduces variance at the cost of bias; the optimal penalty λ^* minimizes total prediction error. The L^1 penalty achieves sparsity (exact zeros); the L^2 penalty achieves shrinkage (approximate zeros).

Mechanism

The penalty $\lambda\Omega(\beta)$ restricts the feasible region. For L^1: the KKT conditions show that the subdifferential at zero absorbs the gradient when the signal is weak, setting the coefficient to exactly zero. For L^2: the smooth penalty applies uniform multiplicative shrinkage via $d_j^2/(d_j^2 + \lambda)$. The soft-thresholding operator is the proximal operator of the L^1 norm.

Cross-System Echo

LASSO is equivalent to basis pursuit (compressed sensing) in signal processing. The sparsity-inducing geometry of the L^1 ball is central to both. In AI, weight decay (ridge) and L^1 penalties are foundational

regularization techniques. In economics, penalized regression enables credible estimation in settings where the number of potential controls exceeds the sample size.

Exercises

1. Derive the ridge estimator $\hat{\beta}^{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y$ by setting the gradient of the penalized objective to zero. Express the solution in terms of the SVD of X.
2. Derive the soft-thresholding operator for the LASSO in the orthogonal design case. Draw the solution path $\hat{\beta}_j(\lambda)$ as a function of λ for $\hat{\beta}_j^{\text{OLS}} = 2.5$.
3. Write out the KKT conditions for the LASSO in the general (non-orthogonal) case. Show that if $|X_j^\top (Y - X\hat{\beta}_{-j})| < \lambda$, then $\hat{\beta}_j = 0$.
4. Prove the grouping property of the elastic net: for $X_j = X_k$, show that $\hat{\beta}_j^{\text{EN}} = \hat{\beta}_k^{\text{EN}}$, whereas the LASSO may assign different values.
5. For ridge regression, show that the effective degrees of freedom is $\text{df}(\lambda) = \sum_{j=1}^{p} d_j^2/(d_j^2 + \lambda)$, where d_j are the singular values of X.
6. Consider $n = 50$, $p = 5$, $X^\top X = 50I$, $\hat{\beta}^{\text{OLS}} = (3.0, 0.2, -2.5, 0.1, 1.8)$, $\sigma^2 = 1$. Compute the LASSO solution for $\lambda = 5$ using the soft-thresholding formula, and compute the ridge solution for $\lambda = 5$. Compare sparsity.
7. Argue, structurally, why the L^1 penalty produces exact zeros while the L^2 penalty does not, using both the geometric (constraint set tangency) and analytic (subdifferential) perspectives.

Statistics · Module 10 — Model Selection Learning · Node 10.4
Concepts: regularization, convexity, maximum-likelihood-estimation, bias-variance-tradeoff
Prerequisites: 10.1, 3.2, 6.1
Enables: 10.5, 10.6

hbar.university — own.your.cognition