Machine Learning

Choice of Linear Model and Regularization principles

Thomas Milcent - 2025

Regularization key principles

Reduce overfitting

- Highly complex models can fit the training data too well, capturing noise instead of underlying patterns.
- Regularization **adds a penalty** to **limit model complexity**, helping it generalize better to new data.

Balance bias and variance

- Regularization helps **reduce variance** (sensitivity to data fluctuations) at the cost of a **slight increase in bias**.
- The goal is a better bias-variance tradeoff, so the model performs well on unseen data.

Add a constraint on model parameters via a penalty (L1 or L2)

- L1 (Lasso): encourages sparsity in coefficients (some become exactly 0).
- L2 (Ridge): prevents coefficients from becoming too large.

Regularization key principles

Minimize

Prediction error + λ penalty on model complexity

- **Biased estimator** when $\lambda \neq 0$.
- Trade bias for a smaller variance.
- λ can be set by cross-validation.

- Simpler model ≈ fewer parameters

→ **shrinkage**: drive the coefficients of the parameters towards 0.

Ridge Regression – L2 Penalty

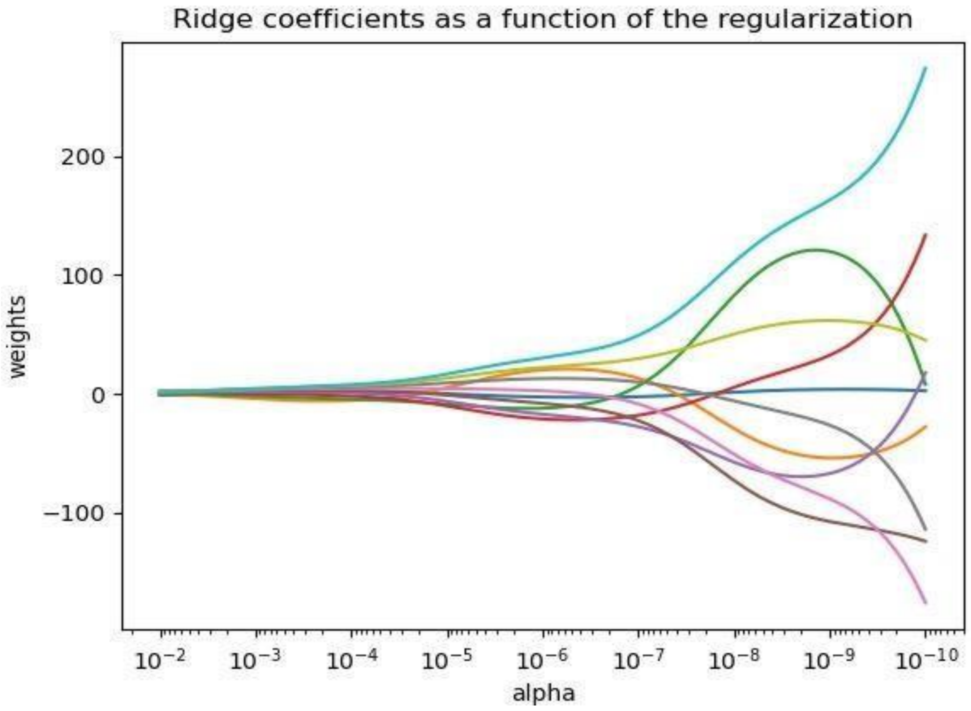
Sum-of-squares penalty $$\hat{\beta}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2$$

Ridge regression estimator $$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top \boldsymbol{y}$$

$$\text{if } (X^\top X + \lambda I) \text{ invertible}$$

Solution Path

[Plot Ridge coefficients as a function of the regularization — scikit-learn 1.3.2 documentation]



Decreasing value of lambda (“alpha” in this example)

Ridge Regression – L2 Penalty

Grouped selection:

- correlated variables get similar weights
- identical variables get identical weights

Ridge regression shrinks coefficients towards 0 but does not result in a **sparse model**.

Sparsity:

- many coefficients get a weight of 0
- they can be eliminated from the model.

Lasso regression – L1 Penalty

L1 penalty

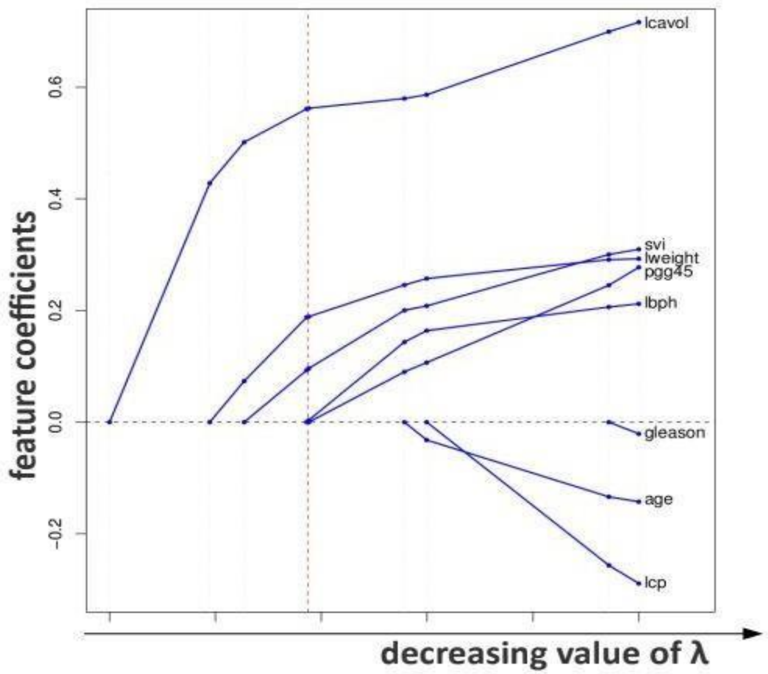
$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_{\boxed{1}}$$

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_p|$$

Tends to give sparse coefficients

No explicit solution: quadratic form under linear constraints to solve this

Solution path



Elastic Net – Mix of L1 and L2 penalty

Combine Lasso & Ridge regression

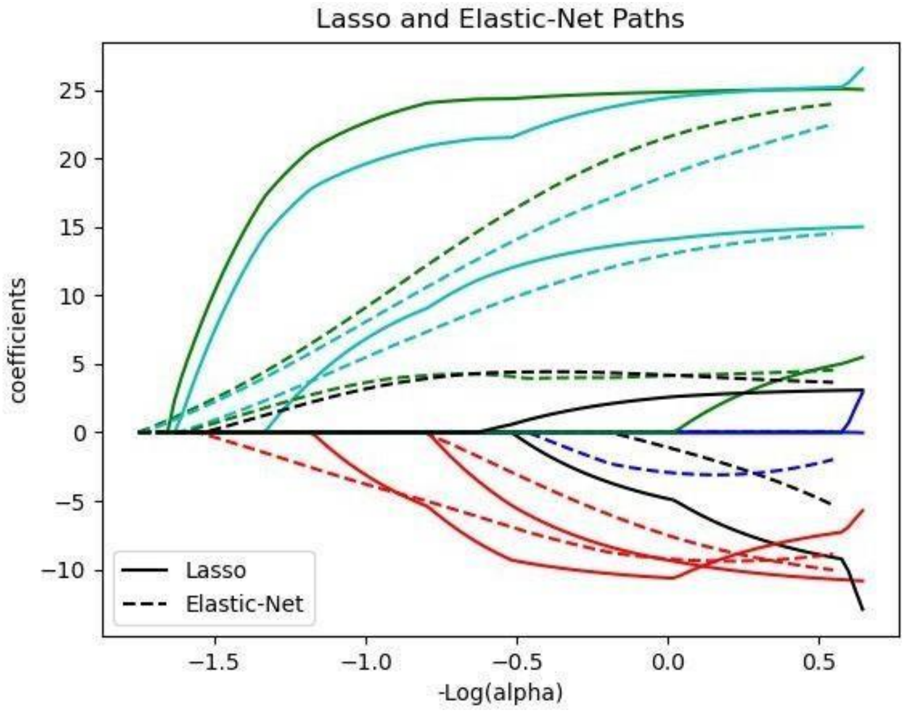
$$\hat{\beta}_{\text{enet}} = \arg\min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2^2 + \lambda \left(\alpha||\boldsymbol{\beta}||_2^2 + (1-\alpha)||\boldsymbol{\beta}||_1\right)$$

The best of the two approach
Selects variables like the Lasso regression
Shrinks together coefficients of correlated variables like the Ridge regression

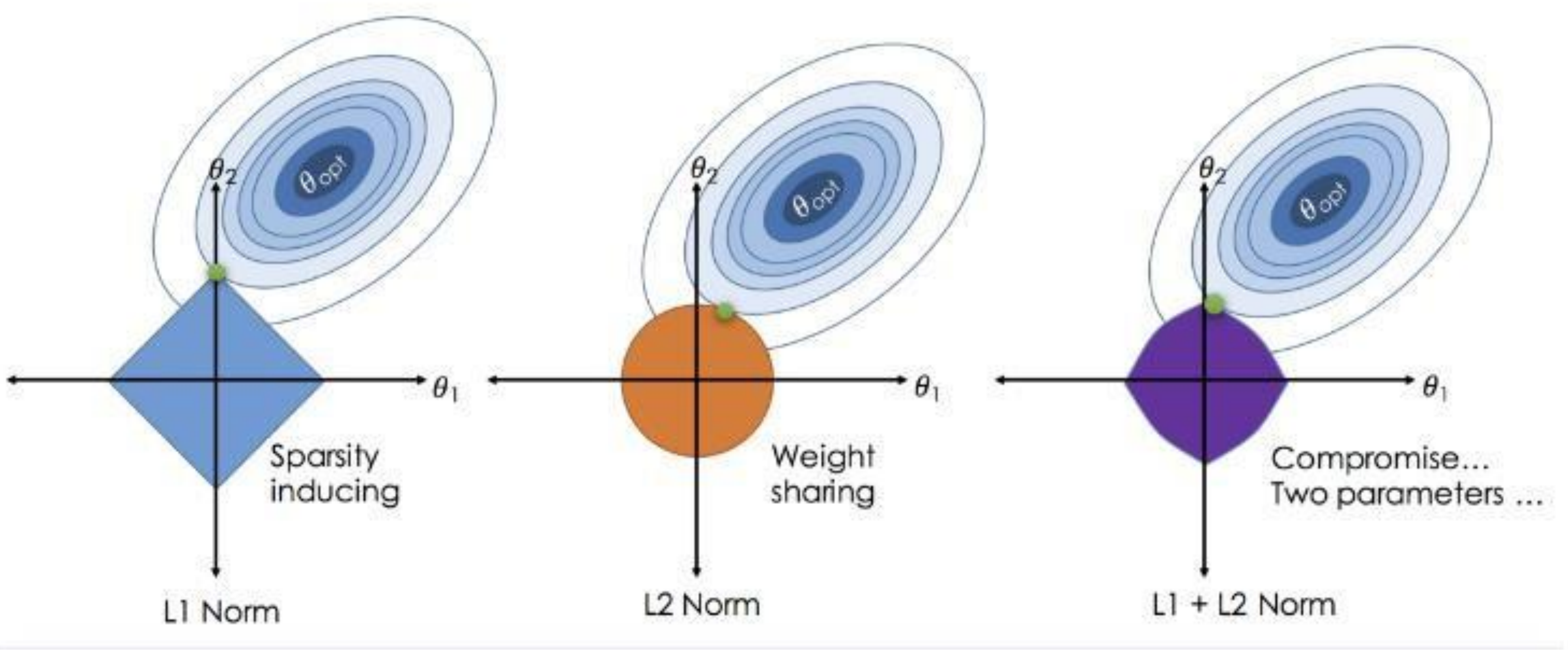
Elastic Net – Mix of L1 and L2 penalty

Example

[Lasso and Elastic Net — scikit-learn 1.3.2 documentation]



To Sum Up : Lasso vs Ridge vs Elastic Net



Play with Lasso and Ridge here : https://www.interactive-ml.com/regularization.html?

To Sum Up : Lasso vs Ridge vs Elastic Net

	Ridge (L2)	**Lasso (L1)**	**Elastic Net (L1 + L2)**
Penalty type	L2: sum of squared coefficients	L1: sum of absolute coefficients	Combination of L1 and L2 penalties
Effect on coefficients	Shrinks coefficients but **never to zero**	Can shrink coefficients **exactly to zero**	Shrinks coefficients and can set some to zero (depending on the mix)
Behavior with correlated predictors	Shares weight across correlated variables (“grouping effect”)	Tends to pick **one** and drop the others	Encourages groups of correlated variables, but can still drop some
When it works best	Many small/medium effects; multicollinearity	Only a few features matter; interpretability needed	Many predictors, possibly correlated; feature selection + stability
Stability of solution	Very stable	Less stable when predictors are correlated	More stable than Lasso, more flexible than Ridge
Bias-variance behavior	More bias than OLS, much lower variance	More bias, variance reduction + sparsity	Tunable trade-off between L1 and L2 effects
Typical hyperparameters	λ (regularization strength)	λ (strength)	λ (strength) + α (mix between L1 and L2)

Generalized Linear Regression – Recap

Understand regularization as a means to control model complexity

Define Lasso, ridge regression, elastic net

Understand the role of the l1 and l2 norms in regularization

Interpret solution paths for Lasso and ridge regression