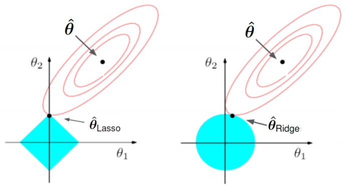
Introduction to Machine Learning

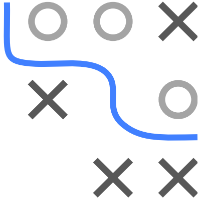
Regularization
Lasso vs. Ridge



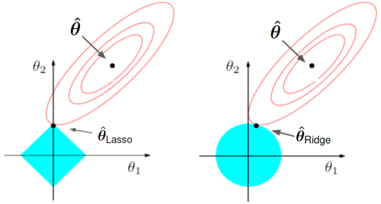
Learning goals

- Properties of ridge vs. lasso
- Coefficient paths
- What happens with corr. features
- Why we need feature scaling

LASSO VS. RIDGE GEOMETRY



$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 \qquad \text{s.t. } \|\boldsymbol{\theta}\|_p^p \leq t$$

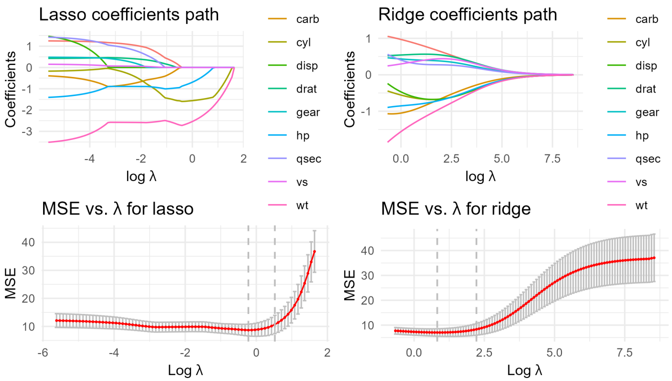


- In both cases (and for sufficiently large λ), the solution which minimizes $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is always a point on the boundary of the feasible region.
- As expected, $\hat{\boldsymbol{\theta}}_{\text{lasso}}$ and $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ have smaller parameter norms than $\hat{\theta}$.
- For lasso, solution likely touches a vertex of constraint region. Induces sparsity and is a form of variable selection.
- For $p > n$: lasso selects at most n features ▸ Zou and Hastie 2005.

COEFFICIENT PATHS AND 0-SHRINKAGE

Example 1: Motor Trend Car Roads Test (mtcars)

We see how only lasso shrinks to exactly 0.



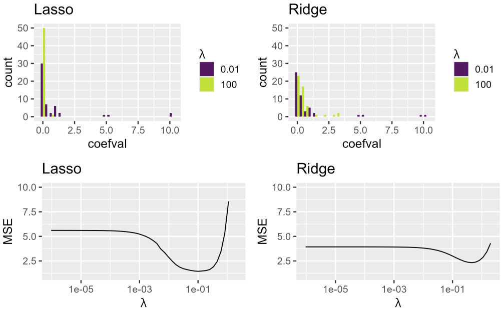
NB: No real overfitting here, as data is so low-dim.

COEFFICIENT PATHS AND 0-SHRINKAGE / 2

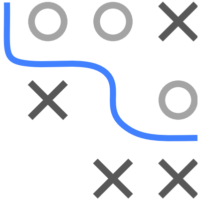
Example 2: High-dim., corr. simulated data: $p = 50$; $n = 100$

$$y = 10 \cdot (x_1 + x_2) + 5 \cdot (x_3 + x_4) + 1 \cdot \sum_{j=5}^{14} x_j + \epsilon$$

36/50 vars are noise; $\epsilon \sim \mathcal{N}(0, 1)$; $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$; $\Sigma_{k,l} = 0.7^{|k-l|}$



REGULARIZATION AND FEATURE SCALING



- Typically we omit θ_0 in penalty $J(\boldsymbol{\theta})$ so that the “infinitely” regularized model is the constant model (but can be implementation-dependent).
- Unregularized LM has **rescaling equivariance**, if you scale some features, can simply "anti-scale" coefs and risk does not change.
- Not true for Reg-LM: if you down-scale features, coeffs become larger to counteract. They are then penalized stronger in $J(\boldsymbol{\theta})$, making them less attractive without any relevenat reason.
- **So: usually standardize features in regularized models, whether linear or non-linear!**

REGULARIZATION AND FEATURE SCALING / 2

- Let the DGP be $y = \sum_{j=1}^{5} \theta_j x_j + \varepsilon$ for $\boldsymbol{\theta} = (1, 2, 3, 4, 5)^\top$, $\varepsilon \sim \mathcal{N}(0, 1)$
- Suppose x_5 was measured in *m* but we change the unit to *cm* ($\tilde{x}_5 = 100 \cdot x_5$):

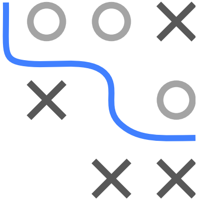
Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	MSE
OLS	0.984	2.147	3.006	3.918	**5.205**	0.812
OLS Rescaled	0.984	2.147	3.006	3.918	**0.052**	0.812

- Estimate $\hat{\theta}_5$ gets scaled by $1/100$ while other estimates and MSE are invariant
- Running ridge regression with $\lambda = 10$ on same data shows that rescaling of of x_5 does not result in inverse rescaling of $\hat{\theta}_5$ (everything changes!)
- This is because $\hat{\theta}_5$ now lives on small scale while $L2$ constraint stays the same. Hence remaining estimates can "afford" larger magnitudes.

Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	MSE
Ridge	0.709	1.874	2.661	3.558	**4.636**	1.366
Ridge Rescaled	0.802	1.943	2.675	3.569	**0.051**	1.08

- For lasso, especially for very correlated features, we could arbitrarily force a feature out of the model through a unit change.

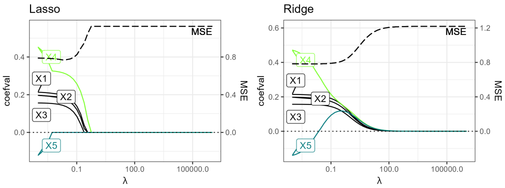
CORRELATED FEATURES: $L1$ VS $L2$



Simulation with $n = 100$:

$$y = 0.2x_1 + 0.2x_2 + 0.2x_3 + 0.2x_4 + 0.2x_5 + \epsilon$$

x_1-x_4 are independent, but x_4 and x_5 are strongly correlated.



- L1 removes x_5 early, L2 has similar coeffs for x_4, x_5 for larger λ
- Also called “grouping property”: for ridge highly corr. features tend to have equal effects; lasso however “decides“ what to select
- L1 selection is somewhat “arbitrary”

CORRELATED FEATURES: $L1$ VS $L2$ / 2

More detailed answer: The “random” decision is in fact a complex deterministic interaction of data geometry (e.g., corr. structures), the optimization method, and its hyperparamters (e.g., initialization). The theoretical reason for this behavior relates to the convexity of the penalties ▸ Zou and Hastie 2005.

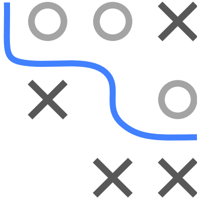
Considering perfectly collinear features $x_4 = x_5$ in the last example, we can obtain some more formal intuition for this phenomenon:

- Because $L2$ penalty is *strictly* convex:
  $x_4 = x_5 \implies \hat{\theta}_{4,ridge} = \hat{\theta}_{5,ridge}$ (grouping prop.)
- $L1$ penalty is not *strictly* convex. Hence, no unique solution exists if $x_4 = x_5$, and sum of coefficients can be arbitrarily allocated to both features while remaining minimizers (no grouping property!): For any solution $\hat{\theta}_{4,lasso}, \hat{\theta}_{5,lasso}$, equivalent minimizers are given by

$$\tilde{\theta}_{4,lasso} = s \cdot (\hat{\theta}_{4,lasso} + \hat{\theta}_{5,lasso}) \text{ and } \tilde{\theta}_{5,lasso} = (1-s) \cdot (\hat{\theta}_{4,lasso} + \hat{\theta}_{5,lasso}) \, \forall s \in [0,1]$$

SUMMARY

▸ Tibshirani 1996 ▸ Zou and Hastie 2005



- Neither ridge nor lasso can be classified as better overall
- Lasso can shrink some coeffs to zero, so selects features;
  ridge usually leads to dense solutions, with smaller coeffs
- Lasso likely better if true underlying structure is sparse
  ridge works well if there are many (weakly) influential features
- Lasso has difficulties handling correlated predictors;
  for high correlation, ridge dominates lasso in performance
- Lasso: for (highly) correlated predictors, usually an “arbitrary” one
  is selected, with large coeff, while the others are (nearly) zeroed
- Ridge: coeffs of correlated features are similar