Regularization

Contents map

developed content units	taxonomy levels
regularization	u1, e1
regularization path	u1, e1
ridge regression	u1, e1
Lasso	u1, e1

prerequisite content units	taxonomy levels
bias variance tradeoff	u1, e1

Main ILO of sub-module "Regularization"

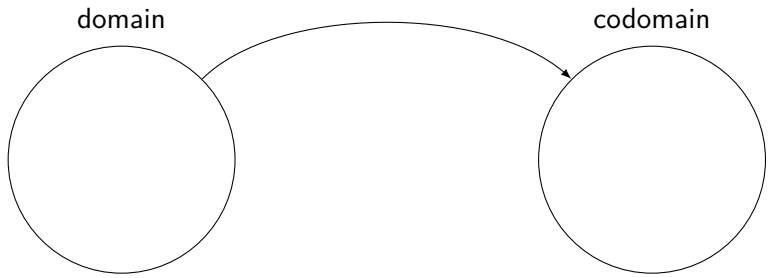
Compare different regularization techniques (ridge, lasso, and elastic net) by evaluating their mathematical formulations, graphical interpretations, and practical implications

Interpret regularization paths from Lasso regression plots to identify the relative importance of features in predictive modeling

Regularization = trading off variance with some bias

Regularization = trading off variance with some bias

main intuition: if $\widehat{\theta}$ has a variance V, then $0.9\widehat{\theta}$ has a variance $0.81V$

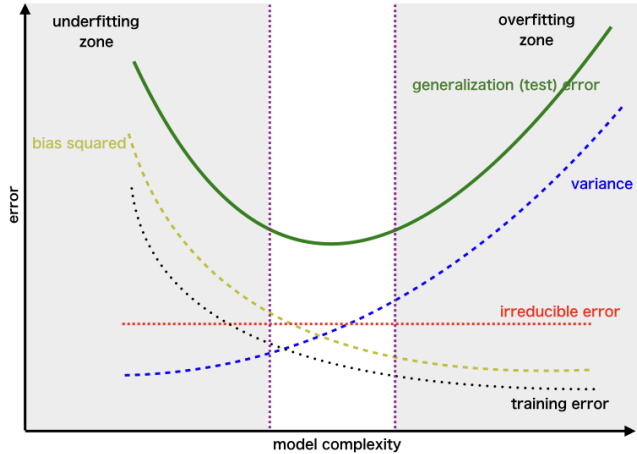


Regularization = trading off variance with some bias

main intuition: if $\widehat{\theta}$ has a variance V, then $0.9\widehat{\theta}$ has a variance $0.81V$ at the same time this will **likely** increase the bias:

Regularization = trading off variance with some bias

if $\widehat{\theta}$ has a variance V and bias B, then there will be a specific $\gamma\widehat{\theta}$ that minimizes $V + B^2$ (but we can't know a priori which γ is best):



the Stein's effect

An interesting example: the Stein's effect (in words)

(caveat: the next 3 slides are just motivational, not for the exam)

> *when estimating several parameters simultaneously, it's possible to improve overall estimation accuracy by borrowing strength across parameters, even if individual estimators may appear less accurate when considered in isolation*

An example of the Stein's effect in formulas

Given

$$y_t = \theta_t + e_t \qquad e_t \sim \mathcal{N}\left(0, \sigma^2\right) \text{ i.i.d.} \qquad \theta_t \in \mathbb{R} \qquad \boldsymbol{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \boldsymbol{\theta} := \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}$$

Then

$$\boldsymbol{\theta}_{LS} = \boldsymbol{y} \quad \text{does not minimize} \quad \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2\right]$$

and

$$\boldsymbol{\theta}_{JS} := \left(1 - \frac{N-2}{\|\boldsymbol{y}\|_2^2}\sigma^2\right)\boldsymbol{y}$$

has lower MSE than the LS solution

What is happening?

In this case

$$\boldsymbol{\theta}_{JS} = \left(1 - \frac{N-2}{\|\boldsymbol{y}\|_2^2}\sigma^2\right)\boldsymbol{y}$$

is a "regularized" version of

$$\boldsymbol{\theta}_{LS} = \boldsymbol{y}$$

ridge regularization

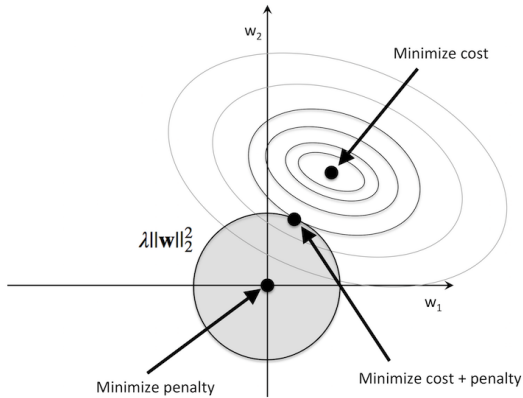
One of the most used regularization techniques: L_2 (a.k.a. "ridge")

$$J(\theta) = J_{\text{original}}(\theta) + \gamma \|\theta\|_2^2$$

Animation: `https://www.geogebra.org/m/myfghjzg`

Ridge regression = the most common approach to regularization

$$J(\theta) = \sum_{i=1}^{m} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$



Lasso regularization

The second most used regularization technique: L_1 (a.k.a. "lasso")

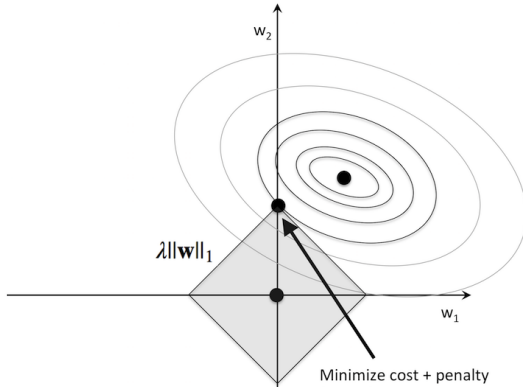
(actually this one typically works better than ridge!)

$$J(\theta) = J_{\text{original}}(\theta) + \gamma \|\theta\|_1$$

Animation: `https://www.geogebra.org/m/gaujemka`

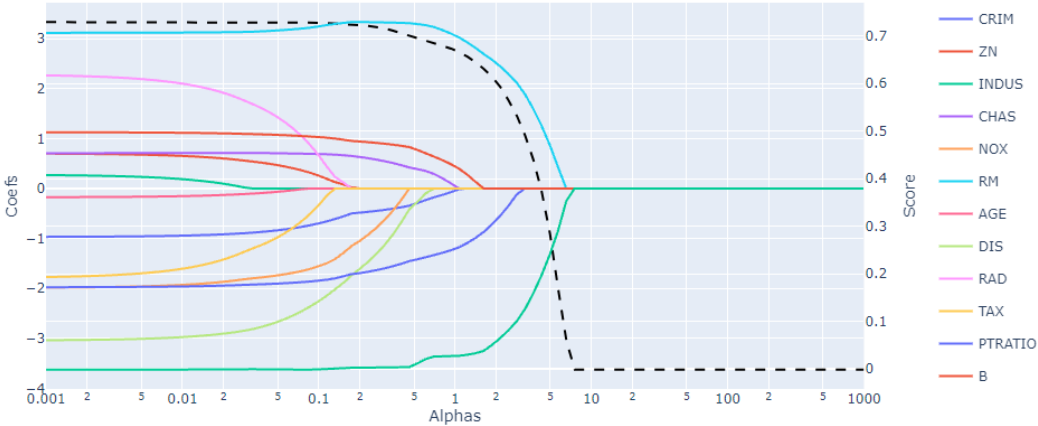
Lasso regression = the most common approach to regularization when one wants to promote sparsity (i.e., parsimonious models)

$$J(\theta) = \sum_{i=1}^{m} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$



A plot you should always include in your reports: the L_1 regularization path

(this is an implicit way to understand the relative importance of the features)



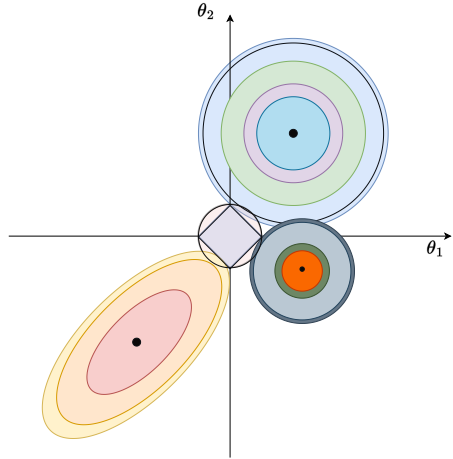
extensions

Elastic net = ridge + lasso

Last most-common regularization approach: combine the two into $\lambda_1 \|\boldsymbol{\theta}\|_1^2 + \lambda_2 \|\boldsymbol{\theta}\|_2^2$

A common way to represent regularization graphically

(Damiano's opinion: not as good as the 3D ones in geogebra)



Bayesian interpretation

Regularization may be seen with the Bayesian googles

(also this part is not for the exam)

Interesting mathematical objects:

- regularized optimization: $\min_{\theta} J(\theta) + \gamma R(\theta)$
- bayesian MAP estimation: $\max_{\theta} p(\theta|y) \propto p(y|\theta)p(\theta)$

Regularization may be seen with the Bayesian googles

(also this part is not for the exam)

Interesting mathematical objects:

- regularized optimization: $\min_{\theta} J(\theta) + \gamma R(\theta)$
- bayesian MAP estimation: $\max_{\theta} p(\theta|y) \propto p(y|\theta)p(\theta)$

Actually sometimes they coincide!

- L_2 regularization $\Leftrightarrow$ Gaussian prior
- L_1 regularization $\Leftrightarrow$ Laplace prior

From regularization to MAP estimation

Example for linear regression with L_2 regularization

$$\text{Ridge:} \quad \min_{\theta} \|y - X\theta\|^2 + \lambda\|\theta\|^2$$

$$\text{MAP:} \quad \max_{\theta} \mathcal{N}(y|X\theta, \sigma^2 I) \cdot \mathcal{N}(\theta|0, \tau^2 I) \quad \text{with } \lambda = \sigma^2/\tau^2$$

From regularization to MAP estimation

Example for linear regression with L_1 regularization

$$\text{Lasso:} \quad \min_{\theta} \|y - X\theta\|^2 + \lambda\|\theta\|_1$$

$$\text{MAP:} \quad \max_{\theta} \mathcal{N}(y|X\theta, \sigma^2 I) \cdot \text{Laplace}(\theta|0, b)$$

Important implication

if you have your own prior for your own problem, you should design your regularization term in an ad-hoc way

Important implication

if you have your own prior for your own problem, you should design your regularization term in an ad-hoc way

Very interesting example of this: "stable splines kernels" for system identification

Summarizing

Compare different regularization techniques (ridge, lasso, and elastic net) by evaluating their mathematical formulations, graphical interpretations, and practical implications

Interpret regularization paths from Lasso regression plots to identify the relative importance of features in predictive modeling

- L2 and L1 regularization techniques place different additional weights to the cost functions
- graphically seeing how they work is essential to fix the understanding behind them
- L1 is especially useful to perform features selection

Most important python code for this sub-module

Core Scientific Computing

- **NumPy**: Fundamental package for numerical computations
- **SciPy**: Advanced scientific computing (optimization, linear algebra)
- **Matplotlib**: Publication-quality visualization
- **Pandas**: Data manipulation and analysis

Machine Learning Focus

- **scikit-learn**: Main library for LS implementations
  - Ridge/Lasso/ElasticNet implementations
  - Cross-validation tools
  - Regularization path visualization
- **statsmodels**: Formal statistical modeling
- **autograd/JAX**: For advanced gradient computations

Specialized Visualization

- **Seaborn**: Enhanced statistical visuals
- **Plotly**: Interactive regularization path plots
- **mpld3**: D3.js integration for matplotlib

Teaching-Specific Tools

- **ipywidgets**: Interactive demonstrations
- `sklearn-evaluation`: Enhanced model evaluation
- **alive-progress**: For long computations during demos

Self-assessment material

Question 1

What is the primary purpose of regularization in statistical learning?

Potential answers:

I: To increase model complexity and fit training data perfectly
II: To reduce overfitting by trading some bias for lower variance
III: To eliminate all bias from the model estimates
IV: To make computations faster by reducing matrix dimensions
V: I do not know

Question 2

In ridge regression, what Bayesian prior does the L2 penalty term correspond to?

Potential answers:

I: Uniform prior over all parameters
II: Laplace (double exponential) prior
III: Gaussian prior centered at zero
IV: Poisson prior with $=1$
V: I do not know

Question 3

Why does L1 regularization (lasso) tend to produce sparse solutions with exactly zero coefficients?

Potential answers:

I: Because it uses a logarithmic penalty term
II: Due to the sharp corners of the L1 constraint region
III: Because it maximizes the likelihood more aggressively
IV: It doesn't - this is a common misconception
V: I do not know

Question 4

What surprising result does the James-Stein estimator demonstrate about maximum likelihood estimation?

Potential answers:

I: LS estimators always have minimum variance

II: LS can be dominated by shrinkage estimators when estimating multiple parameters

III: LS becomes biased when sample size exceeds 30

IV: LS requires normally distributed errors

V: I do not know

Question 5

When examining a lasso regularization path plot, how should you interpret features whose coefficients become non-zero earliest as decreases?

Potential answers:

I: They are likely measurement errors
II: They should be removed from the model
III: They are the most important predictors
IV: They have the smallest scale
V: I do not know

Recap of sub-module "Regularization"

- adding regularization and non-L2 costs noticeably extends capabilities of estimators, at the cost though of introducing some hyperparameters that need to be tuned too from the data

?