Regularization

UE de Master 2, AOS1
Fall 2022

S. Rousseau

General definition

What is regularization? From Goodfellow et al. 2016:

> *Regularization is any modification we make to a learning algorithm that is intented to reduce its* ***generalization error*** *but not its* ***training error****.*

- Adding a penalty term to a loss function
- Data augmentation
- Early stopping
- ...

Today's course is focused on the first point

What is learning

Learning consists in minimizing a training objective

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \widehat{L}(f)$$

- $\mathcal{H}$ is the set of admissible classifier/regression function
- $\widehat{L}$ is an empirical loss function (computed on a train set)
- $\widehat{f}$ is the learnt solution

Most of the time the set $\mathcal{H}$ is parametrized by a parameter $\boldsymbol{\theta} \in \Theta$

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \widehat{L}(\boldsymbol{\theta})$$

Application to linear regression

- Set of admissible solutions is linear functions

$$\mathcal{H} = \{\text{linear functions}\}$$

- Linear functions on $\mathbb{R}^p$ are parametrized by $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\mathcal{H} = \{\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle , x \in \mathbb{R}^p\}$$

- Training objective is the residual sum of squares (RSS)
  Empirical loss function based on square loss, prediction is $\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle$, observed is y_i

$$\widehat{L}(\boldsymbol{\beta}) = \mathrm{RSS}\,(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)^2$$

Application to linear regression (2)

- The learning algorithm is then

$$\widehat{\beta}^{\text{ols}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \beta \rangle)^2 \qquad \text{(ordinary least square)}$$

- Compact matrix formulation $X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}$ and $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$\widehat{\beta}^{\text{ols}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|\boldsymbol{y} - X\beta\|^2$$

Regularization

- Unregularized objective

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{L}(\boldsymbol{\theta}) \tag{1}$$

- Regularized objective

$$\widehat{\boldsymbol{\theta}}_{\lambda}^{\text{reg}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{L}(\boldsymbol{\theta}) + \lambda \cdot R(\boldsymbol{\theta})$$

  - Training objective
  - Tuning parameter
  - Regularization term

Regularization (2)

Regularized objective

$$\widehat{\boldsymbol{\theta}}^{\text{reg}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, \widehat{L}(\boldsymbol{\theta}) + \lambda \cdot R(\boldsymbol{\theta})$$

- $R(\boldsymbol{\theta})$ penalizes some $\boldsymbol{\theta}$'s
- $\lambda \geqslant 0$ is the strength of the penalty
    - $\lambda = 0$: no penality: regular solution
    - $\lambda \longrightarrow +\infty$, solution tends to $\underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, R(\boldsymbol{\theta})$
    - Some tradeoff has to be found between the two extreme cases

Ridge regularization

Most simple regularization we can think of:

- We choose $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} \theta_i^2$
- Penalizes large parameter: prevents the β_i's from exploding
- Ridge regularization is then

$$\widehat{\boldsymbol{\theta}}_\lambda^{\text{ridge}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \widehat{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|^2 \qquad \text{(ridge regularization)}$$

Also known as L_2-regularization, Tikhonov regularization or weight decay (neural network)

Application to ridge regression

- Previous linear regression learning algorithm was

$$\widehat{\boldsymbol{\beta}}^{\text{ols}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$$

- Adding the ridge regularizer term yields

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} &= \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \qquad \text{(ridge regression)} \\
&= \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)^2 + \lambda \sum_{i=1}^{p} \beta_i^2
\end{aligned}$$

Solution to ridge regression

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\| \boldsymbol{y} - X\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|^2 \qquad \text{(ridge regression)}$$

- Define the penalized residual sum of squares (PRSS) as

$$\text{PRSS}\left(\boldsymbol{\beta}\right) = \left\| \boldsymbol{y} - X\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|^2$$

- PRSS is (strictly) convex w.r.t. $\boldsymbol{\beta}$: unique solution
- By differentiating w.r.t. $\boldsymbol{\beta}$ we get

$$\nabla_{\boldsymbol{\beta}} \text{PRSS} = -2X^T(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} \tag{2}$$

Solution to ridge regression

- Setting the derivative to zero we finally get

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \left(X^T X + \lambda I_p\right)^{-1} X^T \boldsymbol{y} \tag{3}$$

- Fitted values are then

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

- For $\lambda = 0$ we have the OLS solution

$$\boldsymbol{\beta}_{\lambda=0}^{\text{ridge}} = \boldsymbol{\beta}^{\text{ols}}$$

$$\widehat{\boldsymbol{y}}^{\text{ols}} = X\left(X^T X\right)^{-1} X^T \boldsymbol{y}$$

Caveats

- The intercept (if present) should not be part of the regularizing parameter. Two possible strategies:
  - center the design matrix X so there is no intercept
  - or we remove the intercept from the regularizing parameter (set $\boldsymbol{\beta}^{\star}$ as $\boldsymbol{\beta}$ without β_0)
- The features should be on the **same scale**; unlike linear regression ridge regression predictions are sensitive to features rescaling

Properties of ridge regression

- Unlike linear regression, there is always a solution (when $\lambda > 0$)
  - $X^T X$ is positive **semi-definite**
  - $X^T X + \lambda I_p$ is then positive definite when $\lambda > 0$ hence is invertible
- It improves the **conditioning** of the problem
- Like linear regression but unlike Lasso, it admits a **closed form solution**

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

- Invariant to rotation: if $Y = XU^T$ is a rotation of the samples then $\widehat{\boldsymbol{\beta}}_Y = U \widehat{\boldsymbol{\beta}}_X$
- Unlike linear regression, both the $\boldsymbol{\beta}_i$'s estimate and predictions are **biased**
- The $\boldsymbol{\beta}_i$'s estimate are **drawn toward zero** w.r.t the OLS solution
- Might have lower variance than OLS

$\widehat{\beta}_\lambda^{\text{ridge}}$ is biased

Suppose that the design matrix X is fixed (conditioning on it)

Linear case:

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}}^{\text{ols}}\right) &= \mathbb{E}\left(\left(X^T X\right)^{-1} X^T \boldsymbol{y}\right) \\
&= \left(X^T X\right)^{-1} X^T \mathbb{E}(\boldsymbol{y}) \\
&= \left(X^T X\right)^{-1} X^T X \boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

Unbiased!

Ridge case:

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}\right) &= \mathbb{E}\left(\left(X^T X + \lambda I_p\right)^{-1} X^T \boldsymbol{y}\right) \\
&= \left(X^T X + \lambda I_p\right)^{-1} X^T X \boldsymbol{\beta} \\
&= \left(\left(X^T X\right)^{-1}\left(X^T X + \lambda I_p\right)\right)^{-1} \boldsymbol{\beta} \\
&= \left(I_p + \lambda\left(X^T X\right)^{-1}\right)^{-1} \boldsymbol{\beta} \neq \boldsymbol{\beta}
\end{aligned}
$$

Biased!

Data augmentation interpretation

- Let's rewrite the PRSS

$$
\begin{aligned}
\mathrm{PRSS}(\boldsymbol{\beta}) &= \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \\
&= \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)^2 + \sum_{i=1}^{p} \left(0 - \left\langle \sqrt{\lambda}\boldsymbol{e}_i, \boldsymbol{\beta} \right\rangle\right)^2
\end{aligned}
$$

- Same as adding p extra samples in addition to the n $\boldsymbol{x}_i$'s
- Additional samples and observations are $\left(\sqrt{\lambda}\boldsymbol{e}_i, 0\right)$ for $i = 1, \ldots, p$
- Same as adding an observation on each axis that is zero

Data augmentation interpretation (cont'd)

- Switching back to matrix form we define

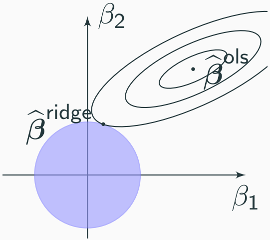
$$X_\lambda = \begin{pmatrix} X & & & \\ \sqrt{\lambda} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \sqrt{\lambda} \end{pmatrix} \quad \text{and} \quad \boldsymbol{y}' = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

- And the PRSS can now be written

$$\left\| \boldsymbol{y}' - X_\lambda \beta \right\|^2 \quad \text{s.t.} \quad X_\lambda = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix} \quad \boldsymbol{y}' = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}$$

Geometric interpretation $\widehat{\beta}^{\text{ridge}}$

- Suppose 2-dimensional case ($p = 2$), $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ is the OLS solution
- Ellipses are level line of the RSS: $\|\boldsymbol{y} - X\beta\|^2$
- A solution for some λ is at the intersection of the L_2 ball and a level line
- Whatever the form of ellipses, ridge solution is systematically drawn toward zero

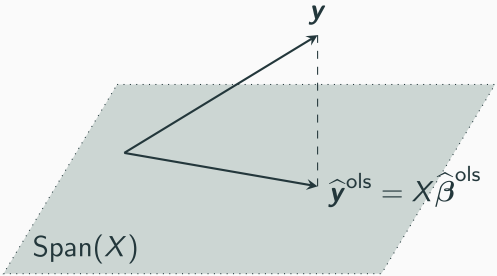


Geometric interpretation of $\widehat{y}^{\text{ols}}$

First see the OLS case

- Let $X = USV^T$ the SVD of X
- Matrix U gathers the (unit) principal components $\mathrm{u}_1, \ldots, \mathrm{u}_k$
- Ordinary least squares orthogonally projects $\boldsymbol{y}$ onto the space spanned by the columns of X:

$$
\begin{aligned}
\widehat{\boldsymbol{y}}^{\text{ols}} &= X\left(X^T X\right)^{-1} X^T \boldsymbol{y} = UU^T \boldsymbol{y} \\
&= \sum_{i=1}^{p} \left(\mathrm{u}_i^T \boldsymbol{y}\right) \mathrm{u}_i
\end{aligned}
$$



Geometric interpretation of $\widehat{y}^{\text{ridge}}$ (cont'd)

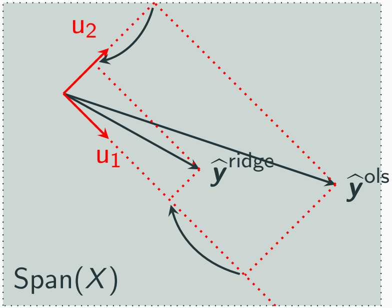
- Ridge regression is doing the same thing plus an additional "shrinking" step

$$
\begin{aligned}
\widehat{\boldsymbol{y}}^{\text{ridge}} &= X\left(X^TX + \lambda I_p\right)^{-1} X^T \boldsymbol{y} = U\left(S\left(S^2 + \lambda I_p\right)^{-1} S\right) U^T \boldsymbol{y} \\
&= \sum_{i=1}^{p} \left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda} \mathrm{u}_i^T \boldsymbol{y} \right) \mathrm{u}_i
\end{aligned}
$$

- Coordinates are now shrunk towards zero since $\dfrac{\sigma_i^2}{\sigma_i^2 + \lambda} < 1$
- Remember that the u_i are here the (unit) principal components of X
- The lesser the variance of the principal component is the greater it is shrunk
- Smooth version of PCA followed by linear regression

Geometric interpretation of $\widehat{y}^{\text{ridge}}$ (cont'd)

- Coordinate of $\widehat{\boldsymbol{y}}^{\text{ols}}$ along u_1 is shrunk by $\dfrac{\sigma_1^2}{\sigma_1^2 + \lambda}$
- Coordinate of $\widehat{\boldsymbol{y}}^{\text{ols}}$ along u_2 is shrunk by $\dfrac{\sigma_2^2}{\sigma_2^2 + \lambda}$



Geometric interpretation of $\widehat{\beta}_\lambda^{\text{ridge}}$

What is the link between the unknown parameter β and its ridge estimate $\widehat{\beta}_\lambda^{\text{ridge}}$?

- Let $X = USV^T$ the SVD of X
- Recall that V gathers the principal directions (new basis of representation)
- Let's look at $\widehat{\beta}_\lambda^{\text{ridge}}$ in basis V, we can show that

$$V^T\widehat{\beta}_\lambda^{\text{ridge}} = \left(I_p + \lambda S^{-2}\right)^{-1} V^T\widehat{\beta}^{\text{ols}}$$

  In the basis defined by V, β is shrunk

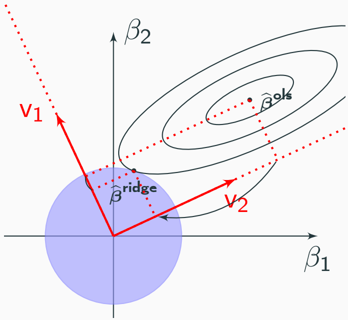
- If we look at the ith coordinate

$$\left(V^T\widehat{\beta}_\lambda^{\text{ridge}}\right)_i = \frac{\sigma_i^2}{\lambda + \sigma_i^2}\left(V^T\widehat{\beta}^{\text{ols}}\right)_i$$

Geometric interpretation of $\widehat{\beta}_\lambda^{\text{ridge}}$

$$\left(V^T\widehat{\beta}_\lambda^{\text{ridge}}\right)_i = \frac{\sigma_i^2}{\lambda+\sigma_i^2}\left(V^T\widehat{\beta}^{\text{ols}}\right)_i$$

- Coordinate of $\widehat{\beta}_\lambda^{\text{ridge}}$ along v_1 is shrunk by $\dfrac{\sigma_1^2}{\sigma_1^2+\lambda}$
- Coordinate of $\widehat{\beta}_\lambda^{\text{ridge}}$ along v_2 is shrunk by $\dfrac{\sigma_2^2}{\sigma_2^2+\lambda}$



Gradient descent interpretation

- General gradient descent update of step η

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} - \eta \nabla L\left(\boldsymbol{\theta}^{k}\right),$$

- Gradient descent update for OLS ($L = \mathrm{RSS}$)

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^{k} - \eta \nabla \,\mathrm{RSS}\left(\boldsymbol{\beta}^{k}\right) \tag{4}$$

- Gradient descent update for ridge regression

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^{k} - \eta \nabla \,\mathrm{PRSS}\left(\boldsymbol{\beta}^{k}\right)$$

- Which writes

$$\boldsymbol{\beta}^{k+1} = (1 - 2\eta\lambda)\boldsymbol{\beta}^{k} - \eta \nabla \,\mathrm{RSS}\left(\boldsymbol{\beta}^{k}\right) \tag{5}$$

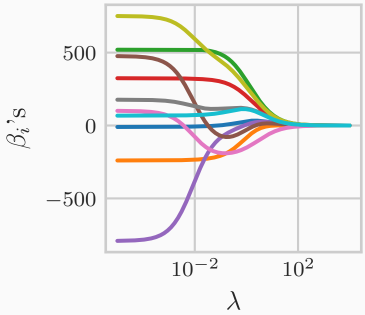
  The update use a shrunk version of $\boldsymbol{\beta}^{k}$

Effect on the β_i's: the regularization path

Regularization path : plot of β_i's against regularization parameter λ.

Here for some centered dataset with 10 covariates:

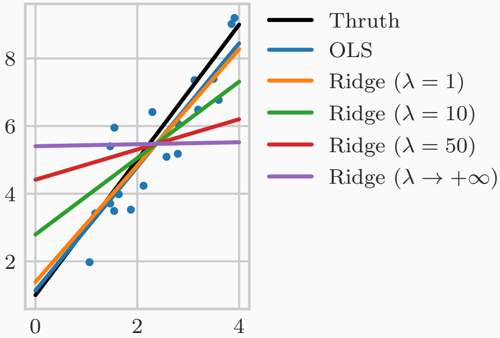
- Linear regression estimates at $\lambda = 0$
- When regularization is too strong, we fit the constant function
- β_i's are shrunk smoothly towards 0 as λ increases
- All the β_i's might not be non-increasing but $\|\boldsymbol{\beta}\|$ is
- The β_i's are never zero



Effect on the fitted line

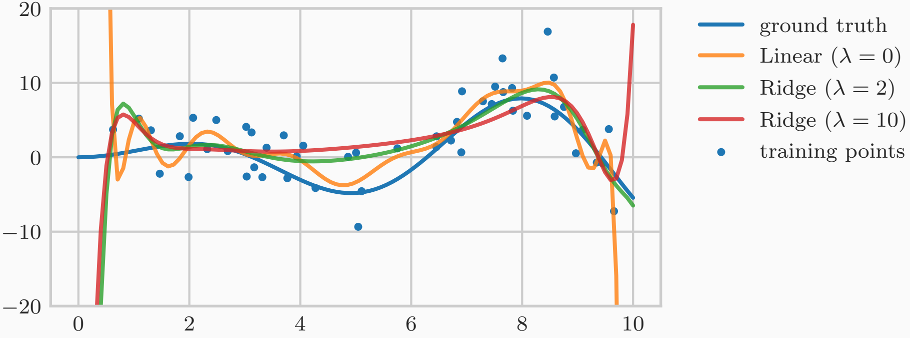
In the case of simple linear regression ($p = 1$)

- Regularizing is like adding the point $(\sqrt{\lambda}, 0)$
- As λ goes to infinity, we fit a constant function



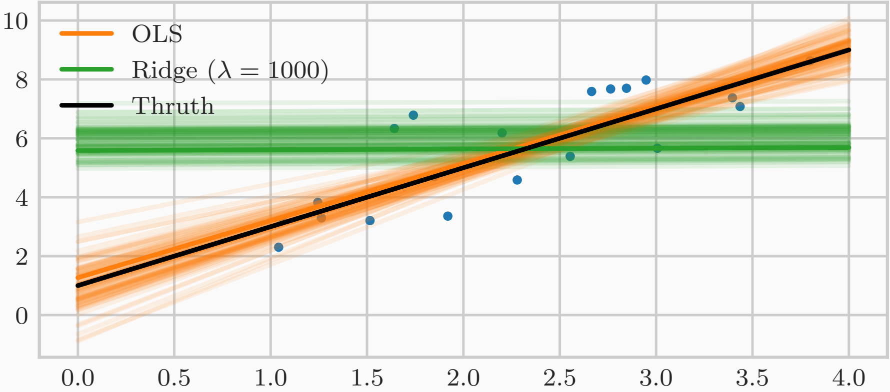
Effect on fitted curve

Adding polynomial features (degree 15): $X_i, X_i^2, \ldots, X_i^{15}$



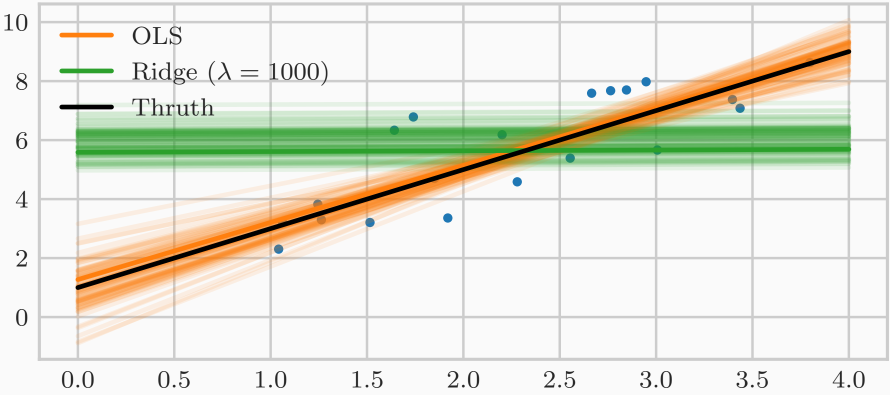
Effect on the β_i's estimates: bias–variance tradeoff

The slope β_1 is biased but variance of estimations is smaller



Effect on predictions: bias–variance tradeoff

Predictions at 0 (for example) are biased but less spread out



Lasso regularization

Other famous regularization

- We choose $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{p} |\theta_i|$
- Penalizes large parameter: prevents the β_i's from exploding
- Lasso regularization is then

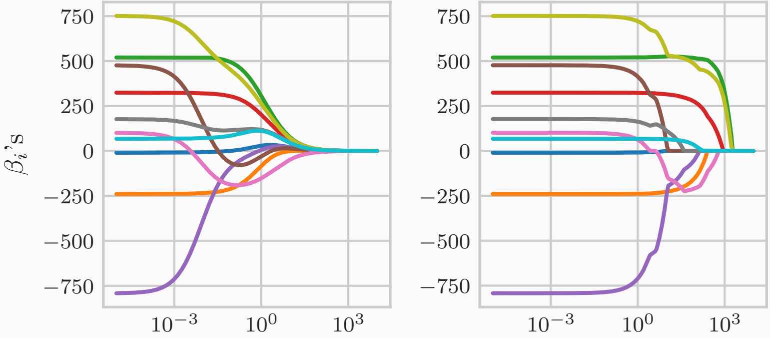
$$\widehat{\boldsymbol{\theta}}_\lambda^{\text{lasso}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \widehat{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 \qquad \text{(Lasso regularization)}$$

- Lasso linear regression

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \qquad \text{(Lasso regression)}$$

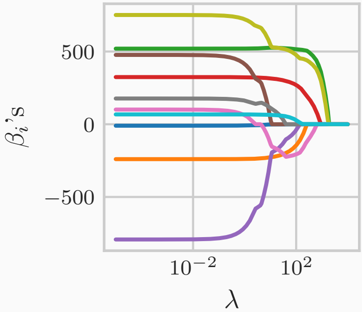
Sparsity promoting property

Compare the coefficients β_i's with ridge and lasso regularization



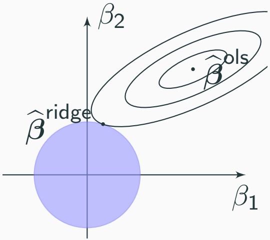
Effect on the β_i's: the regularization path

- Linear regression estimates at $\lambda = 0$
- Piecewise linear regularization path
- β_i's are shrunk as λ increases
- All the β_i's are shrunk to exactly zero at some point: **sparsity promoting effect**
- When regularization is too strong, we fit the constant function

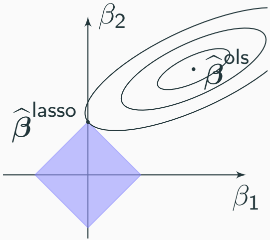


Explaining the sparsity property

- $\widehat{\beta}^{\text{ols}}$ is the ordinary least square solution



Ridge can be anywhere on the L_2 ball



Lasso solution lies on edge of L_1 ball (sparse solution)

Geometric interpretation of $\widehat{\beta}_\lambda^{\text{lasso}}$

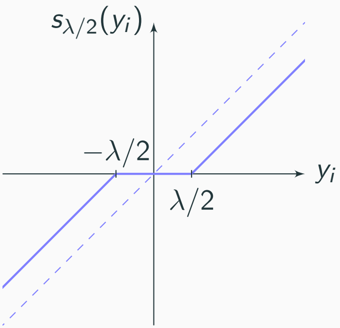
For simplicity, suppose that $n = p$ and X is the identity matrix

- OLS solution is: $\widehat{\boldsymbol{\beta}}^{\text{ols}} = \boldsymbol{y}$
- Lasso regularization reads: $\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$

$$\left(\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}\right)_i = \arg\min_{\beta_i \in \mathbb{R}} (y_i - \beta_i)^2 + \lambda |\beta_i|$$

$$\left(\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}\right)_i = \begin{cases} \max(y_i - \lambda/2, 0) & \text{if } y_i \geqslant 0 \\ \max(y_i + \lambda/2, 0) & \text{if } y_i < 0 \end{cases}$$

(soft thresholding)



Gradient descent interpretation

- Using subgradient to differentiate $\|\cdot\|_1$

$$\nabla_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 = \operatorname{sign}(\boldsymbol{\beta})$$

- Gradient descent update for lasso regression

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \eta \nabla \operatorname{RSS}\left(\boldsymbol{\beta}^k\right) - \eta\lambda \operatorname{sign}\left(\boldsymbol{\beta}^k\right)$$

- Shrink the β_i's regardless of the β_i's magnitude

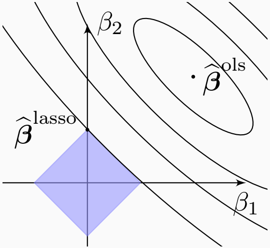
Lasso properties

- No closed from solution
- Convex problem
- Feature selection ability
- Biased predictions and parameter estimate
- Might be unstable if highly correlated variables

Elastic net

Why elastic-net?

- Ridge regression is not selecting variables (all the $\boldsymbol{\beta}_i$'s are nonzero)
- Lasso regression is but in an unstable way
- Small changes in X might lead to entirely different set of selected predictors

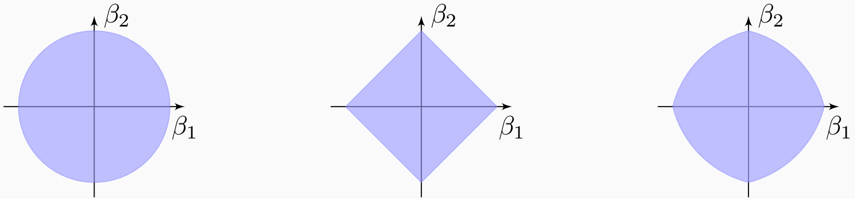


Elastic net

Mixing the two strategies

$$\widehat{\boldsymbol{\theta}}^{\text{elastic}}_{\lambda,\alpha} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \widehat{L}(\boldsymbol{\theta}) + \lambda\Big(\alpha \left\|\boldsymbol{\theta}\right\|_1 + (1-\alpha)\left\|\boldsymbol{\theta}\right\|_2^2\Big) \qquad \text{(Elastic net regularization)}$$

- λ is the regularizing parameter
- α controls the balance between L_1 and L_2 regularizing terms

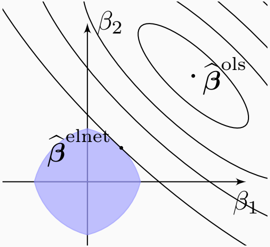


(a) L_2: no selection **(b)** L_1: unstable selection **(c)** Elastic net: stable selection

Elastic net

Explaining the stability of elastic net regularization

- Corners are still sharp: elastic net is still encouraging sparsity
- Elastic net ball is also round (4 portions of (big) circles): stable when some variables are strongly correlated



Ridge/Lasso/Elastic net regression in Python and Scikit–Learn

- Import the PCA module

```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
```

- Instantiate (no parameter), fit and predict

```
lr = LinearRegression()
lr.fit(X, y)
res = lr.predict(new_X)
```

- Instantiate with tuning parameter `alpha`

```
lr = Ridge(alpha=1.0)
lr.fit(X, y)
res = lr.predict(new_X)
```

References i

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics New York, 2001.

[2] Ian Goodfellow et al. *Deep Learning*. Vol. 1. MIT press Cambridge, 2016.