Ridge regularization: an essential concept in data science

Based on the article of Trevor Hastie 2020

04-2021

Bascou Florent
Lefort Tanguy

University of Montpellier

Content

Ridge [1] regularization

Why use it?

We denote $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ such that $y \simeq X\beta$.
Ordinary least squares:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 \iff \hat{\beta} = (X^\top X)^{-1} X^\top y,$$

Problems that can happen:

- $X^\top X$ may be ill conditioned ($\kappa = \frac{\text{largest singular value}}{\text{smallest singular value}} \gg$)
- $p > n$ leads to infinite number of solutions for OLS.

[1] Tikhonov (1943); Hoerl and Kennard (1970)

Ridge[1] regularization
Why use it?

We denote $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ such that $y \simeq X\beta$.
Ordinary least squares:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 \iff \hat{\beta} = (X^\top X)^{-1} X^\top y,$$

Problems that can happen:

- $X^\top X$ may be ill conditioned ($\kappa = \frac{\text{largest singular value}}{\text{smallest singular value}} \gg$)
  - Shift spectrum by λ using $X^\top X + \lambda \mathrm{Id}$.
- $p > n$ leads to infinite number of solutions for OLS.
  - Add penalty to recover unicity.

[1] Tikhonov (1943); Hoerl and Kennard (1970)

Ridge[1] regularization

Why use it?

We denote $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ such that $y \simeq X\beta$.
Ordinary least squares:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 \iff \hat{\beta} = (X^\top X)^{-1} X^\top y,$$

Problems that can happen:

- $X^\top X$ may be ill conditioned ($\kappa = \frac{\text{largest singular value}}{\text{smallest singular value}} \gg$)
  - Shift spectrum by λ using $X^\top X + \lambda \mathrm{Id}$.
- $p > n$ leads to infinite number of solutions for OLS.
  - Add penalty to recover unicity.

Ridge estimator

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \iff \hat{\beta}_{ridge} = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y.$$

[1] Tikhonov (1943); Hoerl and Kennard (1970)

- Coefficients with smaller ℓ_2 norm,
- Handles the multicolinearity issue,
- Don't apply only to linear models, in general:

$$\arg\min_{w} f(w) + \lambda \|w\|_2^2 \ .$$

- Coefficients with smaller ℓ_2 norm,
- Handles the multicolinearity issue,
- Don't apply only to linear models, in general:

$$\arg\min_w f(w) + \lambda \|w\|_2^2 \quad .$$

Some cons (because there must be some)

- Alone won't give sparse solutions, . . .
- **BUT** can be combined with a LASSO[2] which gives the Elastic-Net [3]:

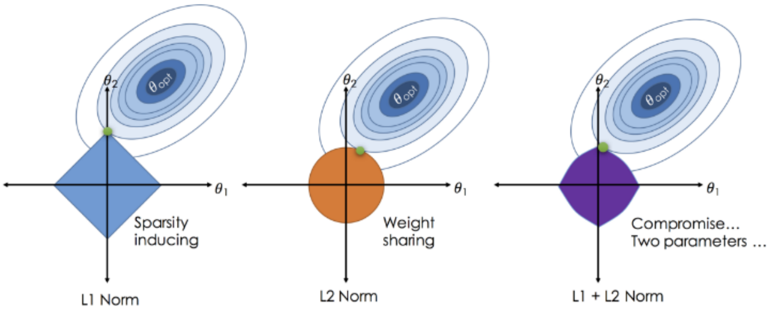
$$\arg\min_\beta \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad .$$

- Introduce an hyperparameter that needs tuning.

[2] Tibshirani (1996)
[3] Zou and Hastie (2005)

Different penalties

And different shapes



Source: `https://towardsdatascience.com/`

Solution is given by : $\hat{\beta}_\lambda = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y$
Problem : λ is a tuning parameter $\Longrightarrow$ computing many $\hat{\beta}_\lambda$

Solution is given by : $\hat{\beta}_\lambda = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y$
Problem : λ is a tuning parameter $\Longrightarrow$ computing many $\hat{\beta}_\lambda$

Use ONE SVD ($X = UDV^\top$) to compute many estimations

$$\hat{\beta}_\lambda = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y = V(D^\top D + \lambda \mathrm{Id})^{-1} D^\top U^\top y = \sum_{j=1}^{rg(X)} v_j \frac{d_j}{d_j^2 + \lambda} \langle u_j | y \rangle$$

Ridge estimator

How to compute it efficiently

Solution is given by : $\hat{\beta}_\lambda = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y$
Problem : λ is a tuning parameter $\Longrightarrow$ computing many $\hat{\beta}_\lambda$

Use ONE SVD ($X = UDV^\top$) to compute many estimations

$$\hat{\beta}_\lambda = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y = V(D^\top D + \lambda \mathrm{Id})^{-1} D^\top U^\top y = \sum_{j=1}^{rg(X)} v_j \frac{d_j}{d_j^2 + \lambda} \langle u_j | y \rangle$$

Use ONE SVD ($X = UDV^\top$) to compute many predictions

$$\hat{y}_\lambda = X\hat{\beta}_\lambda = UD(D^\top D + \lambda \mathrm{Id})^{-1} D^\top U^\top y = \sum_{j=1}^{rg(X)} u_j \frac{d_j^2}{d_j^2 + \lambda} \langle u_j | y \rangle$$

Let us denote :

- $\hat{\beta}_\lambda^{(-i)}$ the estimated coefficients without using the pair (x_i, y_i).
- $R^\lambda = X(X^\top X + \lambda \mathrm{Id})^{-1} X^\top$ the Ridge operator matrix

Let us denote :

- $\hat{\beta}_\lambda^{(-i)}$ the estimated coefficients without using the pair (x_i, y_i).
- $R^\lambda = X(X^\top X + \lambda \mathrm{Id})^{-1} X^\top$ the Ridge operator matrix

Easy computation for LOO-CV

$$LOO_\lambda = \sum_{i=1}^{n} (y_i - x_i^\top \hat{\beta}_\lambda^{(-i)})^2 = \sum_{i=1}^{n} \frac{(y_i - x_i^\top \hat{\beta}_\lambda)^2}{(1 - R_{ii}^\lambda)^2}$$

Let us denote :

- $\hat{\beta}_{\lambda}^{(-i)}$ the estimated coefficients without using the pair (x_i, y_i).
- $R^{\lambda} = X(X^\top X + \lambda \mathrm{Id})^{-1} X^\top$ the Ridge operator matrix

Easy computation for LOO-CV

$$LOO_{\lambda} = \sum_{i=1}^{n} (y_i - x_i^\top \hat{\beta}_{\lambda}^{(-i)})^2 = \sum_{i=1}^{n} \frac{(y_i - x_i^\top \hat{\beta}_{\lambda})^2}{(1 - R_{ii}^{\lambda})^2},$$

$$R^{\lambda} = X(X^\top X + \lambda \mathrm{Id})^{-1} X^\top = UD(D^\top D + \lambda \mathrm{Id})^{-1} D^\top U = US(\lambda)U.$$

with $S(\lambda)$ the diagonal shrinkage matrix with elements $\frac{d_j^2}{d_j^2 + \lambda}$.

Suppose that the data arises from a linear model with *i.i.d* centered errors ε_i

$$y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n$$

Then, the Ridge estimate $\hat{\beta}_\lambda$ is a biased estimate of β.

Suppose that the data arises from a linear model with *i.i.d* centered errors ε_i

$$y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \dots, n$$

Then, the Ridge estimate $\hat{\beta}_\lambda$ is a biased estimate of β.

Bias - Covariance matrix

If the x_i are assumed fixed, $n > p$ and X has full column rank, we get :

$$\mathrm{Bias}(\hat{\beta}_\lambda) = \sum_{j=1}^{p} v_j \frac{\lambda}{d_j^2 + \lambda} \langle v_j | \beta \rangle,$$

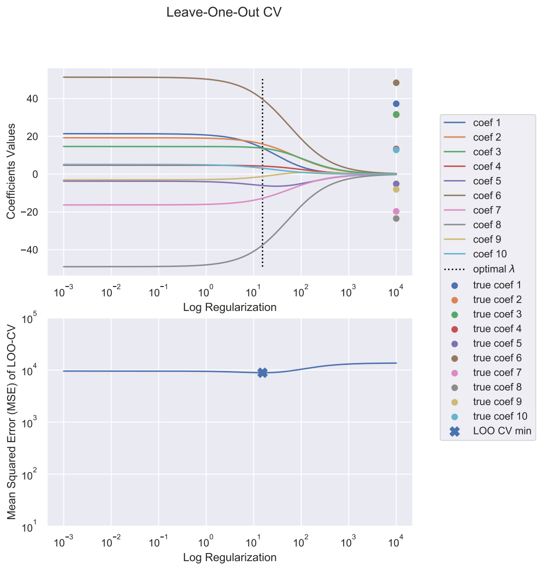
$$\mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2} v_j v_j^\top \quad \text{with } \sigma^2 = \mathrm{Var}(\varepsilon_i).$$

The smaller the j^{th} singular value is, the bigger the shrinkage associated is.

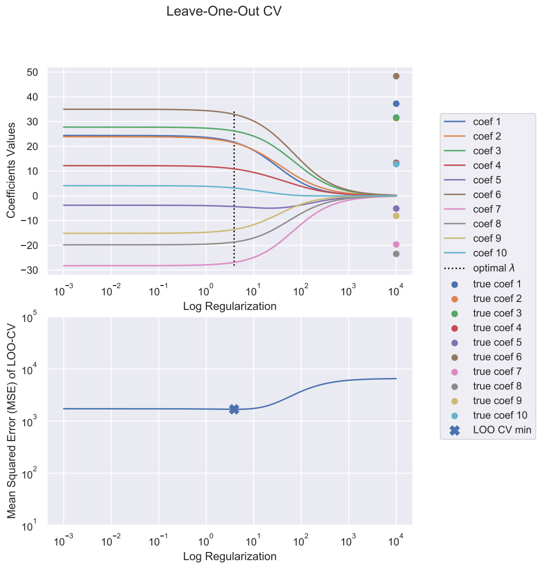
Ridge estimator

An example of LOO CV and bias

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 1$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$



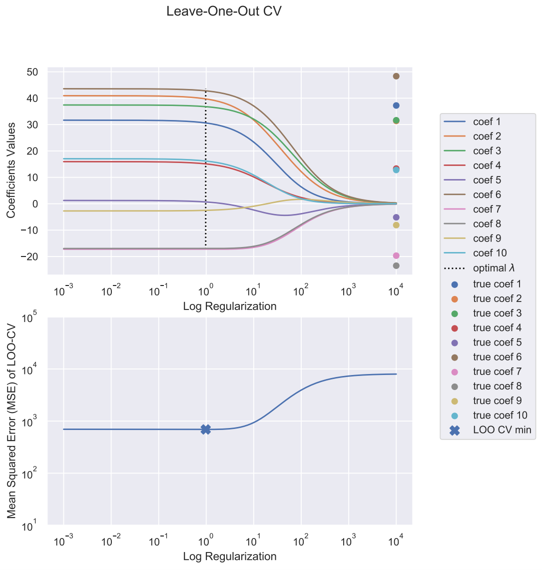
- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 2$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$



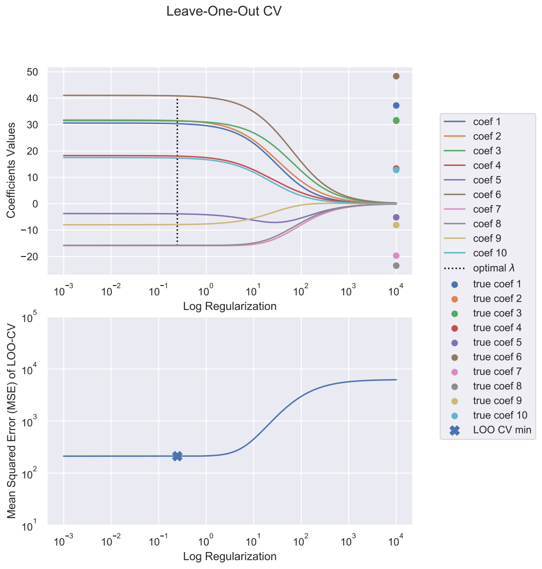
Ridge estimator

An example of LOO CV and bias

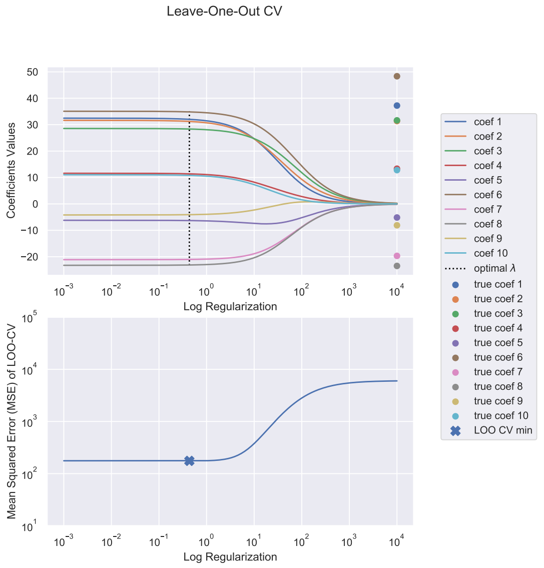
- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 3$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$



- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 5$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

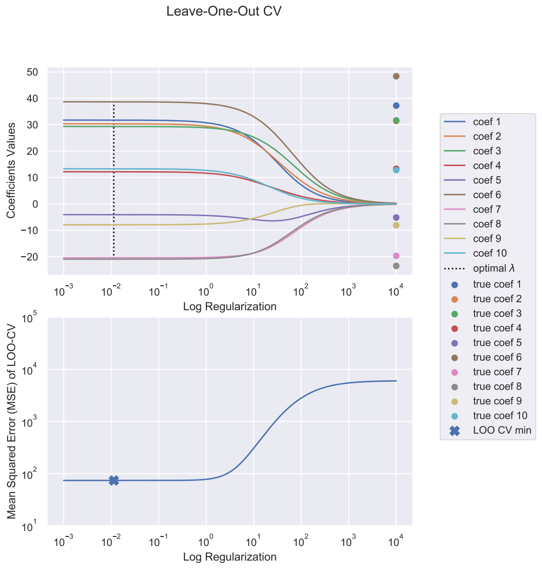


- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 7$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$



Ridge estimator
An example of LOO CV and bias

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 50, p = 10$
- SNR $= 10$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$



Reduce the complexity

If $p > n$, $(X^\top X + \lambda \mathrm{Id})^{-1} \in \mathbb{R}^{p \times p}$ is costly. But we can actually only solve a $n \times n$ system thanks to the relation *(proof with SVD)*:

$$X^\top (XX^\top + \lambda \mathrm{Id})^{-1} y = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y \quad .$$

Reduce the complexity

If $p > n$, $(X^\top X + \lambda \mathrm{Id})^{-1} \in \mathbb{R}^{p \times p}$ is costly. But we can actually only solve a $n \times n$ system thanks to the relation *(proof with SVD)*:

$$X^\top (XX^\top + \lambda \mathrm{Id})^{-1} y = (X^\top X + \lambda \mathrm{Id})^{-1} X^\top y \ .$$

So we can write $\hat{\beta} = X^\top u$ with $u \in \mathbb{R}^n$. Denote $K = XX^\top$ the **Gram** matrix,

$$\begin{aligned} \hat{y} &= X\hat{\beta} = XX^\top u \\ &= K(K + \lambda \mathrm{Id})^{-1} y \end{aligned}$$

- New ridge problem smaller to solve.
- **Opens the door for a lot more!**

Suppose $y_i = \varphi(X_i)$, then $\hat{\varphi} = \arg\min_{\varphi \in \mathcal{H}} \|y - \varphi(X)\|_2^2 + \lambda \|\varphi\|_{\mathcal{H}}$.

Representer theorem [4]

Take $\mathcal{H}$ RKHS of kernel K of map Φ *s.t.* $K(x,y) = \langle \Phi(x), \Phi(y) \rangle = (\Phi(x))(y)$.
For $f \in \mathcal{H}$, $f(x) = \langle f, \Phi(x) \rangle$. Then,

$$\varphi = \sum_{i=1}^{n} \alpha_i K(x_i, \bullet) \quad .$$

[4] Schölkopf et al. (2001)

Suppose $y_i = \varphi(X_i)$, then $\hat{\varphi} = \arg\min_{\varphi \in \mathcal{H}} \|y - \varphi(X)\|_2^2 + \lambda \|\varphi\|_{\mathcal{H}}$.

Representer theorem [4]

Take $\mathcal{H}$ RKHS of kernel K of map Φ *s.t.* $K(x, y) = \langle \Phi(x), \Phi(y) \rangle = (\Phi(x))(y)$.
For $f \in \mathcal{H}$, $f(x) = \langle f, \Phi(x) \rangle$. Then,

$$\varphi = \sum_{i=1}^{n} \alpha_i K(x_i, \bullet) \quad .$$

Denote $\mathbf{K} = (\mathbf{K}_{ij})_{ij} = (K(x_i, x_j))_{ij}$. The problem is now

$$\hat{\alpha} = \arg\min_{\alpha} \|y - \mathbf{K}\alpha\|_2^2 + \lambda \alpha^\top \mathbf{K} \alpha.$$

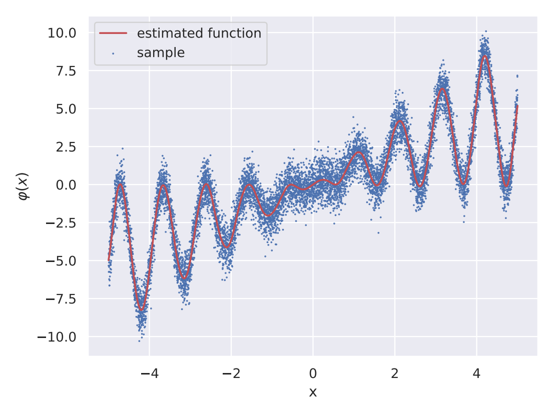
- first order conditions: $\nabla = (\mathbf{K}\mathbf{K} + \lambda \mathbf{K})\hat{\alpha} - \mathbf{K}y = 0$,
- solution:

$$\hat{\alpha} = (\mathbf{K} + \lambda \mathrm{Id})^{-1} y \quad .$$

[4] Schölkopf et al. (2001)

Gaussian kernel to estimate $\varphi(t) = t + t\cos(6t)$ on noised data ($\sigma = 0.8$),

$$K(x, y) = \exp\left\{-\gamma \|x - y\|_2^2\right\}, \gamma = \frac{1}{2 \cdot 0.2^2} \quad .$$

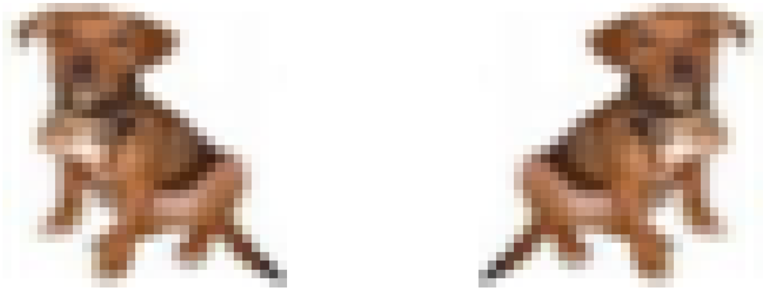


Data augmentation[5]

With images

Example of a small dog from CIFAR10 dataset

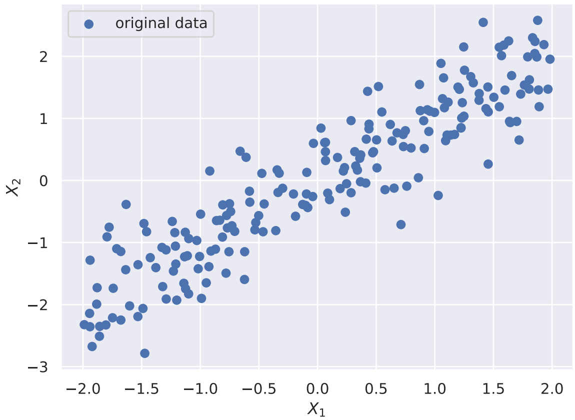
Example of a small dog from CIFAR10 dataset



[5] Chollet and Allaire (2018)

Example of a small dog from CIFAR10 dataset

[5] Chollet and Allaire (2018)



- **Goal:** Use the data we have to create new points,
- **But** can't flip it / make a small rotation, . . .

One way to do so: create perturbated points from the data we have:

$$x_{ij} = x_i + \varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}\left(0, \frac{\lambda}{n}\mathrm{Id}\right), i \in [n], j \in [m] \quad ,$$

with associated response $y_{ij} = y_i \quad .$

Added points compensate each other

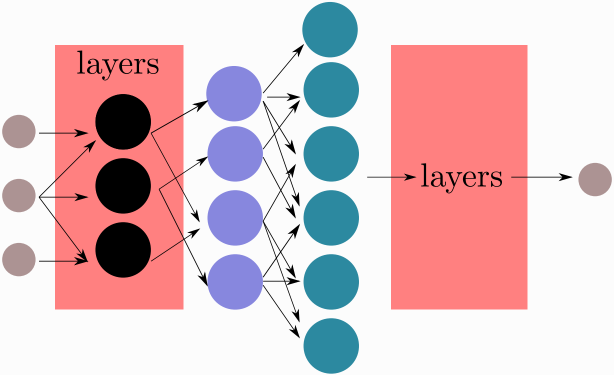
$$\sum_i \frac{1}{m} \sum_j x_{ij} x_{ij}^\top \simeq (XX^\top + \lambda \mathrm{Id})$$

OLS with $X^{augmented} \simeq$ Ridge with X

Dropout [6]

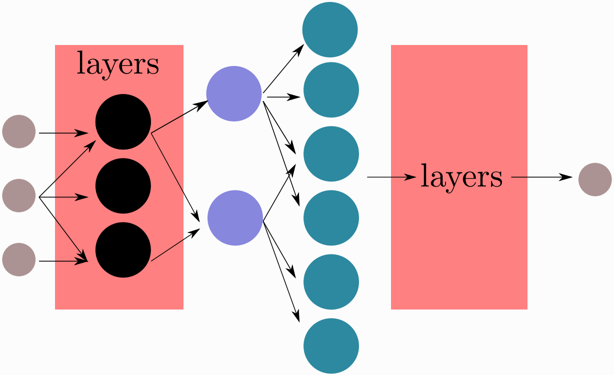
Randomized drop

- Randomly set units of a layer to 0 with probability ϕ to avoid overfitting,
- Inflate surviving ones by $1/(1-\phi)$ factor as compensation.



[6] Srivastava et al. (2014)

- Randomly set units of a layer to 0 with probability ϕ to avoid overfitting,
- Inflate surviving ones by $1/(1-\phi)$ factor as compensation.



[6] Srivastava et al. (2014)

Denoting $(I_{ij})_{ij}$ the dropout mask on $X = (x_{ij})_{ij}$ then cost function is:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij} I_{ij} \beta_j\right)^2 \quad .$$

Note that $\mathbb{E}[I_{ij}] = 0\phi + \frac{1}{1-\phi}(1-\phi) = 1$, thus:

$$\mathbb{E}\left[\frac{\partial L(\beta)}{\partial \beta}\right] = XX^\top \beta - X^\top y + \frac{\phi}{1-\phi}\mathrm{diag}(\|x_j\|^2)_{j=1}^{p}\beta \quad ,$$

[7] Wager et al. (2013)

Denoting $(I_{ij})_{ij}$ the dropout mask on $X = (x_{ij})_{ij}$ then cost function is:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij} I_{ij} \beta_j\right)^2 \quad .$$

Note that $\mathbb{E}[I_{ij}] = 0\phi + \frac{1}{1-\phi}(1-\phi) = 1$, thus:

$$\mathbb{E}\left[\frac{\partial L(\beta)}{\partial \beta}\right] = XX^\top \beta - X^\top y + \frac{\phi}{1-\phi}\mathrm{diag}(\|x_j\|^2)_{j=1}^{p}\beta \quad ,$$

Link with ridge

Solving first order conditions:

$$\hat{\beta}^{dropout} = \left(X^\top X + \frac{\phi}{1-\phi}\,\mathrm{diag}(\|x_j\|^2)\right)^{-1} X^\top y \quad .$$

Normalizing out data leads **in average** to the ridge estimator for $\lambda = \frac{\phi}{1-\phi}$.

[7] Wager et al. (2013)

Dropout

In practice with PyTorch

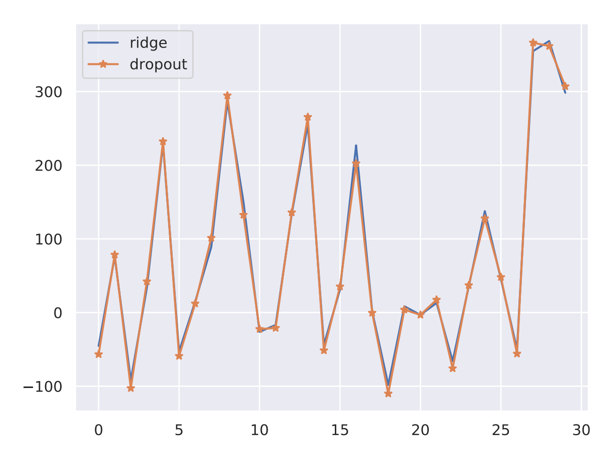
No hastle to implement a dropout layer in a Neural Network with PyTorch with this inflation rate:

```
class MyModel(nn.Module):
    def __init__(self, phi):
        super(MyModel, self).__init__()
        # all your great layers
        self.drop = nn.Dropout(phi)

    def forward(self, x):
        # forward x until the desired layer
        out = self.drop(current_x)
        # forward out until the output layer
        return out
```

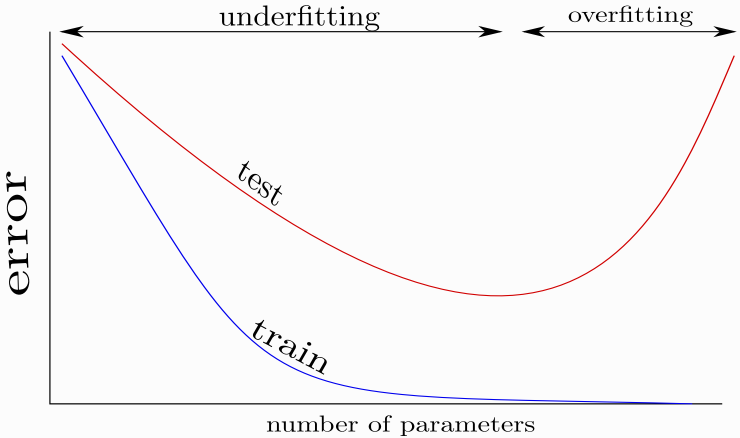
Dropout experiment

With $\phi = 0.5$ (generally $0.3 \leqslant \phi \leqslant 0.5$) and 5 repetitions:

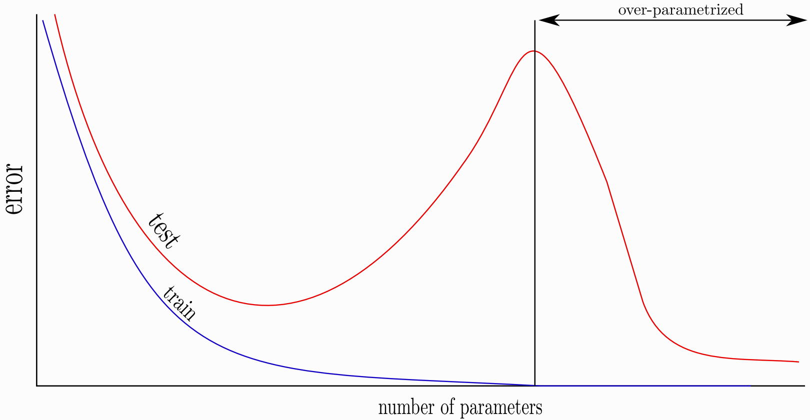


Coefficients of β for the ridge and dropout method with $n = 80$, $p = 30$ and ridge penalty $\lambda = \phi/(1-\phi) = 1$ are close.

Phenomenon observed in Deep Learning, Random Forests …



Phenomenon observed in Deep Learning, Random Forests ...

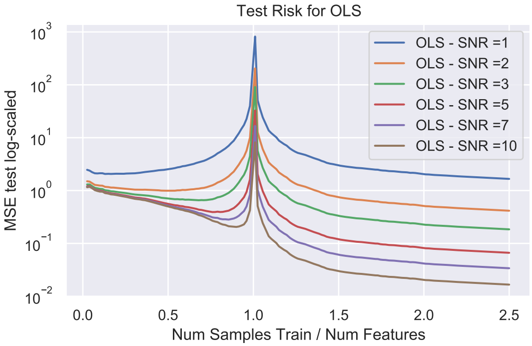


- interpolate the data $\Longrightarrow$ ℓ_2 norm of $\hat{\beta}$ is high,
- smaller ℓ_2 norm for $\hat{\beta}$ generalizes better and we keep zero training error.

Double descent

Example with samples (Nakkiran et al., 2020) - 25 repetitions

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 1000, p = 200$, $\|\beta^*\|_2 = 1$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$,

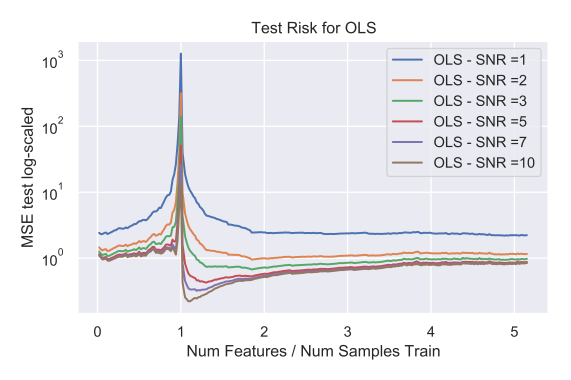


Ordinary Least Square produces a double descent phenomenon when training set has same size as the number of features.

Double descent

Example with features (Nakkiran et al., 2020) - 25 repetitions

- To generate data :
  - Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
  - $n = p = 100$, $\|\beta^*\|_2 = 1$
  - $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$,
- To learn model :
  - if $p \leqslant n$, use X from data ;
  - if $p > n$, merge columns to X from random distribution.



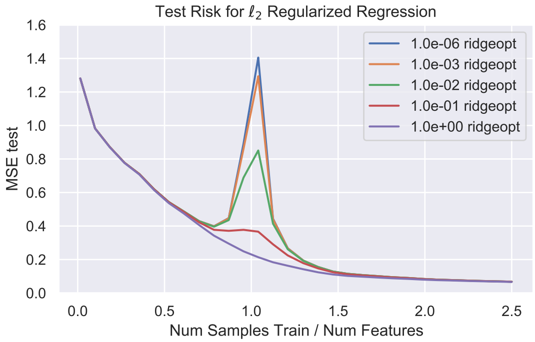
Ordinary Least Square produces a double descent phenomenon when the number of features is the same as the number of samples.

Double descent

With samples (Nakkiran et al., 2020) - 25 repetitions

In some cases ridge regularization can help get a monotonous error curve.

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 1000, p = 200$, $\|\beta^*\|_2 = 1$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$,
- $\lambda_{opt} = \sigma^2 p$

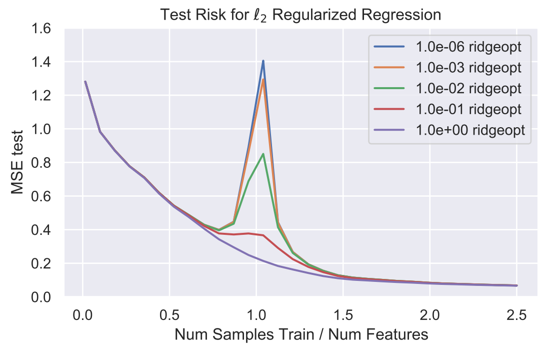


Double descent

With samples (Nakkiran et al., 2020) - 25 repetitions

In some cases ridge regularization can help get a monotonous error curve.

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 1000, p = 200$, $\|\beta^*\|_2 = 1$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$,
- $\lambda_{opt} = \sigma^2 p$



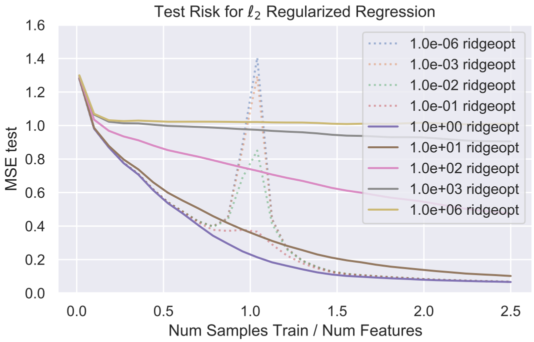
We can achieve monotonic test error decrease with ridge regularization varying p or n for the linear models with isotropic covariates.

Double descent

With samples (Nakkiran et al., 2020) - 25 repetitions

In some cases ridge regularization can help get a monotonous error curve.

- Isotropic data: $X \sim \mathcal{N}(0, \mathrm{Id})$,
- $n = 1000, p = 200$, $\|\beta^*\|_2 = 1$
- $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$,
- $\lambda_{opt} = \sigma^2 p$



This property remains true if $\lambda \geqslant \lambda_{opt}$ varying n for the linear models with isotropic covariates.

Rank selection for matrix problem

Suppose $X \in \mathbb{R}^{m \times n}$. Then we get the rank selection for matrix problem :

$$\min_{M} \|X - M\|_F^2 \quad \text{s. t.} \quad rank(M) \leqslant q.$$

Rank selection for matrix problem

Suppose $X \in \mathbb{R}^{m \times n}$. Then we get the rank selection for matrix problem :

$$\min_{M} \|X - M\|_F^2 \quad \text{s. t.} \quad rank(M) \leqslant q.$$

- Solution: set all the singular values of D to zero, except the q largest.
- Problem: the rank constraint makes this problem non-convex

Rank selection for matrix problem

Suppose $X \in \mathbb{R}^{m\times n}$. Then we get the rank selection for matrix problem :

$$\min_{M} \|X - M\|_F^2 \quad \text{s. t.} \quad rank(M) \leqslant q.$$

- Solution: set all the singular values of D to zero, except the q largest.
- Problem: the rank constraint makes this problem non-convex

Convex relaxation [8]- LASSO version of rank selection matrix

$$\widetilde{M} = \arg\min_{M} \|X - M\|_F^2 + \lambda \|M\|_*$$

with $\|M\|_*$ denoting the nuclear norm - sum of singular values.

[8] Fazel (2002)

Rank selection for matrix problem

Suppose $X \in \mathbb{R}^{m \times n}$. Then we get the rank selection for matrix problem :

$$\min_{M} \|X - M\|_F^2 \quad \text{s. t.} \quad rank(M) \leqslant q.$$

- Solution: set all the singular values of D to zero, except the q largest.
- Problem: the rank constraint makes this problem non-convex

Convex relaxation [8]- LASSO version of rank selection matrix

$$\widetilde{M} = \arg\min_{M} \|X - M\|_F^2 + \lambda \|M\|_*$$

with $\|M\|_*$ denoting the nuclear norm - sum of singular values.

- Solution: soft-tresholding the singular values: $\max(d_i - \lambda, 0)$

[8] Fazel (2002)

Double Ridge Problem [9]

Let $X \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times q}$ and $B \in \mathbb{R}^{n \times q}$. We get the double ridge problem :

$$\widetilde{A}, \widetilde{B} = \arg\min_{A,B} \|X - AB^\top\|_F^2 + \lambda \|A\|_F^2 + \lambda \|B\|_F^2$$

[9] Srebro et al. (2005)

Double Ridge Problem [9]

Let $X \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{m\times q}$ and $B \in \mathbb{R}^{n\times q}$. We get the double ridge problem :

$$\widetilde{A}, \widetilde{B} = \arg\min_{A,B} \|X - AB^\top\|_F^2 + \lambda\|A\|_F^2 + \lambda\|B\|_F^2$$

- This is a ℓ_2 bi-convex problem.
- The solution is the same as the ℓ_1 convex relation problem: $\widetilde{A}\widetilde{B}^\top = \widetilde{M}$

Interest of this property - when SVD fail

If X is massive and sparse, usefull to compute a low-rank matrix approximation by alternating ridge regression.
Given A, we get B by:

$$B^\top = \left(A^\top A + \frac{\lambda}{2}\mathrm{Id}_q\right)^{-1} A^\top X$$

[9] Srebro et al. (2005)

Conclusion

The Ridge regularization is linked to several modern techniques, sometimes hidden behind them.

- LASSO methods are powerful, but SO ARE RIDGE'S,
- it is easy to implement,
- has computational speed-ups from theoretical results.

https://github.com/tanglef/ml_mtp

Conclusion

The Ridge regularization is linked to several modern techniques, sometimes hidden behind them.

- LASSO methods are powerful, but SO ARE RIDGE'S,
- it is easy to implement,
- has computational speed-ups from theoretical results.

One or the other?

The best of both worlds can be used with Elastic-Net regularization.

`https://github.com/tanglef/ml_mtp`

References

François Chollet and Joseph J Allaire. Deep learning with r, ch. 5.4, 2018.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent, 2020.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39:176–179, 1943.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. *arXiv preprint arXiv:1307.1493*, 2013.

H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.