Regularisation

Hamid Dehghani
School of Computer Science
Birmingham

September 2020

Slides adapted from Iain Styles, School of Computer Science

UNIVERSITY OF BIRMINGHAM

Intended Learning Outcome

- Understand the effect of regularisation on regression problems

Regularised Loss Functions

Generalised regularised loss function

$$\mathcal{L}(\mathbf{w}) = \mathcal{L}_{\mathrm{err}}(\mathbf{w}) + \lambda R(\mathbf{w}) \tag{1}$$

Model-data mismatch plus “penalty” term

For the specific case of LSE + Ł$_2$ penalty (Gaussian prior)

$$\mathcal{L}(\mathbf{w}) = (\mathbf{y} - \mathbf{\Phi w})^{\mathrm{T}} (\mathbf{y} - \mathbf{\Phi w}) + \lambda \mathbf{w}^{\mathrm{T}} \mathbf{w} \tag{2}$$

Minimising loss requires both terms to be minimised

λ controls the width of the Gaussian prior and hence the balance between model fitting and parameter shrinkage.

Solving Regularised Least Squares

One reason L_2 is common is its closed-form analytic solution

Differentiate loss function with respect to **w** and set to zero to minimise

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \qquad \left(\mathbf{y} - \mathbf{\Phi}\mathbf{w}\right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{\Phi}\mathbf{w}\right) + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w} \\ \frac{\partial \mathcal{L}_{\mathrm{LSE}}(\mathbf{w})}{\partial \mathbf{w}} &= \qquad -2\mathbf{\Phi}^{\mathrm{T}} \left(\mathbf{y} - \mathbf{\Phi}\mathbf{w}\right) + 2\lambda\mathbf{w} \end{aligned}$$

Set to zero to minimise

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} - \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} - \lambda\mathbf{I}\right)\mathbf{w}^{*} = 0$$

Modified normal equations, can be solved in the same way

UNIVERSITY OF
BIRMINGHAM

Terminology

Ridge regression.

L_2 regularisation, because $\sum_i w_i^2$ is the L_2 norm of $\mathbf{w}$, written as $||\mathbf{w}||_2^2$.

Weight decay, because it pushes weights towards zero.

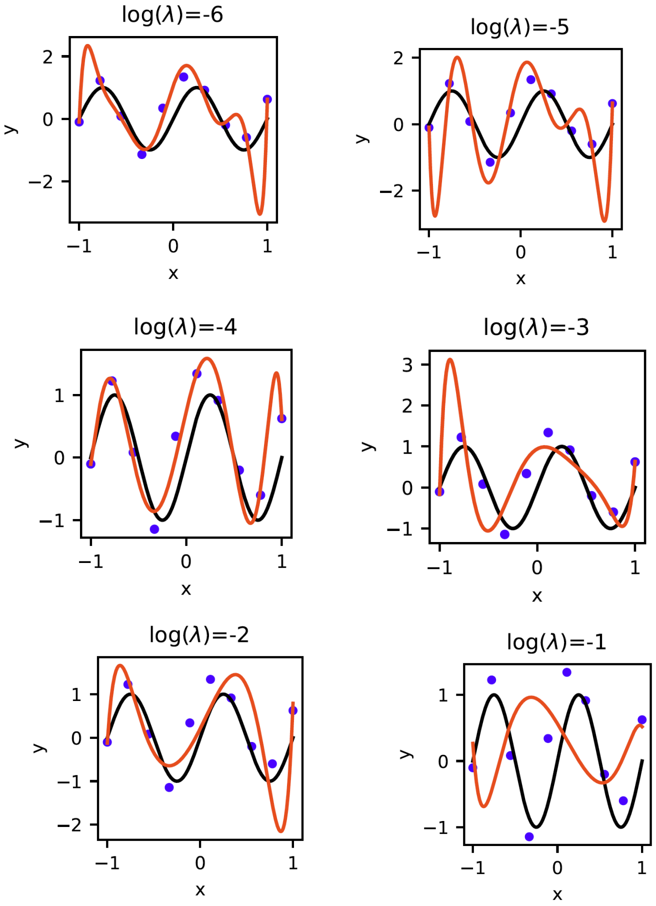
Tikhonov regularisation, of which it is a special case. Tikhonov regularisation uses $R(\mathbf{p}) = ||\mathbf{\Gamma w}||$. Here, we have $\mathbf{\Gamma} = \lambda \mathbf{I}$.

Regularisation in Practice

L_2 regression used to prevent overfitting

Prevents model weight growing large to fit noise.

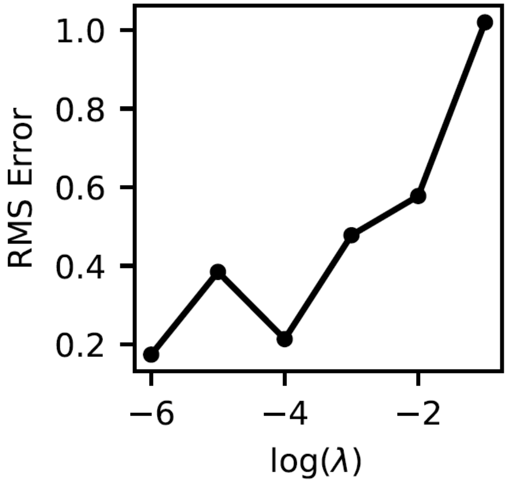
Large values of λ “smooth” fluctuations



UNIVERSITY OF
BIRMINGHAM

Effect of Regularisation

- Regularisation increases training error but improves generalisation
- Penalty term reduce model weights



$\log_{10}\lambda$	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
-6	1.06	8.31	-17.92	-76.20	76.16	250.58	-112.92	-350.94	53.88	168.61
-5	1.67	5.77	-41.44	-29.12	212.84	31.81	-356.24	1.30	183.44	-9.40
-4	0.74	6.51	-5.21	-33.75	1.46	33.96	20.81	13.89	-17.55	-20.25
-3	0.94	1.29	-9.30	5.42	15.29	-21.02	5.50	-4.36	-12.20	19.08
-2	0.20	4.26	2.26	-10.00	-5.63	-4.88	-0.83	2.60	4.33	8.47
-1	0.56	-2.12	-1.35	4.00	-0.63	1.44	0.50	-0.84	1.30	-2.35

Limitations

L_2 regularisation tends to “smooth” the fit

Popular for image denoising

But does not work well for data that really does have “fast” fluctuations.

Other choices can overcome this

Most general form: $R(\mathbf{w}) = \sum_i |w_i|^p$

$p = 1$ is very common - L_1 or *lasso* regularisation

L_1 Regularisation

Also known as the Lasso methods

$R(\mathbf{w}) = \sum_i |w_i|$

Tends to promote *sparsity* in model parameters

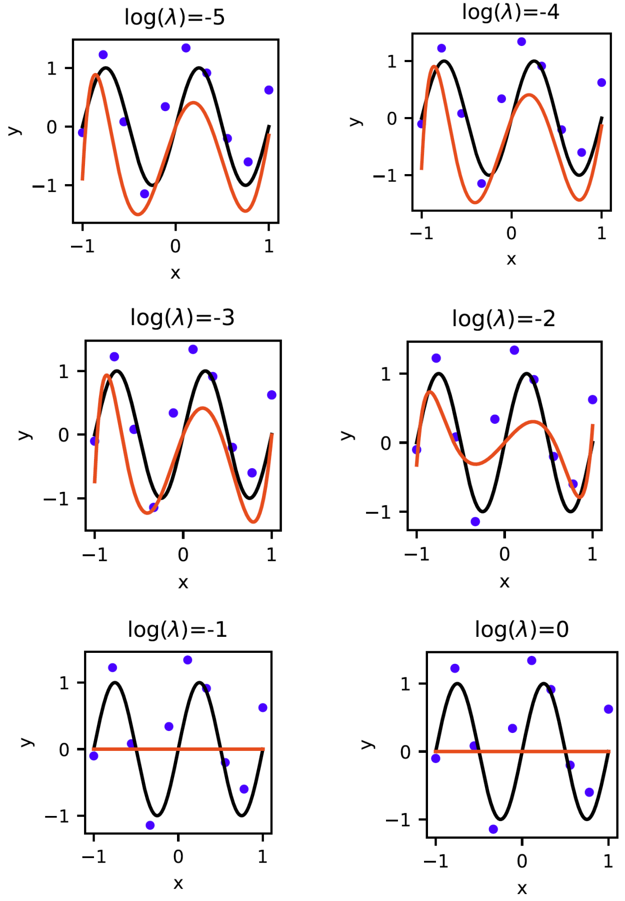
No closed form solution

Turn to `scikit-learn` for implementation

L_1 Regularisation

$\log_{10}\lambda$	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
-5	0.00	3.59	-0.00	-15.72	0.00	11.78	0.00	1.96	0.00	-1.60
-4	0.00	3.54	0.00	-15.27	0.00	11.05	0.00	1.93	0.00	-1.24
-3	0.00	3.10	-0.00	-12.02	-0.00	5.44	-0.00	3.49	-0.00	0.00
-2	0.00	0.95	-0.00	-3.76	-0.00	-0.00	-0.00	0.00	-0.00	2.74
-1	0.00	-0.05	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
0	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00

- Progressice sparsification of model parameters



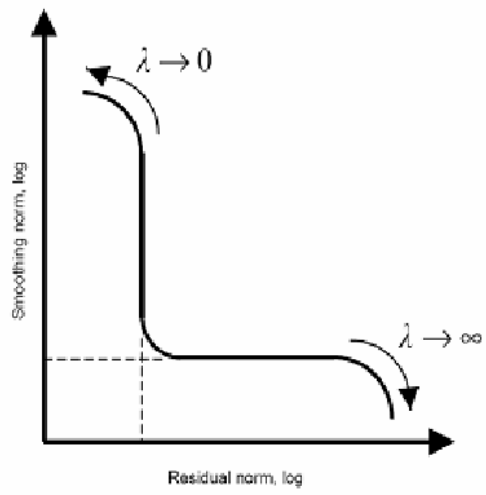
Regularisation in Practice

The correct regulariser and choice of λ is very problem dependent

Smooth or sparse solution?

Rigorous cross-validation often needed to selection λ

Enables complex models to be used and then controlled


$\lambda \to 0$
$\lambda \to \infty$
Smoothing norm, log
Residual norm, log