# Ridge Regression

Mohammad Emtiyaz Khan
EPFL

Oct 1, 2015

# Motivation

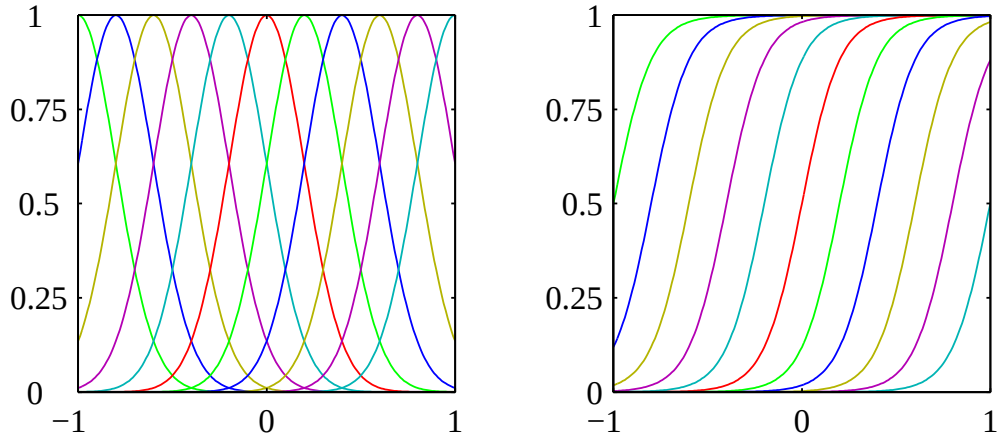
Linear models can be too limited and usually underfit. One way is to use nonlinear basis functions instead.

# Nonlinear basis functions

In general, we can use *basis functions* that are (known and fixed) nonlinear transformations applied to input variables.

Given a $D$ dimensional input vector $\mathbf{x}$, let us say that we have $M$ basis functions $\phi_j(\mathbf{x})$, indexed by $j$.

Example of basis functions: Polynomial, Splines (see Wikipedia), Gaussian, sigmoid, Fourier, Wavelets.




Taken from Bishop, Chapter 3

# Linear basis function model

The model is given as follows:

$$y_n = \beta_0 + \sum_{j=1}^{M} \beta_j \phi_j(\mathbf{x}_n) = \widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T \boldsymbol{\beta}$$

where $\widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T = [1,\, \phi_1(\mathbf{x}_n),\, \phi_2(\mathbf{x}_n), \ldots, \phi_M(\mathbf{x}_n)]$

This model *is* linear in $\boldsymbol{\beta}$ but nonlinear in $\mathbf{x}$. Note that the dimensionality is now $M$, not $D$.

Defining matrix $\widetilde{\boldsymbol{\Phi}}$ with $\widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T$ as rows,

$$\widetilde{\boldsymbol{\Phi}} := \begin{bmatrix} 1 & \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \ldots & \phi_M(\mathbf{x}_1) \\ 1 & \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \ldots & \phi_M(\mathbf{x}_2) \\ 1 & \phi_1(\mathbf{x}_3) & \phi_2(\mathbf{x}_3) & \ldots & \phi_M(\mathbf{x}_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \ldots & \phi_M(\mathbf{x}_N) \end{bmatrix}$$

The least square solution is

$$\boldsymbol{\beta}^*_{lse} = (\widetilde{\boldsymbol{\Phi}}^T \widetilde{\boldsymbol{\Phi}})^{-1} \widetilde{\boldsymbol{\Phi}}^T \mathbf{y}$$

# Two Issues

The model can potentially overfit. The Gram matrix can be ill-conditioned.

# Regularization and ridge regression

Through regularization, we can penalize complex models and favor simpler ones:

$$\min_{\beta} \quad \mathcal{L}(\boldsymbol{\beta}) + \frac{\lambda}{2N} \sum_{j=1}^{M} \beta_j^2$$

The second term is a regularizer.

The main point is that an input variable weighted by a small $\beta_j$ will have less influence on the output.

Note that $\beta_0$ is not penalized. Setting $\lambda = 0$, we get back to least-squares when $\mathcal{L}$ is MSE.

# Ridge regression

When $\mathcal{L}$ is MSE, this is called the ridge regression:

$$\min_{\beta} \quad \frac{1}{2N} \sum_{n=1}^{N} [y_n - \widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T \boldsymbol{\beta}]^2 \quad + \quad \frac{\lambda}{2N} \sum_{j=1}^{M} \beta_j^2$$

Differentiating and setting to zero:

$$\boldsymbol{\beta}_{ridge} = (\widetilde{\boldsymbol{\Phi}}^T \widetilde{\boldsymbol{\Phi}} + \lambda \mathbf{I}_M)^{-1} \widetilde{\boldsymbol{\Phi}}^T \mathbf{y}$$

## Ridge regression to fight ill-conditioning

The eigenvalues of $(\widetilde{\mathbf{\Phi}}^T\widetilde{\mathbf{\Phi}} + \lambda\mathbf{I}_M)$ is at least $\lambda$. This is also referred to as *lifting the eigenvalues*.
Proof: Write the eigenvalue decomposition of $\widetilde{\mathbf{\Phi}}^T\widetilde{\mathbf{\Phi}}$ as $\mathbf{QSQ}^T$ and use the fact that $\mathbf{QQ}^T = \mathbf{I}_M$.

Therefore ridge regression improves the condition number of the Gram matrix.

## Ridge regression as MAP estimator

Assume $\beta_0 = 0$ for this discussion. Recall that least-squares can be interpreted as the maximum likelihood estimator.

$$\boldsymbol{\beta}_{lse} = \arg\max_{\beta} \quad \log\left[\prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{x}_n^T\boldsymbol{\beta}, \sigma^2)\right]$$

Ridge regression has a very similar interpretation:

$$\boldsymbol{\beta}_{ridge} = \arg\max_{\beta} \quad \log\left[\prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{x}_n^T\boldsymbol{\beta}, \boldsymbol{\lambda}) \times \mathcal{N}(\boldsymbol{\beta}|0, \mathbf{I})\right]$$

This is called a Maximum-a-posteriori (MAP) estimate.

# Additional Notes

## Other type of regularization

Popular methods such as shrinkage, weight decay (in the context of neural networks), early stopping are all different forms of regularization.

Another view of regularization: the regularized optimization problem is similar to the following constrained optimization (for some $\tau > 0$).

$$\min_{\beta} \quad \frac{1}{2N} \sum_{n=1}^{N} (y_n - \widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T \boldsymbol{\beta})^2, \quad \text{such that } \boldsymbol{\beta}^T \boldsymbol{\beta} \leq \tau \tag{1}$$

The following picture illustrates this.

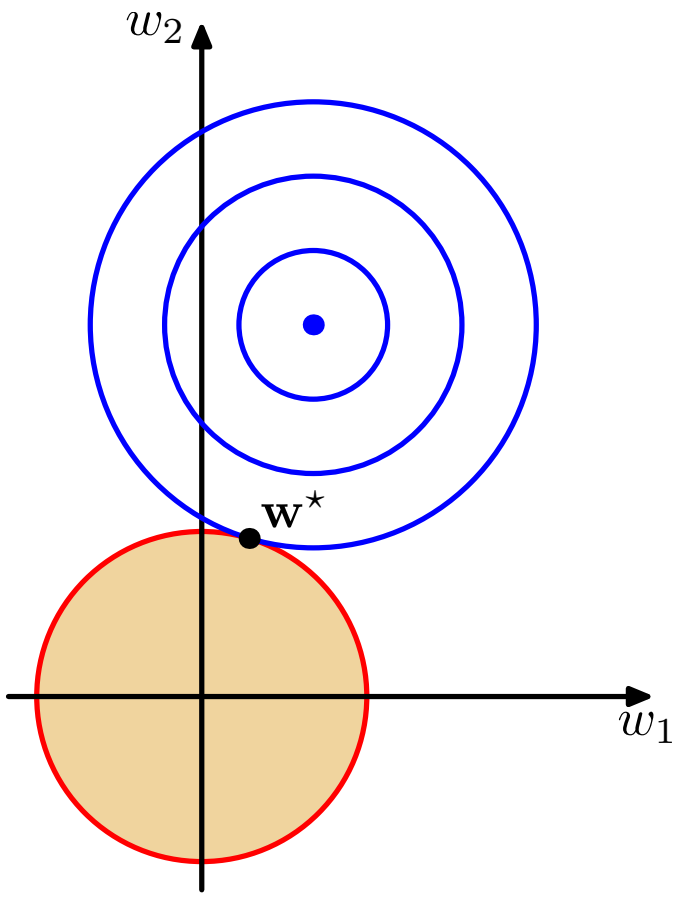



Figure 1: Geometric interpretation of Ridge Regression.

Another popular regularization is the $L_1$ regularizer (also known as lasso)

$$\min_{\beta} \quad \frac{1}{2N}\sum_{n=1}^{N}(y_n - \widetilde{\boldsymbol{\phi}}(\mathbf{x}_n)^T\boldsymbol{\beta})^2, \quad \text{such that} \quad \sum_{i=1}^{M}|\beta_i| \leq \tau \tag{2}$$

This forces some of the elements of $\boldsymbol{\beta}$ to be strictly 0 and therefore forces sparsity in the model (some features are not used since their coefficients are zero).

- Why does $L_1$ regularizer forces sparsity? Hint: Draw the picture similar to above for lasso and look at the optimal solution.
- Why is it good to have sparsity in the model? Is it going to be better than least-squares? When and why?

## MAP estimate and the Bayes rule

In general, a MAP estimate maximizes the product of the likelihood *and* the prior, as opposed to a ML estimate that maximizes only the likelihood.

$$\boldsymbol{\beta}_{lik} = \arg\max_{\beta} p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\lambda}) \tag{3}$$

$$\boldsymbol{\beta}_{map} = \arg\max_{\beta} p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\lambda})p(\boldsymbol{\beta}|\boldsymbol{\theta}) \tag{4}$$

Here $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\lambda})$ defines the *likelihood* of observing the data $\mathbf{y}$ given $\mathbf{X}, \boldsymbol{\beta}$ and some likelihood parameters $\lambda$.

Similarly, $p(\boldsymbol{\beta}|\boldsymbol{\lambda})$ is the *prior* distribution. This incorporates our prior knowledge about $\boldsymbol{\beta}$. Using the Bayes rule,

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A) \tag{5}$$

we can rewrite the MAP estimate, as follows:

$$\boldsymbol{\beta}_{map} = \arg\max_{\beta} p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{6}$$

Therefore, the MAP estimate is the maximum of the posterior distribution, which explains the name.

## To do

1. Read Section 3.1 and 3.3 of Bishop (3.1.1 might be confusing and can be skipped at the first reading).

2. Implement ridge regression and understand how it solves the two problems.

3. (Difficult) Derive the update rule for Ridge regression using the Gaussian formula (Eq. 2.113-2.117 in Bishop's book). This is detailed in Section 3.3.1 of Bishop's book.