These are my technical notes on various topics that I've come across and needed to understand better. Last updated: MG, December 25, 2017

Contents

1 Duality and KKT **3**

1.1 Primal problem . . . 3

1.2 Lagrange dual problem . . . 3

1.2.1 Weak duality . . . 3

1.2.2 Strong duality . . . 4

1.3 Otimality conditions . . . 4

1.3.1 KKT conditions . . . 5

2 Support vector machines and regression **6**

2.1 Support vector machines . . . 6

2.1.1 Introduction . . . 6

2.1.2 Optimal margin hyperplanes . . . 6

2.1.3 Soft margin SVM . . . 7

2.2 Support vector regression . . . 9

3 Learning functions in RKHS **11**

3.1 Learning for scalar output with squarred norm regularization . . . 11

3.1.1 Kernel ridge regression . . . 12

3.2 Learning for vector-output . . . 12

3.2.1 Representer theorem for vector-output problems . . . 12

3.2.2 Solving with least squares and squarred regularizer . . . 13

3.2.3 Kernel multioutput ridge regression . . . 14

3.2.4 Output kernel vs. covariance matrix for Gaussian errors and Gaussian priors on the parameters . . . 15

4 Multiple kernel learning **17**

4.1 Kernel normalization . . . 17

4.2 Solving MKL . . . 17

4.2.1 Only need 1 regularization hyper-parameter . . . 17

4.2.2 Alternating minimisation . . . 18

4.2.3 Change of variables . . . 19

5 Double regularization with single tunning parameter **23**

6 Grids for hyperparameters **24**

6.1 Lasso example . . . 24

7 Trace norm regularization **25**

8 Subgradient and subdifferential **27**

8.1 Absolute value $|x|$ . . . 27

8.2 ℓ_2 norm $||\mathbf{x}||_2$ . . . 27

8.3 Generalised ℓ_2 norm . . . 28

9 Training with noise is equivalent to Tikhonov regularisation **29**

10 Early stopping equivalence to Tikhonov regularization **31**

11 Some useful inequatlities **33**

12 Learning theory **38**

12.1 Generalization . . . 38

12.2 Bounds . . . 38

12.3 Empirical error bound . . . 39

12.4 Uniform deviations . . . 40

12.5 Estimation error . . . 41

12.6 Uncountable hypotheses set . . . 41

12.7 Rademacher complexity . . . 41

12.8 Rademacher bounds . . . 42

12.9 Regression . . . 42

13 Basics of prediction theory **44**

13.1 Prediction as linear projection . . . 44

13.2 Link to ordinary least squares . . . 45

13.3 Vector outputs . . . 46

14 Linear regression basics **47**

14.1 Ordinary least squares . . . 47

14.2 Ridge regression . . . 49

Index **52**

1 Duality and KKT

1.1 Primal problem

The primal optimisation problem (not necessarily convex) for $\mathbf{x} \in \mathbb{R}^n$ is

$$\begin{aligned} &\text{minimize} && f_0(\mathbf{x}) \\ &\text{subject to} && f_i(\mathbf{x}) \le 0,\ i = 1, \dots, m \\ & && h_i(\mathbf{x}) = 0,\ i = 1, \dots, p \end{aligned} \tag{1.1}$$

We assume its domain $\mathcal{D} = \cap_i^m\, dom\, f_i \cap \cap_j^p\, dom\, h_j$ is nonempty, and we denote the optimal solution by $p^* = f_0(\mathbf{x}^*)$.

We define the *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (1.1)

$$L(\mathbf{x}, \lambda, \mu) = f_0(\mathbf{x}) + \sum_i^m \lambda_i f_i(\mathbf{x}) + \sum_i^p \mu_i h_i(\mathbf{x}), \tag{1.2}$$

where λ and μ are the Lagrange multiplier vectors.

1.2 Lagrange dual problem

We define the *Lagrange dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (1.1)

$$g(\lambda, \mu) = \inf_{x \in \mathcal{D}} L(\mathbf{x}, \lambda, \mu) = \inf_{x \in \mathcal{D}} \Big(f_0(\mathbf{x}) + \sum_i^m \lambda_i f_i(\mathbf{x}) + \sum_i^p \mu_i h_i(\mathbf{x}) \Big) \tag{1.3}$$

which is always a concave function (irrespective of the problem (1.1) being convex or not).

The dual function gives a lower bound on the optimal value p^* of problem (1.1) (see [1] sec.5.1.3 for proof).

$$g(\lambda, \mu) \le p^*, \quad \forall \lambda \succeq 0,\, \mu \tag{1.4}$$

When $g(\lambda, \mu) = -\infty$ this is not very useful. So lagrange multipliers such that $\lambda \succeq 0$ and $g(\lambda, \mu) > -\infty$ are called *dual feasable*.

The *Lagrange dual problem* answers the question "What is the best lower bound that we can get from the Lagrange dual function?"

$$\begin{aligned} &\text{maximize} && g(\lambda, \mu) \\ &\text{subject to} && \lambda \succeq 0 \end{aligned} \tag{1.5}$$

This is a convex problem since we maximize a concave function and the constraint is convex. We denote the optimal solution with *optimal Lagrange multipliers* by $d^* = g(\lambda^*, \mu^*)$.

1.2.1 Weak duality

From eq. (1.4) if follows that

$$d^* \le p^* \tag{1.6}$$

The difference $p^* - d^*$ is called the *optimal duality gap* (always non-negative). It also follows that for $d^* = \infty$ we must have $p^* = \infty$ and therefore the primal problem is infeasable; for $p^* = -\infty$ we must have $d^* = -\infty$ and therefore the dual problem is infeasable. The optimal dual bound can sometimes be used as proxy for the optimal solution of the primal problem if it is difficult to solve since the dual is always convex and may be easier.

1.2.2 Strong duality

We say that *strong duality* holds if

$$d^* = p^* \tag{1.7}$$

Generally, the conditions on convex primal problems that have strong duality are called *constraint qualifications*.

For convex problems in the form

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m \\ & \mathbf{A}\mathbf{x} = b,\ i = 1, \ldots, p \end{array} \tag{1.8}$$

the *Slater's condition* (condition for strict feasibility)

$$f_i(\mathbf{x}) < 0,\ i = 1, \ldots, m, \qquad \mathbf{A}\mathbf{x} = b \tag{1.9}$$

ensures strong duality. But affine inequailty constraints do not have to hold with strict inequalities so if all the constraints are affine the Slater's conditions reduce to feasibility conditions.

1.3 Otimality conditions

If we can find dual feasable (λ, μ) we establish a proof (or certificate) that the primal optimal solution is $p^* \geq g(\lambda, \mu)$. In case of strong duality we can find arbitrarily good certificates.

If we find a primal feasable point $\mathbf{x}$ and (λ, μ) are dual feasable then

$$f_0(\mathbf{x}) - p^* \leq f_0(\mathbf{x}) - g(\lambda, \mu) \tag{1.10}$$

and we refer to the difference between the primal and dual objectives $\epsilon = f_0(\mathbf{x}) - g(\lambda, \mu)$ as the *duality gap* associated with these points (and say that the solutions are ϵ-suboptimal). The optimal solutions are always within the intervals specified by the primal and dual feasable points

$$p^* \in [g(\lambda, \mu), f_0(\mathbf{x})], \qquad d^* \in [f_0(\mathbf{x}), g(\lambda, \mu)] \tag{1.11}$$

If the duality gap ϵ is zero than the feasable points are optimal. The duality gap ϵ can be used in algorithms as a stopping criterion.

For the optimal points $\mathbf{x}^*, \lambda^*, \mu^*$ with strong duality we get

$$\begin{array}{rcll} f_0(\mathbf{x}^*) & = & g(\lambda^*, \mu^*) & \text{(strong duality)} \\ & = & \inf_x \Big(f_0(\mathbf{x}) + \sum_i^m \lambda_i^* f_i(\mathbf{x}) + \sum_i^p \mu_i^* h_i(\mathbf{x}) \Big) & \text{(dual function)} \\ & \leq & \Big(f_0(\mathbf{x}^*) + \sum_i^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_i^p \mu_i^* h_i(\mathbf{x}^*) \Big) & \text{(infemum)} \\ & \leq & f_0(\mathbf{x}^*) & \big(\lambda_i^* \geq 0, f_i(\mathbf{x}^*) \leq 0\big) \end{array} \tag{1.12}$$

For this to be valid, the last two lines need to hold as equality and in result we must have

$$\sum_i^m \lambda_i^* f_i(\mathbf{x}^*) = 0 \tag{1.13}$$

Since each term in the sum is non-negative we must have

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, m \tag{1.14}$$

These are known as the *complementary slackness* conditions which must hold for any primal and dual optimal points in case of strong duality. Roughly speaking, they mean that the ith Lagrange multiplier is zero unless the ith constraint is active at the optimum. More formally,

$$\begin{aligned} \lambda_i^* > 0 &\implies f_i(\mathbf{x}^*) = 0 \\ f_i(\mathbf{x}^*) < 0 &\implies \lambda_i^* = 0 \end{aligned} \tag{1.15}$$

1.3.1 KKT conditions

We assume that all functions in the primal problem (1.1) are differentiable (though not necessarily convex). At optimal points $\mathbf{x}^*, \lambda^*, \mu^*$ with zero duality gap the optimal $\mathbf{x}*$ minimizes the Lagrangian $L(\mathbf{x}, \lambda^*, \mu)^*$ so its gradient at $\mathbf{x}*$ must vanish

$$\nabla L(\mathbf{x}^*, \lambda^*, \mu)^* = \nabla f_0(\mathbf{x}^*) + \sum_i^m \lambda_i \nabla f_i(\mathbf{x}^*) + \sum_i^p \mu_i \nabla h_i(\mathbf{x}^*) = 0 \tag{1.16}$$

If we put this together with the primal and dual feasibility conditions and the complementary slackness conditions we get the *Karush-Kuhn-Tucker* (KKT) conditions which must hold for any optimisation problem with strong duality (and differentiable objective and constraints).

$$\begin{array}{rcll} f_i(\mathbf{x}^*) & \leq & 0, \quad i = 1, \ldots, m & \text{(primal feasibility)} \\ h_i(\mathbf{x}^*) & = & 0, \quad i = 1, \ldots, p & \text{(primal feasibility)} \\ \lambda_i^* & \geq & 0, \quad i = 1, \ldots, m & \text{(dual feasibility)} \\ \lambda_i f_i(\mathbf{x}^*) & = & 0, \quad i = 1, \ldots, m & \text{(complementary slackness)} \\ \nabla f_0(\mathbf{x}^*) + \sum_i^m \lambda_i \nabla f_i(\mathbf{x}^*) + \sum_i^p \mu_i \nabla h_i(\mathbf{x}^*) & = & 0 & \text{(vanishing gradient of Lagrangian)} \end{array} \tag{1.17}$$

For convex primal problmes the KKT conditions are also sufficient for any points that satisfy them to be the primal and dual optima with zero duality gap. If a convex optimisation problem satisfies Slater's condition, than the KKT conditions provide necessary and sufficient conditions for optimality.

We can also use this to solve the primal problem by solving first the dual instead and then recovering the optimal primal by minimising the lagrangian at the optimal Lagrange multipliers.

References

[1] S. Boyd and L. Vendenberghe, Convex Optimization. Cambridge University Press, 2004.

2 Support vector machines and regression

2.1 Support vector machines

This is based mainly on [1].

2.1.1 Introduction

We have got a set of patterns $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{H}$ in a dot product space $\mathcal{H}$. Any hyperplane in this space can be written as

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0, \qquad \mathbf{w}, \mathbf{x} \in \mathcal{H}, b \in \mathbb{R} \tag{2.1}$$

Here, $\mathbf{w}$ is a vector orthogonal to the hyperplane (pick two vectors on the hyperplane $\mathbf{x}_1$ and $\mathbf{x}_2$ and obseve that $\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = \langle \mathbf{w}, \mathbf{x}_2 \rangle + b = 0$ so that $\langle \mathbf{w}, \mathbf{x}_1 \rangle - \langle \mathbf{w}, \mathbf{x}_2 \rangle = 0$ and $\langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 0$ where $(\mathbf{x}_1 - \mathbf{x}_2)$ is a vector within the hyperplane).

The length r of any vector $\mathbf{x}$ along $\mathbf{w}$ is given by $r = \langle \mathbf{w}, \mathbf{x} \rangle / ||\mathbf{w}||_2^2$. (To get this observe that $\langle \mathbf{w}, \mathbf{x} - r\mathbf{w} \rangle = 0$ where $r\mathbf{w}$ is the orthogonal projection of $\mathbf{x}$ on $\mathbf{w}$.) All points on the plane eq. (2.1) have the same length along $\mathbf{w}$, they all project onto the same point on the line spanned by $\mathbf{w}$.

The distance d of a point $\mathbf{x}$ from the plane (2.1) is given by $d = |\langle \mathbf{w}, \mathbf{x} \rangle + b| / ||\mathbf{w}||_2^2$. (To get this, observe that $\mathbf{x} = \mathbf{x}_p + d\mathbf{w}$, where $\mathbf{x}_p$ is the orthogonal projection of $\mathbf{x}$ onto the plane. Premultiplying both sides by $\mathbf{w}'$ and adding b yields $\mathbf{w}'\mathbf{x} + b = \mathbf{w}'\mathbf{x}_p + b + d\mathbf{w}'\mathbf{w}$, where $\mathbf{w}'\mathbf{x}_p + b = 0$ from eq. (2.1).) If we multiply the $\mathbf{w}$ and b by the same constant the plane defined by eq. (2.1) and all the distance calculations do not change.

We call the pair $\mathbf{w}, b$ scaled so that

$$\min_{i=1,\dots,n} |\langle \mathbf{w}, \mathbf{x} \rangle + b| / ||\mathbf{w}||_2^2 = 1 \tag{2.2}$$

the *cannonical form* of the hyperplane with respect to $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{H}$. Basically, this says we scale $\mathbf{w}$ and b so that the closest point to the hyperplane has distance from it $d = 1/||\mathbf{w}||_2^2$. We call this distance the *margin*. There are technically still 2 such hyperplanes $\mathbf{w}, b$ and $-\mathbf{w}, -b$ which conincide but have different directions.

If in addition to the patterns we have got the class labels $\{\mathbf{y}_i\}_{i=1}^n \in \{\pm 1\}$ we wish to distniguish these two cases to help us classify the patterns by a decision function

$$f(\mathbf{x}) = sgn\big(\langle \mathbf{w}, \mathbf{x} \rangle + b\big) \tag{2.3}$$

In a classification problem, the margin from the separating hyperplane shall be as big as possible. As stated above the margin is $d = 1/||\mathbf{w}||_2^2$ and therfore we can maximize it by minimising $||\mathbf{w}||_2^2$.

2.1.2 Optimal margin hyperplanes

For a classificatin problem for a set of examples $\{\mathbf{x}_i, y_i\}_{i=1}^n \in \mathcal{H} \times \{\pm 1\}$ we wish to find a decision function (2.3) satisfying $f(\mathbf{x_i}) = y_i$. If such a function exists (and using the canonical form of eq. (2.2)) we have

$$y_i\big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) \geq 1, \qquad i = 1, \dots, n \tag{2.4}$$

Note that this actually helps us to distinquish the two canonical forms $\mathbf{w}, b$ and $-\mathbf{w}, -b$ because only one of those will satisfy eq. (2.4).

In result, a seperating hyperplane that generalizes well can be found by solving the following optimisation problem

$$\begin{aligned} \text{minimize} \quad & \tau(\mathbf{w}) := 1/2\, ||\mathbf{w}||_2^2 \\ \text{subject to} \quad & y_i\big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) \geq 1 \qquad \forall i = 1, \dots, n \end{aligned} \tag{2.5}$$

We formulate the Lagrangian of the problem (2.5) as

$$L(\mathbf{w}, b, \alpha) = 1/2\, ||\mathbf{w}||_2^2 - \sum_i^n \alpha_i \Big(y_i \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \big) - 1 \Big) \tag{2.6}$$

and the corresponding dual function

$$g(\alpha) = \inf_{w,b} 1/2\, ||\mathbf{w}||_2^2 - \sum_i^n \alpha_i \Big(y_i \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \big) - 1 \Big) \tag{2.7}$$

.

Minimizing the Lagrangian with respect $\mathbf{w}$ and b yields

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i^n \alpha_i y_i \mathbf{x}_i = 0 \qquad \Longrightarrow \qquad \mathbf{w} = \sum_i^n \alpha_i y_i \mathbf{x}_i \tag{2.8}$$

and

$$\frac{\partial L}{\partial b} = \sum_i^n \alpha_i y_i = 0 \tag{2.9}$$

and therefore the SVM dual problem is (a convex problem)

$$\begin{aligned} \text{maximize} \quad & \varphi(\alpha) := \sum_i^n \alpha_i - 1/2 \sum_{i,j}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle && (2.10) \\ \text{subject to} \quad & \alpha_i \geq 0 && (2.11) \\ & \sum_i^n \alpha_i y_i = 0 && (2.12) \end{aligned}$$

The decision function (2.3) can now be written as

$$f(\mathbf{x}) = sgn\big(\sum_i^n \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \big), \tag{2.13}$$

where we can replace the inner product by a kernel function

$$f(\mathbf{x}) = sgn\big(\sum_i^n \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \big), \tag{2.14}$$

From the KKT conditions we have for every i that $\alpha_i \Big(y_i \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \big) - 1 \Big) = \alpha_i \Big(y_i \big(\sum_j^n \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) + b \big) - 1 \Big) = 0$ and therefore once solved for α from (2.10) we can solve for b by for example (though other options may be more advantageous see [1] section 7.4) averaging over all $i \in \mathcal{S}$ where $\mathcal{S}$ is the set of support vectors for which $\alpha_i > 0$ (and observing that $1/y_i = y_i$)

$$b = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \Big(\sum_j^n \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) - y_i \Big) \tag{2.15}$$

2.1.3 Soft margin SVM

If the patterns are not separable eg. because of outlier etc. we may want to use a weaker constraint for the separation instead of (2.4)

$$y_i \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \big) \geq 1 - \xi_i, \quad \xi \geq 0, \qquad i = 1, \ldots, n \tag{2.16}$$

Note the link to the *hinge-loss* formulation

$$\xi_i = \max\{1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big), 0\} \tag{2.17}$$

Whenever pattern is on the correct side of the decision surface beyond the margin it does not incur any loss and it does not carry any information about the decision surface.

Clearly, if ξ_i could be arbitrarily large condition (2.16) would be always satisfied. To avoid this trivial solution we penalize their size in the objective function so that the primal problem of *soft-margin SVM* is

$$\begin{aligned}
\text{minimize} \quad & \tau(\mathbf{w}, \xi) := 1/2\,||\mathbf{w}||_2^2 + C\sum_i^n \xi_i, \qquad C > 0 \\
\text{subject to} \quad & y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n \\
& \xi_i \geq 0, \qquad \forall i = 1, \ldots, n
\end{aligned} \tag{2.18}$$

As before we formulate the Lagrangian of the problem (2.5) as

$$L(\mathbf{w}, b, \xi, \alpha, \nu) = 1/2\,||\mathbf{w}||_2^2 + C\sum_i^n \xi_i - \sum_i^n \alpha_i\Big(y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big) - 1 + \xi_i\Big) - \sum_i^n \nu_i\,\xi_i \tag{2.19}$$

and the corresponding dual function

$$g(\alpha, \nu) = \inf_{w,b}\ 1/2\,||\mathbf{w}||_2^2 - \sum_i^n \alpha_i\Big(y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big) - 1 + \xi_i\Big) + \sum_i^n (C - \nu_i)\,\xi_i \tag{2.20}$$

.

Minimizing the soft-margin SVM Lagrangian with respect $\mathbf{w}$, b and ξ yields

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i^n \alpha_i y_i \mathbf{x}_i = 0 \qquad \Longrightarrow \qquad \mathbf{w} = \sum_i^n \alpha_i y_i \mathbf{x}_i \tag{2.21}$$

$$\frac{\partial L}{\partial b} = \sum_i^n \alpha_i y_i = 0 \tag{2.22}$$

and

$$\frac{\partial L}{\partial \xi_i} = -\alpha_i + C - \nu_i = 0 \qquad \Longrightarrow \qquad \nu_i = C - \alpha_i \tag{2.23}$$

and because $\nu_i \geq 0$ we get $\alpha_i \leq C$.

The soft-margin SVM dual prolem is therefore

$$\text{maximize} \quad \varphi(\alpha) := \sum_i^n \alpha_i - 1/2\sum_{i,j}^n \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{2.24}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \tag{2.25}$$

$$\sum_i^n \alpha_i y_i = 0 \tag{2.26}$$

From the KKT conditions we have for every i that $\alpha_i\Big(y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big) - 1 + \xi_i\Big) = \alpha_i\Big(y_i\big(\sum_j^n \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) + b\big) - 1 + \xi_i\Big) = 0$ and therefore once solved for α from (2.10) we can solve for b by for example (though other options may be more advantageous see [1] section 7.4) averaging over all $i \in \mathcal{S}$ where $\mathcal{S}$ is the set of support vectors for which $0 < \alpha_i \leq C$ and $\xi_i = 0$ so that they lie on the margin (and observing that $1/y_i = y_i$)

$$b = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\Big(\sum_j^n \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) - y_i\Big) \tag{2.27}$$

2.2 Support vector regression

To bring the idea about soft-margin from SVMs to *support vector regression* (SVR) we use the *ϵ-insensitive loss*

$$|y - f(\mathbf{x})|_\epsilon = max\{|y - f(\mathbf{x})| - \epsilon, 0\} \tag{2.28}$$

.

In the SVR problem we search for a linear (affine) function

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \qquad \mathbf{w}, \mathbf{x} \in \mathcal{H}, b \in \mathbb{R} \tag{2.29}$$

based on a set of observations $\{\mathbf{x}_i, y_i\}_{i=1}^n \in \mathcal{H} \times \mathbb{R}$ such that it minimizes the risk (or test error)

$$R[f] := \int \ell(f, x, y) dP(x, y), \tag{2.30}$$

where $\ell(f, x, y)$ is a loss function (such as squared error) and P is the probability of the data generation process. Since we cannot minimize eq. (2.30) directly (we do not know the probability distribution P) we instead minimise the regularised empirical risk

$$R_{emp}[f] := \frac{C}{n} \sum_1^n \ell_{emp}(f, x, y) + \Omega(f), \quad C \geq 0 \tag{2.31}$$

In the case of SVR the regulariser is $\Omega(f) = 1/2||\mathbf{w}||_2^2$ and the empirical loss is the ϵ-insensitive loss of eq. (2.28) (note that it does not have to be the same as the theoretical loss used in eq. (2.30)).

Unlike in SVM we minimise directly the ℓ_2 norm of $\mathbf{w}$ instead of its negative so seemingly making the margin *large* but it somehow works out the right way (see [1] Figure 9.1).

The ϵ-insensitive loss creates a tube around the regression curve $f(\mathbf{x}_i) - \epsilon \leq y_i \leq f(\mathbf{x}_i) + \epsilon$ when the loss is zero. But similarly as in SVMs we may want to further relax this to account for outliers etc. by introducing slack variables $\xi_i \geq 0$ so that now we require that $f(\mathbf{x}_i) - \epsilon - \xi_i^* \leq y_i \leq f(\mathbf{x}_i) + \epsilon + \xi_i$, where for each data point we need to slack variables. As in SVMs, if ξ_i could be arbitrarily large we could get zero errors for all data points and therefore we control the size of the slack variables in the objective function of the primal SVR problem

$$\begin{array}{ll} \text{minimize} & \tau(\mathbf{w}, \xi) := 1/2\,||\mathbf{w}||_2^2 + C \sum_i^n (\xi_i + \xi_i^*), \qquad C > 0 \\ \text{subject to} & y_i - \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) \leq \epsilon + \xi_i \\ & \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{array} \tag{2.32}$$

We construct the Lagrangian of the problem (2.32) as

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \nu) \quad = \quad & 1/2\,||\mathbf{w}||_2^2 + C \sum_i^n (\xi_i + \xi_i^*) - \sum_i^n (\nu_i \, \xi_i + \nu_i^* \, \xi_i^*) \\ & + \sum_i^n \alpha_i \Big(y_i - \big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) - \epsilon - \xi_i \Big) \\ & + \sum_i^n \alpha_i^* \Big(\big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big) - y_i - \epsilon - \xi_i^* \Big) \end{aligned} \tag{2.33}$$

and the corresponding dual function

$$g(\alpha, \nu) = \inf_{w,b,\xi} L(\mathbf{w}, b, \xi, \alpha) \tag{2.34}$$

.

Minimizing the SVR Lagrangian with respect $\mathbf{w}$, b and ξ yields

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \sum_i^n (\alpha_i^* - \alpha_i)\mathbf{x}_i = 0 \qquad \Longrightarrow \qquad \mathbf{w} = \sum_i^n (\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{2.35}$$

$$\frac{\partial L}{\partial b} = \sum_i^n (\alpha_i^* - \alpha_i) = 0 \tag{2.36}$$

and

$$\frac{\partial L}{\partial \xi_i} = C - \nu_i - \alpha_i = 0 \qquad \Longrightarrow \qquad \nu_i = C - \alpha_i \tag{2.37}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \nu_i^* - \alpha_i^* = 0 \qquad \Longrightarrow \qquad \nu_i^* = C - \alpha_i^* \tag{2.38}$$

and because $\nu_i \geq 0$ we get $\alpha_i \leq C$.

Substituting these results into the dual function we get the *SVR dual problem*

$$\begin{aligned} \text{maximize} \qquad & \varphi(\alpha, \alpha^*) := -1/2 \sum_{ij}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & + \sum_i^n (\alpha_i - \alpha_i^*) y_i - \sum_i^n (\alpha_i - \alpha_i^*)\epsilon \\ \text{subject to} \qquad & 0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq C \\ & \sum_i^n (\alpha_i^* - \alpha_i) = 0 \end{aligned} \tag{2.39}$$

The linear function (2.29) can be expressed as

$$f(\mathbf{x}) = \sum_i^n (\alpha_i - \alpha_i^*)\langle \mathbf{x}_i, \mathbf{x} \rangle + b = \sum_i^n (\alpha_i - \alpha_i^*)\, k(\mathbf{x}_i, \mathbf{x}) + b \tag{2.40}$$

After solving for α we can get the solution for b by using the KKT conditions which in this case state that

$$\begin{aligned} \alpha_i \Big(y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - \epsilon - \xi_i \Big) &= 0 \\ \alpha_i^* \Big((\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i - \epsilon - \xi_i^* \Big) &= 0 \end{aligned} \tag{2.41}$$

and

$$\begin{aligned} \nu_i \xi_i = (C - \alpha_i)\xi_i &= 0 \\ \nu_i^* \xi_i^* = (C - \alpha_i^*)\xi_i^* &= 0 \end{aligned} \tag{2.42}$$

From which we can conclude:

- only examples with $C = \alpha_i^{(*)}$ can lie outside the ϵ-insensitive tube with $\xi^{(*)} > 0$
- for $0 < \alpha_i^{(*)} \leq C$ we must have $\xi_i^* = 0$
- $\alpha_i \, \alpha_i^* = 0$, that is they cannot be simultaneously both non-zero (but they can both be zeros)

Therefore we can use the support vectors $i \in \mathcal{S}$ for which $0 < \alpha_i^{(*)} \leq C$ to get b by from (2.41).

References

[1] B. Schlkopf and A. J. Smola, Learning with kernels. The MIT Press, 2002.

3 Learning functions in RKHS

Warning: Tihs does not give the full details on the *RKHS* theory and uses some basics facts without explaining.

3.1 Learning for scalar output with squarred norm regularization

To warm up to the problem, I begin with the simple scalar output problem with squarred norm regularizer. The more complete and general theory is in section 3.2.

From the *representer theorem*, the minimisation problem of a functional

$$\min_{f \in \mathcal{H}} J(f) := \sum_{i}^{n} \big(y_i - f(\mathbf{x}_i)\big)^2 + \lambda ||f||_{\mathcal{H}}^2 \tag{3.1}$$

has a solution f^* which admits a representation

$$f^*(\mathbf{x}) = \sum_{i}^{n} c_i \, k(\mathbf{x}_i, \mathbf{x}), \tag{3.2}$$

where n is the number of instances, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function and $\mathbf{c}$ is a vector of the parameters to be learned.

Introducing the *Gram matrix* $\mathbf{K}$ with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and the n-long vectors $\mathbf{y}$ and $\mathbf{c}$ problem (3.1) can be rewritten as a finite-dimensional optimisation

$$\min_{\mathbf{c}} J(\mathbf{c}) := ||\mathbf{y} - \mathbf{Kc}||_2^2 + \lambda \mathbf{c}'\mathbf{Kc} \tag{3.3}$$

Proof: Using the kernel *reproducing property* $\langle k(\mathbf{x}_i, .), k(\mathbf{x}_j, .)\rangle_{\mathcal{H}} = k(\mathbf{x}_i, \mathbf{x}_j)$ and the linearity of inner product

$$\begin{aligned}
J(f) &= \sum_{i}^{n} \big(y_i - f(\mathbf{x}_i)\big)^2 + \lambda ||f||_{\mathcal{H}}^2 = \\
&= \sum_{i}^{n} \big(y_i - \sum_{j}^{n} c_j \, k(\mathbf{x}_i, \mathbf{x}_j)\big)^2 + \lambda \langle \sum_{i}^{n} c_i \, k(\mathbf{x}_i, .), \sum_{j}^{n} c_j \, k(\mathbf{x}_j, .)\rangle_{\mathcal{H}} \\
&= \sum_{i}^{n} y_i^2 - 2 \sum_{i}^{n} \sum_{j}^{n} y_i c_j \, k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i}^{n} \sum_{j}^{n} c_j \, k(\mathbf{x}_i, \mathbf{x}_j) \sum_{l}^{n} c_l \, k(\mathbf{x}_i, \mathbf{x}_l) + \lambda \sum_{ij}^{n} c_i c_j \langle k(\mathbf{x}_i, .), k(\mathbf{x}_j, .)\rangle_{\mathcal{H}} \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{Kc} + \mathbf{c}'\mathbf{K}'\mathbf{Kc} + \lambda \mathbf{c}'\mathbf{Kc} \qquad \text{(see math cheat-sheet)} \\
&= ||\mathbf{y} - \mathbf{Kc}||_2^2 + \lambda \mathbf{c}'\mathbf{Kc},
\end{aligned}$$

We differentiate and equate to zero

$$\frac{\partial J}{\partial \mathbf{c}} = -2\mathbf{K}'\mathbf{y} + 2\mathbf{K}'\mathbf{Kc} + 2\lambda \mathbf{K}'\mathbf{c} = 0 \tag{3.4}$$

and get a closed form solution for the learned parameters

$$\mathbf{c} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y} \tag{3.5}$$

The minimizier (3.2) then is

$$f(\mathbf{x}) = \mathbf{k}(\mathbf{x})'\mathbf{c} = \mathbf{k}(\mathbf{x})'(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}, \tag{3.6}$$

where the vector $\mathbf{k}(\mathbf{x})$ has elements $\mathbf{k}(\mathbf{x})_i = k(\mathbf{x}_i, \mathbf{x})$.

3.1.1 Kernel ridge regression

We can show the equivalence of the above to *ridge regression*.

We choose a simple *linear kernel* $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. We get for $f(\mathbf{x}_j) = \sum_i^n c_i \, k(\mathbf{x}_i, \mathbf{x}_j) = \sum_i^n c_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \sum_i^n c_i \mathbf{x}_i, \mathbf{x}_j \rangle$. By putting $\mathbf{w} = \sum_i^n c_i \mathbf{x}_i$ we get $f(\mathbf{x}_j) = \langle \mathbf{w}, \mathbf{x}_j \rangle = \mathbf{x}_j' \mathbf{w}$.

The regularizer $||f||_{\mathcal{H}}^2 = \langle \sum_i^n c_i \, k(\mathbf{x}_i, .), \sum_i^n c_j \, k(\mathbf{x}_j, .) \rangle_{\mathcal{H}} = \sum_{ij} c_i c_j \langle k(\mathbf{x}_i, .), k(\mathbf{x}_j, .) \rangle_{\mathcal{H}} = \sum_{ij} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{ij} c_i c_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \sum_i c_i \mathbf{x}_i, \sum_j c_k \mathbf{x}_j \rangle = \langle \mathbf{w}, \mathbf{w} \rangle = ||\mathbf{w}||^2$.

Using these substitutions, we can rewrite problem (3.1) as

$$J := \sum_i^n \left(y_i - \mathbf{x}_j' \mathbf{w} \right)^2 + \lambda ||\mathbf{w}||^2, \tag{3.7}$$

where $\mathbf{w} = \sum_i^n c_i \mathbf{x}_i = \mathbf{X}' \mathbf{c}$.

In this sense, solving problem (3.1) with linear kernel in the form $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ is equivalent to solving the ridge regression problem (3.7) whose minimising solution is

$$\begin{aligned} \mathbf{w} &= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}' \mathbf{y} \\ &= \mathbf{X}' \mathbf{c} \\ &= \mathbf{X}' (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \\ &= \mathbf{X}' (\mathbf{X}\mathbf{X}' + \lambda \mathbf{I})^{-1} \mathbf{y}, \end{aligned} \tag{3.8}$$

where the equality of the first and last line is indeed confirmed by the *inversion identity lemma* for positive definite $\mathbf{P}$ and $\mathbf{R}$ (here $\mathbf{B} = \mathbf{X}$, $\mathbf{P} = 1/\lambda \mathbf{I}_D$ and $\mathbf{R} = \mathbf{I}_N$)

$$(\mathbf{P}^{-1} + \mathbf{B}'\mathbf{R}^{-1}\mathbf{B})^{-1} \mathbf{B}' \mathbf{R}^{-1} = \mathbf{P}\mathbf{B}' (\mathbf{B}\mathbf{P}\mathbf{B}' + \mathbf{R})^{-1} \tag{3.9}$$

3.2 Learning for vector-output

3.2.1 Representer theorem for vector-output problems

From the *representer theorem*, the minimisation of a functional

$$\min_{\mathbf{f} \in \mathcal{H}} J(\mathbf{f}) := \sum_i^n \mathcal{L}\big(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)\big) + \lambda \Omega(||\mathbf{f}||_{\mathcal{H}}) \tag{3.10}$$

has a solution $\mathbf{f}^* \in \mathcal{H}$ which admits a representation

$$\mathbf{f}^*(\mathbf{x}) = \sum_i^n \mathbf{H}(\mathbf{x}_i, \mathbf{x}) \, \mathbf{c}_i = \sum_i^n \sum_j^m c_{ij} \, \mathbf{H}(\mathbf{x}_i, \mathbf{x})_{:j}, \tag{3.11}$$

where n is the number of instances of the input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathcal{Y} \subset \mathbb{R}^m\}$, $\mathcal{L}(.)$ is an arbitrary loss function, $\Omega(.)$ is a monotonically increasing function, $||.||_{\mathcal{H}}$ is a norm in the RKHS $\mathcal{H}$, $\mathbf{H}(.,.) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$ is the *matrix-valued kernel* associated with $\mathcal{H}$, and $\mathbf{H}(\mathbf{x}_i, \mathbf{x})_{:j}$ is its j-th column.

Note that $\mathbf{H}$ is a $m \times m$ matrix with t, s elements being the scalar-valued kernels

$$\mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)_{t,s} = h\big((\mathbf{x}_i, t), (\mathbf{x}_j, s)\big) = \langle \mathbf{e}_t, \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j) \mathbf{e}_s \rangle = \langle \mathbf{H}(\mathbf{x}_i, .) \mathbf{e}_t, \mathbf{H}(\mathbf{x}_j, .) \mathbf{e}_s \rangle_{\mathcal{H}}, \tag{3.12}$$

where $\mathbf{e}_t$ is the t-th standard coordinate basis vector of $\mathbb{R}^m$.

From eq. (3.11) and (3.12) we get that the solution for task t admits the representation

$$f(\mathbf{x}, t) = \sum_i^n \sum_s^m c_{is} \, h\big((\mathbf{x}_i, s), (\mathbf{x}, t)\big), \tag{3.13}$$

where $h : \mathcal{X} \times \mathbb{N}_m \times \mathcal{X} \times \mathbb{N}_m \to \mathbb{R}$ is a scalar-valued kernel.

We will furher assume that the scalar kernels h are *seperable* so that

$$\mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)_{s,t} = h\big((\mathbf{x}_i, s), (\mathbf{x}_j, t)\big) = k(\mathbf{x}_i, \mathbf{x}_j)\, l(s,t), \tag{3.14}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathbb{N}_m \times \mathbb{N}_m \to \mathbb{R}$ are scalar-valued input and output kernels respectively. Equivalently,

$$\mathbf{H}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{L}\, k(\mathbf{x}_i, \mathbf{x}_j), \tag{3.15}$$

with $\mathbf{L}$ being the $m \times m$ Gram matrix with elements $L_{st} = l(s,t)$ (here s, t are the task indices).

3.2.2 Solving with least squares and squarred regularizer

For the simple case

$$\min_{\mathbf{f} \in \mathcal{H}} J(\mathbf{f}) := \sum_i^n ||\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)||_2^2 + \lambda ||\mathbf{f}||_{\mathcal{H}}^2 \tag{3.16}$$

we can learn the functions from the equivalent finite-dimensional problem

$$\min_{\mathbf{C}} J(\mathbf{C}) = ||\mathbf{Y} - \mathbf{KCL}||_F^2 + \lambda\, \langle \mathbf{C}'\mathbf{KC}, \mathbf{L} \rangle_F \tag{3.17}$$

where we use eq. (3.11) and eq. (3.15), and where $\mathbf{K}$ is the Gram matrix with the elements $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{Y}$ is the $n \times m$ output matrix, and $\mathbf{C}$ is the $n \times m$ parameters matrix. (note: the choice of the kernels $k(.,.)$ and $l(.,.)$ specifies the space $\mathcal{H}$ we work with).

Proof: Using the property $\langle \mathbf{H}(\mathbf{x}_i, .)\, \mathbf{z}, \mathbf{H}(\mathbf{x}_j, .)\, \mathbf{y} \rangle_{\mathcal{H}} = \langle \mathbf{z}, \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)\, \mathbf{y} \rangle$

$$\begin{aligned}
J(\mathbf{f}) &= \sum_i^n ||\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)||_2^2 + \lambda ||\mathbf{f}||_{\mathcal{H}}^2 \\
&= \sum_i^n ||\mathbf{y}_i||_2^2 - 2 \sum_i^n \langle \mathbf{y}_i, \sum_j^n \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)\, \mathbf{c}_j \rangle + \sum_i^n || \sum_j^n \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)\, \mathbf{c}_j ||_2^2 \\
&\quad + \lambda\, \langle \sum_i^n \mathbf{H}(\mathbf{x}_i, .)\, \mathbf{c}_i, \sum_j^n \mathbf{H}(\mathbf{x}_j, .)\, \mathbf{c}_j \rangle_{\mathcal{H}} \\
&= ||\mathbf{Y}||_F^2 - 2 \sum_i^n \langle \mathbf{y}_i, \sum_j^n k(\mathbf{x}_i, \mathbf{x}_j) \mathbf{L}\, \mathbf{c}_j \rangle + \sum_i^n || \sum_j^n k(\mathbf{x}_i, \mathbf{x}_j) \mathbf{L}\, \mathbf{c}_j ||_2^2 \\
&\quad + \lambda \sum_{ij}^n \langle \mathbf{c}_i, \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)\, \mathbf{c}_j \rangle \\
&= ||\mathbf{Y}||_F^2 - 2 \sum_{ij}^n K_{ij} \langle \mathbf{y}_i, \mathbf{L}\, \mathbf{c}_j \rangle + \sum_{ijz}^n K_{ij} K_{iz} \langle \mathbf{L}\, \mathbf{c}_j, \mathbf{L}\, \mathbf{c}_z \rangle + \lambda \sum_{ij}^n K_{ij} \langle \mathbf{c}_i, \mathbf{L}\, \mathbf{c}_j \rangle \\
&= ||\mathbf{Y}||_F^2 - 2 \langle \mathbf{Y}, \mathbf{KCL} \rangle_F + ||\mathbf{KCL}||_F^2 + \lambda\, \langle \mathbf{C}'\mathbf{KC}, \mathbf{L} \rangle_F \\
&= ||\mathbf{Y} - \mathbf{KCL}||_F^2 + \lambda\, \langle \mathbf{C}'\mathbf{KC}, \mathbf{L} \rangle_F,
\end{aligned}$$

where we used the symmetry of the Gram metrices $\mathbf{L}' = \mathbf{L}$, $\mathbf{K}' = \mathbf{K}$ and

$$\begin{gathered}
\textstyle\sum_{ij} K_{ij} \langle \mathbf{Y}_{i:}, \mathbf{LC}_{j:} \rangle = \sum_{ij} K_{ij} \langle \mathbf{Y}_{i:}, \sum_k \mathbf{L}_{:k} C_{jk} \rangle = \sum_{ij} K_{ij} \sum_l Y_{il} \sum_k L_{lk} C_{jk} = \sum_{ijkl} K'_{ji} Y_{il} L_{lk} C'_{kj} \\
= tr(\mathbf{K}'\mathbf{YLC}') = tr(\mathbf{YLC}'\mathbf{K}') = \langle \mathbf{Y}, \mathbf{KCL} \rangle_F
\end{gathered}$$

$$\begin{gathered}
\textstyle\sum_{ijz} K_{ij} K_{iz} \langle \mathbf{LC}_{j:}, \mathbf{LC}_{z:} \rangle = \sum_{ijz} K_{ij} K_{iz} \langle \sum_k \mathbf{L}_{:k} C_{jk}, \sum_l \mathbf{L}_{:l} C_{zl} \rangle = \sum_{ijzkl} K_{ij} K_{iz} C_{jk} C_{zl} \langle \mathbf{L}_{:k}, \mathbf{L}_{:l} \rangle \\
\textstyle\sum_{ijzklp} K_{ij} C_{jk} L'_{kp} L_{pl} C'_{lz} K'_{zi} = tr(\mathbf{KCL}'\mathbf{LC}'\mathbf{K}') = \langle \mathbf{KCL}, \mathbf{KCL} \rangle_F = ||\mathbf{KCL}||_F^2
\end{gathered}$$

We will solve the vectorised version of (3.17) with $\tilde{\mathbf{c}} = vec(\mathbf{C})$ and $\tilde{\mathbf{y}} = vec(\mathbf{Y})$

$$J(\tilde{\mathbf{c}}) = ||\tilde{\mathbf{y}} - (\mathbf{L} \otimes \mathbf{K})\,\tilde{\mathbf{c}}||_2^2 + \lambda\,\tilde{\mathbf{c}}'\,(\mathbf{L} \otimes \mathbf{K})\,\tilde{\mathbf{c}} \tag{3.18}$$

In analogy to eq. (3.5) we get (after differentiating and equating to zero) the closed form for the minimiser

$$\tilde{\mathbf{c}} = \big((\mathbf{L} \otimes \mathbf{K}) + \lambda \mathbf{I}\big)^{-1} \tilde{\mathbf{y}} \tag{3.19}$$

3.2.3 Kernel multioutput ridge regression

If the RKHS we chose to work with in problems (3.10) is associated with a simple linear input kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ (with Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}'$, $\mathbf{X}$ being the $n \times d$ input matrix) we essentially choose to work with simple linear functions of the form $\mathbf{f}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{W} \rangle$ with the link between the parameters being $\mathbf{W} = \mathbf{X}'\mathbf{C}\mathbf{L}$.

Proof:

$$\begin{aligned} \mathbf{f}(\mathbf{x}_j) &= \sum_i^n \mathbf{H}(\mathbf{x}_i, \mathbf{x}_j)\mathbf{c}_i = \sum_i^n k(\mathbf{X}_{i:}, \mathbf{X}_{j:})\mathbf{L}\,\mathbf{C}_{i:} = \sum_i^n \langle \mathbf{X}_{i:}, \mathbf{X}_{j:} \rangle\,\mathbf{L}\,\mathbf{C}'_{:i} = \sum_i^n \sum_p^d X_{ip} X_{jp}\,(\mathbf{L}\,\mathbf{C}')_{:i} \\ &= \sum_p^d X_{jp}\,\mathbf{L}\mathbf{C}'\mathbf{X}_{:p} = \mathbf{L}\mathbf{C}'\mathbf{X}\mathbf{X}'_{j:} = \langle \mathbf{X}'\mathbf{C}\mathbf{L}, \mathbf{X}'_{j:} \rangle = \langle \mathbf{W}, \mathbf{x}_j \rangle \quad \text{with } \mathbf{W} = \mathbf{X}'\mathbf{C}\mathbf{L} \end{aligned}$$

For simple linear input kernel, the corresponding squared norm regularizer is

$$||\mathbf{f}||^2_{\mathcal{H}} = \langle \mathbf{C}'\mathbf{K}\mathbf{C}, \mathbf{L} \rangle = tr(\mathbf{C}'\mathbf{X}\mathbf{X}'\mathbf{C}\mathbf{L}) = tr(\mathbf{L}^{\mathbf{-1/2}}\mathbf{L}\mathbf{C}'\mathbf{X}\,\mathbf{X}'\mathbf{C}\mathbf{L}\mathbf{L}^{\mathbf{-1/2}}) = ||\mathbf{W}\mathbf{L}^{-1/2}||^2_F, \tag{3.20}$$

where $\mathbf{L}^{-1/2} = \mathbf{L}'^{-1/2}$ and $\mathbf{L}^{-1/2}\mathbf{L}^{-1/2} = \mathbf{L}^{-1}$ due the the PD property of $\mathbf{L}$.

Problem (3.17) with simple linear kernel is thus equivalent to a *multi-output ridge regression*

$$R(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}\mathbf{W}||^2_F + \lambda||\mathbf{W}\mathbf{L}^{-1/2}||^2_F \tag{3.21}$$

If we further choose the output kernel to be diagonal (with diagonal Gram matrix) so that $l(t,s) = \delta_{ts}L_{ts}$ the link between the parameters further simplifies so that the columns $\mathbf{W}_{:t} = L_{tt}\mathbf{X}'\mathbf{C}_{:t}$. Moreover, for the spherical output kernel $\mathbf{L} = \mu\mathbf{I}$ this reduces to $\mathbf{W} = \mu\mathbf{X}'\mathbf{C}$.

In the spherical case the problem (3.17) is equivalent to a standard *ridge regression*

$$R(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}\mathbf{W}||^2_F + \lambda/\mu\,||\mathbf{W}||^2_F \tag{3.22}$$

To show the equivalence of the solutions of (3.17) and (14.7) we use the vectorisation of the transposed problem using $\tilde{\mathbf{y}} = vec(\mathbf{Y}')$ and $\tilde{\mathbf{w}} = vec(\mathbf{W}')$

$$R(\tilde{\mathbf{w}}) = ||\tilde{\mathbf{y}} - (\mathbf{X} \otimes \mathbf{I})\tilde{\mathbf{w}}||^2_2 + \lambda/\mu\,||\tilde{\mathbf{w}}||^2_2 \tag{3.23}$$

The minimiser of the ridge regression problem (14.7) has a known closed form solution in a vectorised form (easy to derive by differentiating and equating to zero)

$$\tilde{\mathbf{w}} = \big((\mathbf{X}'\mathbf{X} \otimes \mathbf{I}) + \lambda/\mu\,\mathbf{I}\big)^{-1}(\mathbf{X}' \otimes \mathbf{I})\tilde{\mathbf{y}} \tag{3.24}$$

which indeed coincides with the minimising solution for $\tilde{\mathbf{c}}$ with $\tilde{\mathbf{w}} = \mu(\mathbf{X} \otimes \mathbf{I})\tilde{\mathbf{c}}$

Proof: We use $vec(\mathbf{W}') = \mu\,vec(\mathbf{C}'\mathbf{X}) = \mu(\mathbf{X}' \otimes \mathbf{I})\tilde{\mathbf{c}}$ and the matrix inversion identity lemma for positive definite $\mathbf{P}$ and $\mathbf{R}$

$$(\mathbf{P}^{-1} + \mathbf{B}'\mathbf{R}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{R}^{-1} = \mathbf{P}\mathbf{B}'(\mathbf{B}\mathbf{P}\mathbf{B}' + \mathbf{R})^{-1} \tag{3.25}$$

where we put $\mathbf{B} = (\mathbf{X} \otimes \mathbf{I})$, $\mathbf{P} = \mathbf{I}$ and $\mathbf{R} = \lambda/\mu\mathbf{I}$

$$\begin{aligned}
\tilde{\mathbf{w}} &= \big((\mathbf{X}'\mathbf{X} \otimes \mathbf{I}) + \lambda/\mu\,\mathbf{I}\big)^{-1}(\mathbf{X}' \otimes \mathbf{I})\tilde{\mathbf{y}} \\
&= \mu(\mathbf{X}' \otimes \mathbf{I})\tilde{\mathbf{c}} \\
&= \mu(\mathbf{X}' \otimes \mathbf{I})\big((\mathbf{K} \otimes \mathbf{L}) + \lambda\mathbf{I}\big)^{-1}\tilde{\mathbf{y}} \\
&= \mu(\mathbf{X}' \otimes \mathbf{I})\big((\mathbf{X}\mathbf{X}' \otimes \mu\mathbf{I}) + \lambda\mathbf{I}\big)^{-1}\tilde{\mathbf{y}} \\
&= (\mathbf{X}' \otimes \mathbf{I})\big((\mathbf{X}\mathbf{X}' \otimes \mathbf{I}) + \lambda/\mu\mathbf{I}\big)^{-1}\tilde{\mathbf{y}}
\end{aligned}$$

3.2.4 Output kernel vs. covariance matrix for Gaussian errors and Gaussian priors on the parameters

We formulate Guassian linear regression problem with Gaussian priors on the parameters $\mathbf{Z}$

$$\mathbf{y}_i = \mathbf{Z}'\mathbf{x}_i + \mathbf{e}_i,\ \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),\ Cov(\mathbf{e}_i, \mathbf{e}_j) = 0,\ z_{kl} \sim \mathcal{N}(0, \sigma^2),\ Cov(z_{kl}, z_{sr}) = 0,\ \forall i, j \in \mathbb{N}_n \tag{3.26}$$

Multivariate *Gaussian* distribution density is

$$f(\mathbf{y}_i) = (2\pi)^{-0.5k}|\mathbf{\Sigma}|^{-0.5}e^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i)} \tag{3.27}$$

so that the likelihood is

$$\mathcal{L}(\mathbf{Z}|\mathbf{y}_i, \mathbf{x}_i, i \in \mathbb{N}_n) = \prod_{i=1}^{n}(2\pi)^{-0.5k}|\mathbf{\Sigma}|^{-0.5}e^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i)} \tag{3.28}$$

And the posterior is

$$\mathcal{P}(\mathbf{Z}|\mathbf{y}_i, \mathbf{x}_i, i \in \mathbb{N}_n) \propto \mathcal{L}(\mathbf{Z}|\mathbf{y}_i, \mathbf{x}_i, i \in \mathbb{N}_n)\prod_{kl}(2\pi)^{-0.5}\sigma^{-1}e^{-\frac{z_{kl}^2}{2\sigma^2}} \tag{3.29}$$

To find optimal parameters $\mathbf{Z}$ we will minimize the negative log of the posterior (instead of maximizing the posterior directly

$$\begin{aligned}
-\ln\mathcal{P}(\mathbf{Z}|\mathbf{y}_i, \mathbf{x}_i, i \in \mathbb{N}_n) &\propto \sum_{i}^{n}\frac{1}{2}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{Z}'\mathbf{x}_i) + \sum_{kl}\frac{z_{kl}^2}{2\sigma^2} \\
&= \frac{1}{2}tr\Big((\mathbf{Y} - \mathbf{X}\mathbf{Z})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{Z})'\Big) + \frac{1}{2\sigma^2}||\mathbf{Z}||_F^2 \\
&= \frac{1}{2}||(\mathbf{Y} - \mathbf{X}\mathbf{Z})\mathbf{\Sigma}^{-1/2}||_F^2 + \frac{1}{2\sigma^2}||\mathbf{Z}||_F^2, \quad \text{where } \mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}'^{-1/2} \\
&= \frac{1}{2}||(\mathbf{Y}\mathbf{\Sigma}^{-1/2} - \mathbf{X}\mathbf{W})||_F^2 + \frac{1}{2\sigma^2}||\mathbf{W}\mathbf{\Sigma}^{1/2}||_F^2, \quad \text{where } \mathbf{W} = \mathbf{Z}\mathbf{\Sigma}^{-1/2}
\end{aligned}$$

We will use this change of variable to formulate the final optimisation problem

$$\arg\min_{W}\ \frac{1}{2}||(\mathbf{Y}\mathbf{\Sigma}^{-1/2} - \mathbf{X}\mathbf{W})||_F^2 + \frac{1}{2\lambda^2}||\mathbf{W}\mathbf{\Sigma}^{1/2}||_F^2 \tag{3.30}$$

which is similar to (3.21) but not quite.

Next, we will assume a covariance in the Gaussian priors of the parameters $\mathbf{Z}$ so that $\text{Vec}(\mathbf{Z}) = \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_z)$ and the error covariance $\mathbf{\Sigma} = \mathbf{I}$.

The posterior is then

$$\mathcal{P}(\mathbf{Z}|\mathbf{y}_i,\mathbf{x}_i, i\in\mathbb{N}_n) \propto \mathcal{L}(\mathbf{Z}|\mathbf{y}_i,\mathbf{x}_i, i\in\mathbb{N}_n)\,(2\pi)^{-0.5kd}|\mathbf{\Sigma}_z|^{-0.5}e^{-\frac{1}{2}\mathbf{z}'\mathbf{\Sigma}_z^{-1}\mathbf{z}} \tag{3.31}$$

And the negative log-likelihood

$$\begin{aligned} -\ln\mathcal{P}(\mathbf{Z}|\mathbf{y}_i,\mathbf{x}_i, i\in\mathbb{N}_n) &\propto \sum_i^n \frac{1}{2}(\mathbf{y}_i-\mathbf{Z}'\mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{Z}'\mathbf{x}_i)+\frac{1}{2}\mathbf{z}'\mathbf{\Sigma}_z^{-1}\mathbf{z} \\ &= \frac{1}{2}tr\Big((\mathbf{Y}-\mathbf{XZ})\mathbf{I}(\mathbf{Y}-\mathbf{XZ})'\Big)+\frac{1}{2}||\mathbf{\Sigma}_z^{-1/2}\mathbf{z}||_2^2, \quad \text{where } \mathbf{\Sigma}_z^{-1/2}=\mathbf{\Sigma}_z'^{-1/2} \\ &= \frac{1}{2}||(\mathbf{Y}-\mathbf{XZ})||_F^2+\frac{1}{2}||\mathbf{AZB}||_F^2, \quad \text{where } \mathbf{\Sigma}_z^{-1/2}=\mathbf{B}\otimes\mathbf{A} \end{aligned}$$

For $\mathbf{A}=\mathbf{I}$, which is equivalent to the prior on $\mathbf{Z}$ as being composed of d independent identically distributed rows each of which is Gaussian with $\mathbf{Z}_{i:}\sim\mathcal{N}(\mathbf{0},(\mathbf{BB}')^{-1})$, *Note:*

$$||\mathbf{AZB}||_F^2->tr(\mathbf{B}'\mathbf{Z}'\mathbf{A}'\mathbf{AZB})->tr(\mathbf{BB}'\mathbf{Z}'\mathbf{A}'\mathbf{AZ})$$

we get the optimisation problem

$$\arg\min_W \ \frac{1}{2}||(\mathbf{Y}-\mathbf{XZ})||_F^2+\frac{\lambda}{2}||\mathbf{ZB}/\lambda||_F^2 \tag{3.32}$$

By comparison to (3.21) we see that the $\mathbf{L}^{-1/2}=\mathbf{B}/\lambda$ and therefore $\mathbf{L}^{-1}=\mathbf{BB}'/\lambda^2$ and $\mathbf{L}=\lambda^2(\mathbf{BB}')^{-1}$ which is the scaled covariance matrix of the prior on the independent rows of the parameters matrix $\mathbf{Z}$.

4 Multiple kernel learning

The multiple kernel learning (MKL) problem is cast in similar form as the standard function learning in RKHS (section 3, equation (3.1))

$$\min_{\gamma\in\Delta}\min_{f\in\mathcal{H}} J(f) := \sum_i^n \big(y_i - f(\mathbf{x}_i)\big)^2 + \lambda \, ||f||_{\mathcal{H}}^2 \tag{4.1}$$

The RKHS $\mathcal{H}$ is endowed with a kernel function $k(.,.) = \sum_{j=1}^m \gamma_j k_j(.,.)$, where the domain Δ of γ is the simplex $\Delta = \{\gamma \in \mathbb{R}^m : \gamma_j \geq 0, \sum_{j=1}^m \gamma_j = 1\}$

Using the representer theorem, introducing the Gram matrices $\mathbf{K}_j(l,s) = k_j(x_l, x_s)$ and $\mathbf{K} = \sum_{j=1}^m \gamma_j \mathbf{K}_j$, problem (4.1) can be rewritten as a finite-dimensional optimisation problem

$$\min_{\gamma\in\Delta}\min_{\mathbf{c}\in\mathbb{R}^n} J(\mathbf{c},\gamma) := ||\mathbf{y} - \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda \, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} \tag{4.2}$$

Proof: Follows simply from the equivalent proof in section 3.1

4.1 Kernel normalization

The original MKL learning problem in [2] was in the form (**??**) with the following constraints $\sum_j \gamma_j \mathbf{K}_j \succeq 0$, $\gamma_j \geq 0$ and $\text{Tr}(\sum_j \gamma_j \mathbf{K}_j) \leq c$. If all the individual kernels are PSD the first constraint can be dropped due to the non-negative constraint on γ_j. Also, if the kernels are normalised so the $k_j(\mathbf{x}_i, \mathbf{x}_i) = 1$ for all i, j the trace constraint reduces to $\sum_j \gamma_j = 1$ and thus we get the simplex constraint on γ.

How do we normalize the kernels so that we can use this simple constraint?

For a kernel function $k(.,.) = \langle \Phi(.), \Phi(.) \rangle$ and the corresponding gram matrix $\mathbf{K}$ we want to have a normalised kernel function $\tilde{k}(.,.)$ and gram matrix $\widetilde{\mathbf{K}}$ such that $\widetilde{\mathbf{K}}_{ii} = \tilde{k}(\mathbf{x}_i, \mathbf{x}_i) = \langle \widetilde{\Phi}(\mathbf{x}_i), \widetilde{\Phi}(\mathbf{x}_i) \rangle = 1$. This is true, if we define $\widetilde{\Phi}(\mathbf{x}_i)$ as the normalised version of $\Phi(\mathbf{x}_i)$ so that $\widetilde{\Phi}(\mathbf{x}_i) = \Phi(\mathbf{x}_i)/\sqrt{\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \rangle}$.

Now, given a non-normalised gram matrix $\mathbf{K}$ we want to find the corresponding $\widetilde{\mathbf{K}}$.

$$\begin{aligned}
\widetilde{\mathbf{K}}_{ij} &= \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \langle \widetilde{\Phi}(\mathbf{x}_i), \widetilde{\Phi}(\mathbf{x}_j) \rangle = \Big\langle \frac{\Phi(\mathbf{x}_i)}{\sqrt{\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \rangle}}, \frac{\Phi(\mathbf{x}_j)}{\sqrt{\langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_j) \rangle}} \Big\rangle \\
&= \frac{\mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}\mathbf{K}_{jj}}} = \frac{k(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{k(\mathbf{x}_i, \mathbf{x}_i), k(\mathbf{x}_j, \mathbf{x}_j)}} \qquad (4.3)
\end{aligned}$$

4.2 Solving MKL

Instead of the constrained problem (4.2), we will optimise its Lagrangian

$$\min_{\gamma\in\mathbb{R}_+^m}\min_{\mathbf{c}\in\mathbb{R}^n} J(\mathbf{c},\gamma) := ||\mathbf{y} - \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda \, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} + \kappa \sum_{j=1}^m \gamma_j \tag{4.4}$$

4.2.1 Only need 1 regularization hyper-parameter

Next, I'll show that we, in fact, do not need two regularization parameters but should only use one.

$$\begin{array}{rr} ||\mathbf{y} - \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} + \kappa \sum_{j=1}^m \gamma_j = & \\ ||\mathbf{y} - \sum_{j=1}^m \kappa\gamma_j \mathbf{K}_j \mathbf{c}/\kappa||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \kappa\gamma_j \mathbf{K}_j \mathbf{c}/\kappa + \sum_{j=1}^m \kappa\gamma_j = & \\ ||\mathbf{y} - \sum_{j=1}^m \nu_j \mathbf{K}_j \mathbf{q}||_2^2 + \lambda\kappa\, \mathbf{q}^T \sum_{j=1}^m \nu_j \mathbf{K}_j \mathbf{q} + \sum_{j=1}^m \nu_j & \text{where } \mu = \kappa\gamma,\ \mathbf{q} = \mathbf{c}/\kappa \end{array}$$

From this we see that the hyperparameter κ is absorbed into scaling of the γ and $\mathbf{c}$ vectors and the first regularization parameter. Since minimising the problem in the first and the last line is equivalent we can put $\kappa = 1$ and instead of problem (4.4) work with

$$\min_{\gamma \in \mathbb{R}_+^m} \min_{\mathbf{c} \in \mathbb{R}^n} J(\mathbf{c}, \gamma) := ||\mathbf{y} - \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} + \sum_{j=1}^m \gamma_j \tag{4.5}$$

4.2.2 Alternating minimisation

We can alternate between vectors γ and $\mathbf{c}$.

1) For fixed γ the solution for $\mathbf{c}$ is in the closed form

$$(\mathbf{K} + \lambda \mathbf{I}_n)\, \mathbf{c} = \mathbf{y}, \quad \text{where } \mathbf{K} = \sum_{j=1}^m \gamma_j \mathbf{K}_j \tag{4.6}$$

2) For fixed c we will rewrite the J function as

$$\begin{aligned} J(\gamma) &= ||\mathbf{y} - \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} + \sum_{j=1}^m \gamma_j \\ &= ||\mathbf{y} - \mathbf{Z}\,\gamma||_2^2 + \lambda \mathbf{c}^T \mathbf{Z}\,\gamma + ||\gamma||_1 \qquad \text{where } \mathbf{Z}_{:j} = \mathbf{K}_j \mathbf{c} \\ &= f(\gamma) + ||\gamma||_1, \end{aligned} \tag{4.7}$$

where $f(\gamma) = ||\mathbf{y} - \mathbf{Z}\,\gamma||_2^2 + \lambda\, \mathbf{c}^T \mathbf{Z}\,\gamma$ is convex differentiable and the minimisation can be solved by proximal gradient descent with steps

$$\gamma^{k+1} = prox_{\alpha||.||_1}(\gamma^k - \alpha \nabla f(\gamma^k)), \tag{4.8}$$

where α is the step size, the gradient of f is

$$\nabla f(\gamma) = -2\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}\,\gamma) + \lambda\, \mathbf{Z}^T \mathbf{c} \tag{4.9}$$

and the proximal operator is

$$prox_{\alpha||.||_1}(\mathbf{v}) := \arg\min_{\mathbf{x}} ||\mathbf{x}||_1 + \frac{1}{2\alpha} ||\mathbf{x} - \mathbf{v}||_2^2 \tag{4.10}$$

with a solution that can be expressed element-wise

$$[prox_{\alpha||.||_1}(\mathbf{v})]_i = sgn(v_i)(|v_i| - \alpha)_+ \tag{4.11}$$

4.2.3 Change of variables

First we rewrite equation (4.5) as

$$\begin{gathered}
\min_{\gamma \in \mathbb{R}_{+}^{m}, \mathbf{c} \in \mathbb{R}^{n}} ||\mathbf{y} - \textstyle\sum_{j}^{m} \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \textstyle\sum_{j=1}^{m} \gamma_j \mathbf{K}_j \mathbf{c} + \textstyle\sum_{j}^{m} \gamma_j = \\
\min_{\gamma \in \mathbb{R}_{+}^{m}, \mathbf{z} \in \mathbb{R}^{n}m} ||\mathbf{y} - \textstyle\sum_{j}^{m} \gamma_j \mathbf{\Phi}_j \mathbf{\Phi}_j^T \mathbf{c}||_2^2 + \lambda \textstyle\sum_{j=1}^{m} \gamma_j \mathbf{c}^T \mathbf{\Phi}_j \mathbf{\Phi}_j^T \mathbf{c} + \textstyle\sum_{j}^{m} \gamma_j = \\
\min_{\gamma \in \mathbb{R}_{+}^{m}, \mathbf{z} \in \mathbb{R}^{n}m} ||\mathbf{y} - \textstyle\sum_{j}^{m} \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + \lambda \textstyle\sum_{j=1}^{m} \mathbf{z}_j^T \mathbf{z}_j / \gamma_j + \textstyle\sum_{j}^{m} \gamma_j, \quad \text{where } \mathbf{z}_j = \gamma_j \mathbf{\Phi}_j^T \mathbf{c}
\end{gathered} \tag{4.12}$$

We first minimise problem (4.12) with respect to γ.

$$\frac{\partial J(\mathbf{z}, \gamma)}{\partial \gamma_j} = -\lambda \frac{||\mathbf{z}_j||_2^2}{\gamma_j^2} + 1 \tag{4.13}$$

We confirm that J is a convex function of γ_j

$$\frac{\partial^2 J(\mathbf{z}, \gamma)}{\partial \gamma_j \partial \gamma_j} = 2\lambda \frac{||\mathbf{z}_j||_2^2}{\gamma_j^3} > 0 \text{ (since } \gamma_j > 0) \tag{4.14}$$

and find the minimising solution

$$\begin{aligned}
-\lambda \frac{||\mathbf{z}_j||^2}{\gamma_j^2} + 1 &= 0 \\
\gamma_j^2 &= \lambda ||\mathbf{z}_j||_2^2 \\
\gamma_j &= \sqrt{\lambda} ||\mathbf{z}_j||_2
\end{aligned} \tag{4.15}$$

We plug this back to equation (4.12)

$$\begin{gathered}
\min_{\gamma \in \mathbb{R}_{+}^{m}, \mathbf{z} \in \mathbb{R}^{n}m} ||\mathbf{y} - \textstyle\sum_{j}^{m} \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + \lambda \textstyle\sum_{j=1}^{m} \mathbf{z}_j^T \mathbf{z}_j / \gamma_j + \textstyle\sum_{j}^{m} \gamma_j = \\
\min_{\mathbf{z} \in \mathbb{R}^{n}m} ||\mathbf{y} - \textstyle\sum_{j}^{m} \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + 2\sqrt{\lambda} \textstyle\sum_{j=1}^{m} ||\mathbf{z}_j||_2 = \\
\min_{\mathbf{z} \in \mathbb{R}^{n}m} ||\mathbf{y} - \mathbf{\Phi} \mathbf{z}||_2^2 + 2\sqrt{\lambda} \textstyle\sum_{j=1}^{m} ||\mathbf{z}_j||_2 = \\
\min_{\mathbf{z} \in \mathbb{R}^{n}m} f(\mathbf{z}) + g(\mathbf{z}),
\end{gathered} \tag{4.16}$$

where $\mathbf{\Phi} = [\mathbf{\Phi}_1 \mathbf{\Phi}_2 \ldots \mathbf{\Phi}_m]$, $\mathbf{z} = [\mathbf{z}_1^T \mathbf{z}_2^T \ldots \mathbf{z}_m^T]^T$, $f(\mathbf{z}) = ||\mathbf{y} - \mathbf{\Phi}\mathbf{z}||_2^2$ is convex differentiable in $\mathbf{z}$ and $g(\mathbf{z}) = 2\sqrt{\lambda} \sum_{j=1}^{m} ||\mathbf{z}_j||_2$ is convex non-differentiable.
We can solve this by the proximal gradient descent with steps

$$\mathbf{z}^{k+1} = prox_{\alpha g(.)}(\mathbf{z}^k - \alpha \nabla f(\mathbf{z}^k)), \tag{4.17}$$

where α is the step size, the gradient of f is

$$\nabla f(\mathbf{z}) = -2\mathbf{\Phi}^T (\mathbf{y} - \mathbf{\Phi}\, \mathbf{z}) \tag{4.18}$$

and the proximal operator is

$$prox_{\alpha g}(\mathbf{v}) := \arg\min_{\mathbf{x}} 2\sqrt{\lambda} \sum_{j=1}^{m} ||\mathbf{x}_j||_2 + \frac{1}{2\alpha} ||\mathbf{x} - \mathbf{v}||_2^2 \tag{4.19}$$

with a solution that can be expressed element-wise

$$[prox_{\alpha g}(\mathbf{v})]_j = \mathbf{v}_j \Big(1 - \frac{2\sqrt{\lambda}\alpha}{||\mathbf{v}_j||_2}\Big)_+ \tag{4.20}$$

Finally, we recover $\mathbf{c}$ by solving the set of equations $\mathbf{z}_j = \sqrt{\lambda}||\mathbf{z}_j||_2 \mathbf{\Phi}_j^T \mathbf{c}$ for all j.

Let's derive the solution using the standard mechanism (note that we do not need to go through the primal-dual mechanism, we only need to use change of variable) We want to minimise the Lagrangian is

$$L(\mathbf{w}, \lambda) := \sum_i^n \big(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle\big)^2 + \lambda\,||\mathbf{w}||_2^2 \tag{4.21}$$

The optimality condition for the minimum of the primal Lagrangian (4.21) yields

$$\frac{\partial L}{\partial \mathbf{w}} = \sum_i^n (-2y_i\mathbf{x}_i + 2\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}) + 2\lambda\mathbf{w} = 0 \tag{4.22}$$

so that

$$\begin{aligned} \mathbf{w}^* &= (\sum_i^n \mathbf{x}_i\mathbf{x}_i^T + \lambda\mathbf{I}_d)^{-1} \sum_i^n y_i\mathbf{x}_i \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y} \qquad \text{using eq (3.9)} \\ &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_n)^{-1}\mathbf{y} \\ &= \mathbf{X}^T(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{X}^T\mathbf{c} \end{aligned} \tag{4.23}$$

So that the solution is

$$\begin{aligned} f^*(\mathbf{x}_j) &= \langle \mathbf{w}^*, \mathbf{x}_j \rangle = \langle \mathbf{X}^T(\mathbf{K} + \lambda\mathbf{I}_N)^{-1}\mathbf{y},\, \mathbf{x}_i \rangle \\ &= \langle \mathbf{X}^T\mathbf{c},\, \mathbf{x}_j \rangle = \langle \sum_i^n \mathbf{x}_i c_i,\, \mathbf{x}_j \rangle = \sum_i^n c_i \langle \mathbf{x}_i,\, \mathbf{x}_j \rangle = \sum_i^n c_i\, k(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \tag{4.24}$$

where $\mathbf{c} = (\mathbf{K} + \lambda\mathbf{I}_N)^{-1}\mathbf{y}$, which indeed is the solution to problem (3.5).

In the simple linear case we have

$$\begin{aligned} J(\mathbf{w}) &:= ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2 \quad \text{from eq (4.23)} \\ &:= ||\mathbf{y} - \mathbf{X}\mathbf{X}^T(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y}||_2^2 + \lambda||\mathbf{X}^T(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y}||_2^2 \\ &:= ||\mathbf{y} - \mathbf{K}\mathbf{c}||_2^2 + \lambda\mathbf{c}^T\mathbf{K}\mathbf{c} \end{aligned} \tag{4.25}$$

For the MKL case we have

$$\begin{aligned} J(\mathbf{c}, \gamma) &:= ||\mathbf{y} - \sum_{j=1}^m \gamma_j\mathbf{K}_j\mathbf{c}||_2^2 + \lambda\,\mathbf{c}^T \sum_{j=1}^m \gamma_j\mathbf{K}_j\mathbf{c} \\ &:= ||\mathbf{y} - \sum_{j=1}^m \gamma_j\mathbf{X}_j\mathbf{X}_j^T\mathbf{c}||_2^2 + \lambda\,\mathbf{c}^T \sum_{j=1}^m \gamma_j\mathbf{X}_j\mathbf{X}_j^T\mathbf{c} \\ &:= ||\mathbf{y} - \sum_{j=1}^m \mathbf{X}_j\gamma_j\mathbf{X}_j^T\mathbf{c}||_2^2 + \lambda\, \sum_{j=1}^m \gamma_j\mathbf{c}^T\mathbf{X}_j\mathbf{X}_j^T\mathbf{c}\gamma_j/\gamma_j \\ &:= ||\mathbf{y} - \sum_{j=1}^m \mathbf{X}_j\mathbf{w}_j||_2^2 + \lambda\, \sum_{j=1}^m \frac{||\mathbf{w}_j||_2^2}{\gamma_j} \qquad \text{where } \mathbf{w}_j = \gamma_j\mathbf{X}_j^T\mathbf{c} \end{aligned} \tag{4.26}$$

In the final line in the above we have to extend the regularizer function $\Omega : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}_+$ in the form $\Omega(x, y) = \frac{||\mathbf{x}||_2^2}{y}$ to point $(0, 0)$ so that $\Omega(0, 0) = 0$ (by convention).

As explained for example in [1] this is equivalent to a group-lasso penalty since

$$\min_{\gamma \in \Delta_p} \sum_{j=1}^m \frac{||\mathbf{w}_j||_2^2}{\gamma_j} = ||\mathbf{w}||_{1,2}^2 = (\sum_j ||\mathbf{w}_j||_2)^2 \tag{4.27}$$

Proof: From Cauchy-Schwarz inequality we have $(\langle \mathbf{u}, \mathbf{v}\rangle)^2 = (\sum u_i v_i)^2 \leq \langle \mathbf{u}, \mathbf{u}\rangle\langle \mathbf{v}, \mathbf{v}\rangle = \sum u_i^2 \sum v_i^2$. The equality holds only if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent.

We therefore have

$$(\sum_j ||\mathbf{w}_j||_2)^2 = \Big(\sum_j \frac{||\mathbf{w}_j||_2}{\sqrt{\gamma_j}}\sqrt{\gamma_j}\Big)^2 \leq \sum_j \frac{||\mathbf{w}_j||_2^2}{\gamma_j}\sum_j \gamma_j = \sum_j \frac{||\mathbf{w}_j||_2^2}{\gamma_j}, \tag{4.28}$$

where the last equality comes from the simplex constraint. The CS is an equality if $c\,\frac{||\mathbf{w}_j||_2}{\sqrt{\gamma_j}} = \sqrt{\gamma_j}$ for all groups j, that is if $\gamma_j = c\,||\mathbf{w}_j||_2$. Next from the simplex constraint we have $\sum_g \gamma_j = c\sum_j ||\mathbf{w}_j||_2 = 1$ and therefore $c = 1/\sum_j ||\mathbf{w}_j||_2$ and $\gamma_j = ||\mathbf{w}_j||_2/\sum_j ||\mathbf{w}_j||_2$

The question is now how to solve the kernel version and not the feature-space version of the minimisation problem (4.26).

References

[1] Bach, F., Jenatton, R., Mairal, J., & Obozinski, G. Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning, 2012

[2] Lanckriet, G., & Cristianini, N. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research, 2004

[3] Bhlmann, P., Rtimann, P., van de Geer, S., & Zhang, C.-H. Correlated variables in regression: Clustering and sparse estimation. Journal of Statistical Planning and Inference, 2013

[4] Simon, N., & Tibshirani, R. Standardization and the Group Lasso Penalty. Statistica Sinica, 2012

$$\min_{\gamma\in\Delta}\min_{f\in\mathcal{H}} J(f) := \sum_i^n \big(y_i - \sum_j^m \gamma_j f_j(\mathbf{x}_i)\big)^2 + \lambda \sum_j^m \gamma_j ||f_j||_{\mathcal{H}_j}^2 \tag{4.29}$$

$$\begin{aligned} J(f) &= \sum_i^n \big(y_i - \sum_j^m \gamma_j f_j(\mathbf{x}_i)\big)^2 + \lambda \sum_j^m \gamma_j ||f_j||_{\mathcal{H}_j}^2 && (4.30)\\ &= \sum_i^n \big(y_i - \sum_j^m \gamma_j \Phi_j \mathbf{w}_j\big)^2 + \lambda \sum_j^m \gamma_j ||\mathbf{w}_j||_2^2 \\ &= \sum_i^n \big(y_i - \sum_j^m \Phi_j \mathbf{z}_j\big)^2 + \lambda \sum_j^m ||\mathbf{z}_j||_2^2/\gamma_j, \quad \text{where } \mathbf{z}_j = \gamma_j \mathbf{w}_j && (4.31) \end{aligned}$$

$$\begin{gathered} \min_{\gamma\in\Delta}\min_{f\in\mathcal{H}} \textstyle\sum_i^n \big(y_i - \sum_j^m \gamma_j f_j(\mathbf{x}_i)\big)^2 + \lambda \sum_j^m \gamma_j ||f_j||_{\mathcal{H}_j}^2 = \\ \min_{\gamma\in\Delta}\min_{\mathbf{z}\in\mathcal{R}^{nm}} \textstyle\sum_i^n \big(y_i - \sum_j^m \Phi_j(\mathbf{x}_i)^T \mathbf{z}_j\big)^2 + \lambda \sum_j^m ||\mathbf{z}_j||_2^2/\gamma_j = \\ \min_{\mathbf{z}\in\mathcal{R}^{nm}} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + \lambda \sum_j^m ||\mathbf{z}_j||_2^2/\gamma_j \end{gathered} \tag{4.32}$$

$$\begin{gathered} \min_{\gamma\in\Delta,\mathbf{c}\in\mathbb{R}^n} ||\mathbf{y} - \textstyle\sum_j^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} = \\ \min_{\gamma\in\Delta,\mathbf{c}\in\mathbb{R}^n} ||\mathbf{y} - \textstyle\sum_j^m \gamma_j \mathbf{\Phi}_j \mathbf{\Phi}_j^T \mathbf{c}||_2^2 + \lambda \sum_{j=1}^m \gamma_j \mathbf{c}^T \mathbf{\Phi}_j \mathbf{\Phi}_j^T \mathbf{c} = \\ \min_{\gamma\in\Delta,\mathbf{z}\in\mathbb{R}^{nm}} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + \lambda \sum_{j=1}^m ||\mathbf{z}_j||_2^2/\gamma_j, \quad \text{where } \mathbf{z}_j = \gamma_j \mathbf{\Phi}_j^T \mathbf{c} \\ \min_{\gamma\in\Delta,\mathbf{z}\in\mathbb{R}^{nm}} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{\Phi}_j \mathbf{z}_j||_2^2 + \lambda^2 \,(\sum_{j=1}^m ||\mathbf{z}_j||_2)^2 \end{gathered} \tag{4.33}$$

From the two equivalences above we see that the problem (4.2) and (4.29) are equivalent.

Next we define function $g_j = \gamma_j f_j$ so that from problem (4.29) we get

$$\min_{\gamma \in \Delta} \min_{f \in \mathcal{H}} J(f) := \sum_i^n \big(y_i - \sum_j^m g_j(\mathbf{x}_i)\big)^2 + \lambda \sum_j^m ||g_j||^2_{\mathcal{H}_j}/\gamma_j \tag{4.34}$$

When we minimise this with respect to γ_j we find the minimum is achieved at a point

$$\gamma_j = \frac{||g_j||_{\mathcal{H}_j}}{\sum_j^m ||g_j||_{\mathcal{H}_j}} \tag{4.35}$$

$$\begin{aligned} &\min\nolimits_{\gamma \in \mathbb{R}^m_+, \mathbf{c} \in \mathbb{R}^n} ||\mathbf{y} - \textstyle\sum_j^m \gamma_j \mathbf{K}_j \mathbf{c}||_2^2 + \lambda\, \mathbf{c}^T \sum_{j=1}^m \gamma_j \mathbf{K}_j \mathbf{c} + \sum_j^m \gamma_j = \\ &\min\nolimits_{\gamma \in \mathbb{R}^m_+, \mathbf{z} \in \mathbb{R}^n m} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{K}_j \mathbf{z_j}||_2^2 + \lambda \sum_j^m \mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j/\gamma_j + \sum_j^m \gamma_j, \quad \text{where } \mathbf{z}_j = \gamma_j \mathbf{c} \end{aligned} \tag{4.36}$$

We first minimise problem (4.36) with respect to γ.

$$\frac{\partial J(\mathbf{z}, \gamma)}{\partial \gamma_j} = -\lambda \frac{\mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j}{\gamma_j^2} + 1 \tag{4.37}$$

We confirm that J is a convex function in γ_j by the 2nd derivative condition

$$\frac{\partial^2 J(\mathbf{z}, \gamma)}{\partial \gamma_j \partial \gamma_j} = 2\lambda \frac{\mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j}{\gamma_j^3} > 0 \quad (\text{since } \gamma_j > 0) \tag{4.38}$$

So the minimum is attained at point γ_j^*

$$\begin{aligned} -\lambda \frac{\mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j}{\gamma_j^{*2}} + 1 &= 0 \\ \gamma_j^{*2} &= \lambda \mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j \\ \gamma_j * &= \sqrt{\lambda \mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j} \end{aligned} \tag{4.39}$$

We can plug this back to equation (4.36) to get

$$\begin{aligned} &\min\nolimits_{\gamma \in \mathbb{R}^m_+, \mathbf{z} \in \mathbb{R}^n m} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{K}_j \mathbf{z_j}||_2^2 + \lambda \sum_j^m \mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j/\gamma_j + \sum_j^m \gamma_j, = \\ &\min\nolimits_{\mathbf{z} \in \mathbb{R}^n m} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{K}_j \mathbf{z_j}||_2^2 + 2\sqrt{\lambda} \sum_j^m \sqrt{\mathbf{z}_j^T \mathbf{K}_j \mathbf{z}_j} = \\ &\min\nolimits_{\mathbf{z} \in \mathbb{R}^n m} ||\mathbf{y} - \textstyle\sum_j^m \mathbf{K}_j \mathbf{z_j}||_2^2 + 2\sqrt{\lambda} \sum_j^m ||\mathbf{K}_j^{1/2} \mathbf{z}_j||_2 \end{aligned} \tag{4.40}$$

which is a form of groupwise prediction penalty as in [3] or standardised group lasso of [4].

I will use proximal gradient descent and for the proximal use the separability property of proximals.

http://math.stackexchange.com/questions/175263/gradient-and-hessian-of-general-2-norm

$$\frac{\partial J(\mathbf{z})}{\partial \mathbf{z}_i} = -\mathbf{K}_i^T (\mathbf{y} - \sum_j^m \mathbf{K}_j \mathbf{z_j}) + 2\lambda \frac{\mathbf{K}_i^T \mathbf{z}_i}{\sqrt{\mathbf{z}_i^T \mathbf{K}_i \mathbf{z}_i}} \tag{4.41}$$

$$\frac{\partial^2 J(\mathbf{z})}{\partial \mathbf{z}_i \partial \mathbf{z}_i} = \frac{2\lambda \mathbf{K}_i}{\sqrt{\mathbf{z}_i^T \mathbf{K}_i \mathbf{z}_i}} - 2\lambda \mathbf{K}_i^T \mathbf{z}_i \, (\mathbf{z}_i^T \mathbf{K}_i \mathbf{z}_i)^{-3/2} \mathbf{K}_i^T \mathbf{z}_i \tag{4.42}$$

.... hmmm http://math.stackexchange.com/questions/811376/hessian-of-a-square-root-of-a-quadratic-form

Let's do it differently

5 Double regularization with single tunning parameter

The output-kernel learning problem can be written as

$$\min_{\mathbf{C},\mathbf{L}} J(\mathbf{C},\mathbf{L}) = ||\mathbf{Y} - \mathbf{KCL}||_F^2 + \lambda_1 \langle \mathbf{C}'\mathbf{KC}, \mathbf{L}\rangle_F + \lambda_2 \Omega(\mathbf{L}), \tag{5.1}$$

where $\mathbf{Y}$ is the $n \times m$ output data matrix, $\mathbf{K}$ is the $n \times n$ input-kernel Gram matrix, $\mathbf{L}$ is the $m \times m$ output-kernel Gram matrix, $\mathbf{C}$ is the $n \times m$ parameters matrix and $\Omega(.)$ is the regularizer on $\mathbf{L}$ and $\langle \mathbf{A}, \mathbf{B}\rangle_F = tr(\mathbf{A}'\mathbf{B})$.

We see that in (5.1) we use two regularization terms with λ_1 and λ_2 as the regularization parameters. Here I'll show that in fact you only need (and should use) one regularization parameter because the second can be absorbed into the relative scaling of the $\mathbf{L}$ and $\mathbf{C}$ that we learn.

To begin with, let's put $\Omega(\mathbf{L}) = ||\mathbf{L}||_F^2$ so that the minimization problem is

$$\min_{\mathbf{C},\mathbf{L}} J(\mathbf{C},\mathbf{L}) = ||\mathbf{Y} - \mathbf{KCL}||_F^2 + \lambda_1 \langle \mathbf{C}'\mathbf{KC}, \mathbf{L}\rangle_F + \lambda_2 ||\mathbf{L}||_F^2 \tag{5.2}$$

First, we observe that $\lambda_2 ||\mathbf{L}||_F^2 = ||\sqrt{\lambda_2}\,\mathbf{L}||_F^2$. We introduce the following change of variables $\sqrt{\lambda_2}\,\mathbf{L} = \widetilde{\mathbf{L}}$ and $\mathbf{C} = \sqrt{\lambda_2}\widetilde{\mathbf{C}}$ where the matrices with tildes are simply scaled versions of the original matrices. Using these we can rewrite problem (5.2) as

$$\begin{aligned} \min_{\widetilde{\mathbf{C}},\widetilde{\mathbf{L}}} J(\widetilde{\mathbf{C}},\widetilde{\mathbf{L}}) &= ||\mathbf{Y} - \mathbf{K}\sqrt{\lambda_2}\widetilde{\mathbf{C}}\frac{1}{\sqrt{\lambda_2}}\widetilde{\mathbf{L}}||_F^2 + \lambda_1 \langle \sqrt{\lambda_2}\widetilde{\mathbf{C}}\mathbf{K}\sqrt{\lambda_2}\widetilde{\mathbf{C}}, \frac{1}{\sqrt{\lambda_2}}\widetilde{\mathbf{L}}\rangle_F + ||\widetilde{\mathbf{L}}||_F^2 \\ &= ||\mathbf{Y} - \mathbf{K}\widetilde{\mathbf{C}}\widetilde{\mathbf{L}}||_F^2 + \lambda_1\sqrt{\lambda_2}\,\langle \widetilde{\mathbf{C}}\mathbf{K}\widetilde{\mathbf{C}}, \widetilde{\mathbf{L}}\rangle_F + ||\widetilde{\mathbf{L}}||_F^2, \end{aligned} \tag{5.3}$$

where the 2nd regularization parameter λ_2 has been absorbed into the first regularization parameter and the scaling of the $\mathbf{C}$ and $\mathbf{L}$ matrices.

From this we see that we can fix λ_2 arbitrarily (and hence for example set it to $\lambda_2 = 1$) and only grid search for λ_1 to find the optimal combination of the parameter and output-kernel matrices. If we changed the value of λ_2 we could get the same regularization path by adjusting the λ_1 grid accordingly. The minimizing solutions $\widetilde{\mathbf{C}}$ and $\widetilde{\mathbf{L}}$ would be the scaled version of $\mathbf{C}$ and $\mathbf{L}$ but would yield the same objective values $J(.)$.

In consequence, not only we *can* drop the second regularization parameter but we *should* drop it (unless we fix the scale of $\mathbf{C}$). Otherwise, for every combination λ_1 and λ_2 we can find a combination $\widetilde{\lambda}_1$, $\widetilde{\lambda}_2$ which will yield the same minimum of the objective value with different scalings of the learned matrices $\mathbf{C}$ and $\mathbf{L}$.

6 Grids for hyperparameters

How to derive reasonable grids for tunning hyperparameters?

6.1 Lasso example

The Lasso minimisation problem is

$$\arg\min_{w} J(\mathbf{w}) := \frac{1}{n} \frac{||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2}{2\lambda} + ||\mathbf{w}||_1 \tag{6.1}$$

For $\lambda \to \infty$ the parameters $\mathbf{w}$ are shrank towards zero so that eventually (for big enough λ) they are all zero. We can find λ_{max} as the smallest value of λ for which we get $\mathbf{w} = 0$.

The logic comes form the coordinate descent optimisation strategy. We assume we're at a point where all the other parameters have already been shrank (thresholded) to zero and we solve for the last remaining w_i from

$$\begin{aligned}
\arg\min_{w_i} J(w_i) &:= \frac{1}{n} \frac{||\mathbf{y} - \mathbf{X}^{(-i)}\mathbf{w}^{(-i)} - \mathbf{X}_{:i} w_i||_2^2}{2\lambda} + ||\mathbf{w}||_1 \\
&= \frac{1}{n} \frac{||\mathbf{y} - \mathbf{X}_{:i} w_i||_2^2}{2\lambda} + |w_i|, \qquad (6.2)
\end{aligned}$$

where the equivalence comes from the assumption $\mathbf{w}^{(-i)} = 0$.

We solve for w_i as usual by equating to gradient to zero

$$\frac{\partial J(w_i)}{\partial w_i} = \frac{-\mathbf{X}_{:i}\mathbf{y} + \mathbf{X}'_{:i}\mathbf{X}_{:i} w_i}{n\lambda} + \frac{\partial |w_i|}{\partial w_i} = 0 \tag{6.3}$$

From which we get

$$w_i = \begin{cases} \frac{\mathbf{X}_{:i}\mathbf{y} - n\lambda}{\mathbf{X}'_{:i}\mathbf{X}_{:i}} & \text{for } w_i > 0 \\ \frac{\mathbf{X}_{:i}\mathbf{y} + n\lambda}{\mathbf{X}'_{:i}\mathbf{X}_{:i}} & \text{for } w_i < 0 \\ \frac{\mathbf{X}_{:i}\mathbf{y} - [-n\lambda; n\lambda]}{\mathbf{X}'_{:i}\mathbf{X}_{:i}} & \text{for } w_i = 0 \end{cases} \tag{6.4}$$

From this we see that $w_i = 0$ if $|\mathbf{X}_{:i}\mathbf{y}| < n\lambda$ and therefore we set $\lambda_{max} = \max |\mathbf{X}_{:i}\mathbf{y}|/n$

7 Trace norm regularization

This uses a lot from [1].

Trace norm (or nuclear norm) regularization is often times used as convex relaxation of rank constraint. For $(m \times n)$ matrix $\mathbf{X}$ the trace norm is

$$||\mathbf{X}||_* = tr(\sqrt{\mathbf{X}\mathbf{X}'}) = \sum_i^{min\{m,n\}} \sigma_i, \tag{7.1}$$

where σ_i are the singular values of $\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Theta}'$ so that the $p = min\{m,n\}$ singular values σ_i are on the diagonal of the $(m \times n)$ matrix $\mathbf{\Sigma}$, and $\mathbf{\Phi} \in \mathbb{R}^{m\times m}$, $\mathbf{\Theta} \in \mathbb{R}^{n\times n}$ are unitary so that $\mathbf{\Phi}\mathbf{\Phi}' = \mathbf{\Phi}'\mathbf{\Phi} = \mathbf{I}_m$ and $\mathbf{\Theta}\mathbf{\Theta}' = \mathbf{\Theta}'\mathbf{\Theta} = \mathbf{I}_n$.

The trace norm can be seen as the ℓ_1 norm of the singular values of the matrix.

Proof:

$$||\mathbf{X}||_* = tr(\sqrt{\mathbf{X}\mathbf{X}'}) = tr(\sqrt{\mathbf{\Phi}\mathbf{\Sigma}^2\mathbf{\Phi}'}) = tr(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}') = tr(\mathbf{\Sigma}) = \sum_i^{min\{m,n\}} \sigma_i,$$

where $\mathbf{X}\mathbf{X}' = \mathbf{\Phi}\mathbf{\Sigma}^2\mathbf{\Phi}'$ is the eigenvalue decomposition.

There is an important lemma, useful for low-rank matrix factorization

Lemma: For any matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, $||\mathbf{X}||_* \leq t$ iff there exists $\mathbf{A} \in \mathbb{R}^{m\times m}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that

$$\begin{bmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{bmatrix} \succeq 0 \qquad \text{and} \qquad tr(\mathbf{A}) + tr(\mathbf{B}) \leq 2t$$

Proof: Every PSD matrix $\mathbf{Z} \succeq 0$ can be factorised as $\mathbf{Z} = \mathbf{\Phi}\mathbf{\Phi}'$. Therefore we can write

$$\begin{bmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{bmatrix} = \mathbf{\Phi}\mathbf{\Phi}' = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}' & \mathbf{V}' \end{bmatrix} = \begin{bmatrix} \mathbf{U}\mathbf{U}' & \mathbf{U}\mathbf{V}' \\ \mathbf{V}\mathbf{U}' & \mathbf{V}\mathbf{V}' \end{bmatrix}, \tag{7.2}$$

where $\mathbf{X} = \mathbf{U}\mathbf{V}'$, and $\mathbf{A} = \mathbf{U}\mathbf{U}'$ and $\mathbf{B} = \mathbf{V}\mathbf{V}'$. We now have

$$tr(\mathbf{A}) + tr(\mathbf{B}) = tr(\mathbf{U}\mathbf{U}') + tr(\mathbf{V}\mathbf{V}') = ||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2 = \sum_i \left(||\mathbf{U}_{:i}||_2^2 + ||\mathbf{V}_{:i}||_2^2\right) \leq 2t \tag{7.3}$$

To find the decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}'$ we use the singular value decomposition $\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Theta}'$ with the properties of postmultiplication by diagonal matrix and the outer product approach to matrix multiplication (see Math cheat-sheet)

$$\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Theta}' = \begin{bmatrix} \sigma_1\mathbf{\Phi}_{:1} & \cdots & \sigma_p\mathbf{\Phi}_{:p} \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}'_{1:} \\ \vdots \\ \mathbf{\Theta}'_{p:} \end{bmatrix} = \sum_i^p \sigma_i \mathbf{\Phi}_{:i}\mathbf{\Theta}_{:i} = \sum_i^r \sigma_i \mathbf{\Phi}_{:i}\mathbf{\Theta}_{:i} = \sum_i^r \mathbf{U}_{:i}\mathbf{V}_{:i} = \mathbf{U}\mathbf{V}', \tag{7.4}$$

where $r = rank(\mathbf{X}) \leq p$ is the number of non-zero singular values and $\mathbf{U} \in \mathbb{R}^{m\times r}$ with columns $\mathbf{U}_{:i} = \nu_i\mathbf{\Phi}_{:i}$, and $\mathbf{V} \in \mathbb{R}^{n\times r}$ with columns $\mathbf{V}_{:i} = \eta_i\mathbf{\Theta}_{:i}$ such that $\nu_i\eta_i = \sigma_i$.

From this we see that the decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}'$ is not unique since for any $\widetilde{\mathbf{U}}_{:i} = c_i\nu_i\mathbf{\Phi}_{:i}$, $\widetilde{\mathbf{V}}_{:i} = c_i^{-1}\eta_i\mathbf{\Theta}_{:i}$ we have $\sum_i^r \widetilde{\mathbf{U}}_{:i}\widetilde{\mathbf{V}}_{:i} = \sum_i^r c_i\nu_i\mathbf{\Phi}_{:i}c_i^{-1}\eta_i\mathbf{\Theta}_{:i} = \sum_i^r \sigma_i\mathbf{\Phi}_{:i}\mathbf{\Theta}_{:i} = \mathbf{X}$.

We now go back and want to check the validity of eq. (7.3). First, we want to find such decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}'$ that minimises the the sum of the ℓ_2 norms in (7.3)

$$\begin{aligned} \min_c J(c) \quad &:= \quad \sum_i \left(||\widetilde{\mathbf{U}}_{:i}||_2^2 + ||\widetilde{\mathbf{V}}_{:i}||_2^2\right) = \sum_i^r \left(||c_i\mathbf{U}_{:i}||_2^2 + ||c_i^{-1}\mathbf{V}_{:i}||_2^2\right) \\ &= \quad \sum_i^r \left(c_i^2||\mathbf{U}_{:i}||_2^2 + ||\mathbf{V}_{:i}||_2^2/c_i^2\right) \end{aligned}$$

We get for c_i (by taking the derivative and putting equal to zero)

$$\begin{aligned} 0 &= 2c_i||\mathbf{U}_{:i}||_2^2 - 2||\mathbf{V}_{:i}||_2^2/c_i^3 \\ c_i^2 &= \frac{||\mathbf{U}_{:i}||_2}{||\mathbf{V}_{:i}||_2} \end{aligned}$$

and therefore the minimal

$$\begin{aligned} J(c^*) &= \sum_i^r \left(||c_i^*\mathbf{U}_{:i}||_2^2 + ||c_i^{*-1}\mathbf{V}_{:i}||_2^2 = \sum_i^r \left(c_i^2||\mathbf{U}_{:i}||_2^2 + ||\mathbf{V}_{:i}||_2^2/c_i^2\right)\right. \\ &= \sum_i^r \left(\frac{||\mathbf{U}_{:i}||_2}{||\mathbf{V}_{:i}||_2}||\mathbf{U}_{:i}||_2^2 + ||\mathbf{V}_{:i}||_2^2\frac{||\mathbf{V}_{:i}||_2}{||\mathbf{U}_{:i}||_2}\right) = 2\sum_i^r ||\mathbf{U}_{:i}||_2||\mathbf{V}_{:i}||_2 \\ &= 2\sum_i^r \nu_i\eta_i = 2\sum_i^r \sigma_i = 2||\mathbf{X}||_* = 2t \end{aligned}$$

where we used the unitarity of $\mathbf{\Phi}$ and $\mathbf{\Theta}$.

To summarise: We can decompose a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ of rank r as $\mathbf{X} = \mathbf{U}\mathbf{V}'$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$. However, this decomposition is not unique since for any vector $\mathbf{c}$ and matrices with columns $\widetilde{\mathbf{U}}_{:i} = \sum_i c_i\mathbf{U}_{:i}$ and $\widetilde{\mathbf{V}}_{:i} = \sum_i c_i^{-1}\mathbf{V}_{:i}$ we also have $\mathbf{X} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}'$. The minimisation problem with respect to $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{U}\mathbf{V}' = \mathbf{X}$

$$\min_{U,V|X=UV} J(\mathbf{U},\mathbf{V}) := \frac{1}{2}\left(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2\right) \tag{7.5}$$

reaches its optimum for matrices with columns $\mathbf{U}^*_{:i} = \nu_i\mathbf{\Phi}_{:i}$ and $\mathbf{V}^*_{:i} = \eta_i\mathbf{\Theta}_{:i}$, where $\nu_i = \eta_i = \sqrt{\sigma_i}$ coming from the SVD $\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Theta}'$ and the minimal value is

$$J(\mathbf{U}^*,\mathbf{V}^*) = ||\mathbf{X}||_* \tag{7.6}$$

Proof:

$$\begin{aligned} J(U,V) &= \frac{1}{2}\left(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2\right) = \frac{1}{2}\sum_i^r \left(||\mathbf{U}_{:i}||_2^2 + ||\mathbf{V}_{:i}||_2^2\right) \\ &= \frac{1}{2}\sum_i^r \left(||\nu_i\mathbf{\Phi}_{:i}||_2^2 + ||\eta_i\mathbf{\Theta}_{:i}||_2^2\right) = \frac{1}{2}\sum_i^r \left(\nu_i^2 + \eta_i^2\right) \end{aligned}$$

$$\min_{\nu,\eta|\nu_i\eta_i=\sigma_i} \frac{1}{2}\sum_i^r \left(\nu_i^2 + \eta_i^2\right) = \min_\nu \frac{1}{2}\sum_i^r \left(\nu_i^2 + \sigma_i^2/\nu_i^2\right) \tag{7.7}$$

From which we have for the minimising $2\nu_i^* - 2\sigma_i^2/\nu_i^{*3} = 0$ and therefore $\nu_i^{*2} = \sigma_i = \eta_i^{*2}$ (from $\nu_i^* = \sqrt{\sigma_i} = \eta_i^*$). In result the minimum of $J()$ is attained at

$$J(\mathbf{U}^*,\mathbf{V}^*) = \frac{1}{2}\sum_i^r \left(\nu_i^{*2} + \eta_i^{*2}\right) = \frac{1}{2}\sum_i^r \left(\sigma_i + \sigma_i\right) = \sum_i^r \sigma_i = ||\mathbf{X}||_* \tag{7.8}$$

If we impose an additional constraint on $||\mathbf{V}_{:i}||_2^2 = 1$, we get $\eta_i = 1$ and therefore $\nu_i = \sigma_i$. The minimum of $J()$ is than attained

$$J(\mathbf{U}^*,\mathbf{V}^*) = \frac{1}{2}\sum_i^r \left(\nu_i^{*2} + \eta_i^{*2}\right) = \frac{1}{2}\sum_i^r \left(\sigma_i^2 + 1\right) = \frac{||\mathbf{X}||_F^2 + rank(\mathbf{X})}{2} \tag{7.9}$$

References

[1] N. Srebro and T. S. Jaakkola, Maximum-Margin Matrix Factorization, in NIPS, 2004.

8 Subgradient and subdifferential

Definition : A vector $\mathbf{v} \in \mathbb{R}^n$ is a **subgradient** of (not necessarily convex) function $f : \mathbb{R}^n \to \mathbb{R}$ at point $\mathbf{x}_0 \in dom f$ if for all $\mathbf{x} \in dom f$

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{v}^T(\mathbf{x} - \mathbf{x}_0) \tag{8.1}$$

Note: If f is convex and differentiable at point $\mathbf{x}_0$ than the subgradient is equal to the gradient $\mathbf{v} = \nabla f(\mathbf{x}_0)$.

Definition : The set of all subgradients of function f at point $\mathbf{x}_0$ is called the **subdifferential** and denoted $\partial f(\mathbf{x}_0)$

$$\partial f(\mathbf{x}_0) = \{\mathbf{v} | f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{v}^T(\mathbf{x} - \mathbf{x}_0)\} \quad \text{for all } \mathbf{x} \in dom f \tag{8.2}$$

Note: The subdifferential $\partial f(\mathbf{x})$ is a closed convex set (though may be empty)

Note: For a convex subdifferentiable function f the standard optimality condition for a minimum $f(\mathbf{x}^*) = \inf_x f(\mathbf{x}) \Leftrightarrow 0 = \nabla f(\mathbf{x})$ changes to $f(\mathbf{x}^*) = \inf_x f(\mathbf{x}) \Leftrightarrow 0 \in \partial f(\mathbf{x})$.

8.1 Absolute value $|x|$

The absolute value $f(x) = |x|$ is differentiable at all points of its domain $dom f = \mathbb{R}$ with the gradient $\nabla f(x) = sign(x)$ except at the point $x = 0$.

At $x = 0$ we use the subgradient definition (8.1) and get

$$\begin{aligned} f(x) - f(0) &\geq v\,(x - 0) \\ |x| &\geq v\,x \quad \text{for all } x \in dom f, \end{aligned} \tag{8.3}$$

which is satisfied if and only if $v \in [-1, 1]$.

The subdifferential of $f(x) = |x|$ is therefore

$$\partial f(x) = \begin{cases} sign(x) & \text{if } x \neq 0 \\ \{v : v \in [-1, 1]\} & \text{if } x = 0 \end{cases} \tag{8.4}$$

8.2 ℓ_2 norm $||\mathbf{x}||_2$

The gradient of the ℓ_2 norm $f(\mathbf{x}) = ||\mathbf{x}||_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ at all points of its domain $dom f = \mathbb{R}^n$ except at the point $\mathbf{x} = 0$ is $\nabla f(\mathbf{x}) = \mathbf{x}/||\mathbf{x}||_2$.

At $\mathbf{x} = 0$ we use the subgradient definition (8.1) and get

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{0}) &\geq \mathbf{v}^T\,(\mathbf{x} - \mathbf{0}) \\ ||\mathbf{x}||_2 &\geq \mathbf{v}^T\,\mathbf{x} \quad \text{for all } \mathbf{x} \in dom f \\ ||\mathbf{x}||_2 &\geq ||\mathbf{v}||_2\,||\mathbf{x}||_2 \quad (\text{from Cauchy-Schwarz inequality } \mathbf{v}^T\,\mathbf{x} \leq ||\mathbf{v}||_2||\mathbf{x}||_2) \\ 1 &\geq ||\mathbf{v}||_2 \end{aligned} \tag{8.5}$$

The subdifferential of $f(\mathbf{x}) = ||\mathbf{x}||_2$ is therefore

$$\partial f(\mathbf{x}) = \begin{cases} \mathbf{x}/||\mathbf{x}||_2 & \text{if } \mathbf{x} \neq 0 \\ \{\mathbf{v} : ||\mathbf{v}||_2 \leq 1\} & \text{if } \mathbf{x} = 0 \end{cases} \tag{8.6}$$

8.3 Generalised ℓ_2 norm

The gradient of the generalised ℓ_2 norm $f(\mathbf{x}) = ||\mathbf{Ax}||_2 = \sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{Ax}}$ at all points of its domain $dom f = \mathbb{R}^n$ except at the point $\mathbf{x} = 0$ is $\nabla f(\mathbf{x}) = \mathbf{A}^T\mathbf{Ax}/||\mathbf{Ax}||_2$.

At $\mathbf{x} = 0$ we use the subgradient definition (8.1)

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{0}) &\geq \mathbf{v}^T(\mathbf{x} - \mathbf{0}) \\ ||\mathbf{Ax}||_2 &\geq \mathbf{v}^T\mathbf{x} \quad \text{for all } \mathbf{x} \in dom f \end{aligned} \tag{8.7}$$

For the left side of the inequality we do the eigen-decomposition

$$||\mathbf{Ax}||_2 = (\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})^{1/2} = (\mathbf{x}^T \sum_i \mathbf{w}_i\lambda_i\mathbf{w}_i^T\mathbf{x})^{1/2} \tag{8.8}$$

Because $\mathbf{A}^T\mathbf{A}$ is a PSD matrix, the set of eigenvectors $\mathbf{w}_i$ forms an orthogonal basis in $\mathbf{R}^n$. We can therefore express any vector $\mathbf{x}$ in this basis as $\mathbf{x} = \sum_i c_{(x)i}\mathbf{w}_i$ and continue from eq. (8.8)

$$(\sum_j c_{(x)j}\mathbf{w}_j^T \sum_i \mathbf{w}_i\lambda_i\mathbf{w}_i^T \sum_l c_{(x)l}\mathbf{w}_l)^{1/2} = (\sum_i c_{(x)i}\lambda_i c_{(x)i})^{1/2} = (\sum_i c_{(x)i}^2\lambda_i)^{1/2} \tag{8.9}$$

where the sums and $\mathbf{w}$'s eliminated due to the orthogonality of the $\mathbf{w}$'s (that is $\mathbf{w}_i^T\mathbf{w}_j = 0$ if $i \neq j$ and $\mathbf{w}_i^T\mathbf{w}_i = 1$).

We use the same eigenbasis for the right side of the inequality where we get

$$\mathbf{v}^T\mathbf{x} = \sum_i c_{(v)i}\mathbf{w}_i^T \sum_j c_{(x)j}\mathbf{w}_j = \sum_i c_{(v)i}\mathbf{w}_i^T c_{(x)i}\mathbf{w}_i = \sum_i c_{(v)i}c_{(x)i} \tag{8.10}$$

We introduce a change of variable $z_i = c_{(x)i}\sqrt{\lambda_i}$ and $s_i = c_{(v)i}/\sqrt{\lambda_i}$ and use these after plugging the expressions from eq. (8.9) and eq. (8.10) back to the inequality (8.7).

$$\begin{aligned} (\sum_i c_{(x)i}^2\lambda_i)^{1/2} &\geq \sum_i c_{(v)i}c_{(x)i} \\ (\sum_i z_i^2)^{1/2} &\geq \sum_i s_i z_i \\ ||\mathbf{z}||_2 &\geq \mathbf{s}^T\mathbf{z} \\ 1 &\geq ||\mathbf{s}||_2 \quad \text{from C-S inequality as in eq. (8.5)} \\ 1 &\geq (\sum_i c_{(v)i}^2/\lambda_i)^{1/2} \end{aligned} \tag{8.11}$$

The subdifferential of $f(\mathbf{x}) = ||\mathbf{Ax}||_2$ is therefore

$$\partial f(\mathbf{x}) = \begin{cases} \mathbf{A}^T\mathbf{Ax}/||\mathbf{Ax}||_2 & \text{if } \mathbf{x} \neq 0 \\ \{\mathbf{v} : \mathbf{v} = \sum_i c_i\mathbf{w}_i,\ (\sum_i c_i^2/\lambda_i)^{1/2} \leq 1,\ \mathbf{A}^T\mathbf{A} = \sum_i \mathbf{w}_i\lambda_i\mathbf{w}_i^T\} & \text{if } \mathbf{x} = 0 \end{cases} \tag{8.12}$$

9 Training with noise is equivalent to Tikhonov regularisation

Bishop has a paper on this [1] which deals with NNs and derives the equivalence for non-linear functions in a rather elaborate way through Taylor expansion. This is to provide some intuition using simple linear regression.

For simplicity, here below I use a very specific example of data and perturbations. But the intuition should extend to more complex cases.

Let's have a dataset of n input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in (\mathbb{R}^n \times \mathbb{R})\}_{i=1}^n$. Note that the dimensionality of the input space is the same as the number of samples so that the design matrix $\mathbf{X}$ is square $n \times n$. The aim is to learn a function $f : \mathbb{R}^n \to \mathbb{R}$ mapping the inputs to the outputs and here we stick to the class of linear functions in the form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$, where $\mathbf{w}$ is the n-dimensional parameters vector.

The classical ordinary least squares method finds the optimal $\widehat{\mathbf{w}}$ from the minimisation problem

$$\widehat{\mathbf{w}} = \arg\min_{w} ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 \tag{9.1}$$

with the analytical form solution

$$(\mathbf{X}^T\mathbf{X})\,\widehat{\mathbf{w}} = \mathbf{X}^T\mathbf{y} \tag{9.2}$$

Bishop discusses the option to perturb the input samples $\mathbf{x}$ by a random noise ξ to stabilise the learning. The noise is a n-dimensinal random variable typically centred at zero ($E(\xi) = 0$) and is uncorrelated between input dimensions $E(\xi_i\xi_j) = \mu\,\delta_{ij}$. He does not speak about symmetry but it helps the intuition in the example below.

Next we do the perturbations to our data. Here, I generate a random noise vector ξ. For each instance the perturbations shall impact only single dimension so that the perturbed inputs matrix is $\mathbf{X}_1 = \mathbf{X} + \Xi$, where Ξ is the diagonal matrix constructed from ξ. For the sake of symmetry, I will perturb the data once more with the negative noise as $\mathbf{X}_2 = \mathbf{X} - \Xi$. Note: This seems a bit as a cheat here (I need it to show what I want) but remember that normally I would do many more perturbations. So each dimension would be perturbed many times, with some positive some negative perturbations with overall $E(\xi_i) = 0$.

The least square problem with these augmented data now amounts to

$$\widehat{\mathbf{w}}_A = \arg\min_{w} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{w} \right\|_2^2 \tag{9.3}$$

and has the OLS solution

$$\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \right) \widehat{\mathbf{w}}_A = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \end{bmatrix} \tag{9.4}$$

Now some linear algebra to rewrite the above

$$\begin{aligned} \left(\mathbf{X}^T\mathbf{X} + \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_2^T\mathbf{X}_2\right)\widehat{\mathbf{w}}_A &= \mathbf{X}^T\mathbf{y} + \mathbf{X}_1^T\mathbf{y} + \mathbf{X}_2^T\mathbf{y} \\ \left(\mathbf{X}^T\mathbf{X} + (\mathbf{X}+\Xi)^T(\mathbf{X}+\Xi) + (\mathbf{X}-\Xi)^T(\mathbf{X}-\Xi)\right)\widehat{\mathbf{w}}_A &= \mathbf{X}^T\mathbf{y} + (\mathbf{X}+\Xi)^T\mathbf{y} + (\mathbf{X}-\Xi)^T\mathbf{y} \\ \left(\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X} + \mathbf{X}^T\Xi + \Xi^T\mathbf{X} + \Xi^T\boldsymbol{\Xi} + \mathbf{X}^T\mathbf{X} - \mathbf{X}^T\Xi - \Xi^T\mathbf{X} + \Xi^T\Xi\right)\widehat{\mathbf{w}}_A &= 3\mathbf{X}^T\mathbf{y} + \Xi^T\mathbf{y} - \Xi^T\mathbf{y} \\ \left(3\mathbf{X}^T\mathbf{X} + 2\Xi^T\Xi\right)\widehat{\mathbf{w}}_A &= 3\mathbf{X}^T\mathbf{y} \qquad (9.5) \end{aligned}$$

The critical point in the above elimination was the positive and negative application of the perturbations to the same set of instances. Without those, this simple example would fall apart.

However, what I show is an example of 1 experiment over 1 data sample with 1 random noise vector. To be rigorous, this analysis should be done in expectations over the sampling and noise distributions. Then the properties of the noise (centered, uncorrelated) should kick in and I would imagine should boil down to something similar. Showing this is too much work if I just need some simple intuition so ... not done here.

Next consider the regularised least squares problem with the generalised norm $||\mathbf{w}||_Q^2 = \langle \mathbf{w}, \mathbf{Q}\mathbf{w} \rangle$

$$\widehat{\mathbf{w}}_R = \arg\min_w ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + ||\mathbf{w}||_Q^2, \tag{9.6}$$

with the analytical form solution

$$(\mathbf{X}^T\mathbf{X} + \mathbf{Q})\,\widehat{\mathbf{w}}_R = \mathbf{X}^T\mathbf{y} \tag{9.7}$$

Clearly, with $\mathbf{Q} = 2/3\,\Xi^T\Xi$ the minimising solutions of the regularised problem (9.6) and the augmented data problem (9.3) coincide.

For the classical ridge regression problem

$$\widehat{\mathbf{w}}_R = \arg\min_w ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \frac{3}{2}\,\lambda\,||\mathbf{w}||_2^2, \tag{9.8}$$

the corresponding data-augmentation matrix would be $\Xi = \sqrt{\lambda} I$. The constant 3/2 is due to the specific data-augmentation strategy consisting of the 3 data parts.

More complicated augmentation strategies and particularly non-stochastic ones (or without the noise being zero centered etc.) may not translate this simply but the intuition should be similar.

References

[1] C. M. Bishop: Training with noise is equivalent to Tikhonov regularization, Neural Computation 1995.

10 Early stopping equivalence to Tikhonov regularization

Based on [1] and me.

This is to show the equivalence between early stopping and l2 regularization on a simple example of linear ridge regression.

The ridge regression problem

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}||\mathbf{X}\mathbf{w}-\mathbf{y}||_2^2 + \lambda\frac{1}{2}||\mathbf{w}||_2^2 \tag{10.1}$$

has a closed form solution

$$\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

The gradient descent approach involves the following steps

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \tau[\mathbf{X}^T(\mathbf{X}\mathbf{w}_{k-1}-\mathbf{y}) - \lambda\mathbf{w}_{k-1}] \ ,$$

where τ is the step size and the term in the bracket after is the gradient of the function being optimised (10.1) with respect to $\mathbf{w}$.

Let's explore the case where $\lambda = 0$ (not a ridge regression but an ordinary least squares problem) with the initial value $\mathbf{w}_0 = 0$.

$$\begin{aligned}
\mathbf{w}_k &= \mathbf{w}_{k-1} - \tau\mathbf{X}^T(\mathbf{X}\mathbf{w}_{k-1}-\mathbf{y}) \qquad \text{(from } \lambda = 0)\\
&= (\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})\mathbf{w}_{k-1} + \tau\mathbf{X}^T\mathbf{y}\\
&= (\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})((\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})\mathbf{w}_{k-2} + \tau\mathbf{X}^T\mathbf{y}) + \tau\mathbf{X}^T\mathbf{y}\\
&\vdots\\
&= (\mathbf{I}-\mathbf{X}^T\mathbf{X})^k\mathbf{w}_0 + \tau\sum_{j=0}^{k-1}(\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})^j\mathbf{X}^T\mathbf{y}\\
\mathbf{w}_k &= \tau\sum_{j=0}^{k-1}(\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})^j\mathbf{X}^T\mathbf{y} \qquad \text{(from } \mathbf{w}_0 = 0) \qquad (10.2)
\end{aligned}$$

Use the Neumann series results.

If the operator norm $||\mathbf{A}|| < 1$ we have

$$\sum_{j=0}^{\infty}\mathbf{A}^j = (\mathbf{I}-\mathbf{A})^{-1} \qquad \sum_{j=0}^{k-1}\mathbf{A}^j = (\mathbf{I}-\mathbf{A}^k)(\mathbf{I}-\mathbf{A})^{-1}$$

In particular for $||\mathbf{I}-\mathbf{A}|| < 1$

$$\sum_{j=0}^{\infty}(\mathbf{I}-\mathbf{A})^j = \mathbf{A}^{-1} \qquad \sum_{j=0}^{k-1}(\mathbf{I}-\mathbf{A})^j = (\mathbf{I}-(\mathbf{I}-\mathbf{A})^k)\mathbf{A}^{-1}$$

Hence if we continue the gradient descent (10.2) to infinity we get

$$\mathbf{w}_\infty = \tau\sum_{j=0}^{\infty}(\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})^j\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad \Rightarrow \textbf{OLS} \tag{10.3}$$

For a limited number of steps - **early stopping**

$$\mathbf{w}_k = \tau\sum_{j=0}^{k-1}(\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})^j\mathbf{X}^T\mathbf{y} = \left(\mathbf{I}-(\mathbf{I}-\tau\mathbf{X}^T\mathbf{X})^k\right)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{10.4}$$

To show the equivalence between early stopping and the regularization we need to show the equivalence between

$$\left(\mathbf{I} - (\mathbf{I} - \tau \mathbf{X}^T\mathbf{X})^k\right)(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \tag{10.5}$$

References

[1] MIT 9.520/6.860: Statistical Learning Theory and Applications, Fall 2017, Class 08: Iterative Regularization via Early Stopping `http://www.mit.edu/~9.520/fall17/Classes/early_stopping.html`

11 Some useful inequatlities

Markov's inequality *Let X be a non-negative random variable such that $P(X \geq 0) = 1$ and $P(X = 0) < 1$ and $EX < \infty$. Then for any $r > 0$*

$$P(X \geq r) \leq \frac{EX}{r} \quad \textit{and} \quad P(X \geq t\,EX) \leq \frac{1}{r}$$

Proof: Following [1]. For a given r it holds that

$$\begin{aligned} r\,I_{(X\geq r)} &\leq X \quad (I_{(X\geq r)} = 1 \text{ if } X \geq r \text{ and zero otherwise}) \\ E(r\,I_{(X\geq r)}) &\leq EX \quad (\text{by monotonicity of expectation}) \\ r\big(1\,P(X \geq r) + 0\,P(X < r)\big) &\leq EX \quad (\text{expanding the expectation}) \\ P(X \geq r) &\leq \frac{EX}{r} \quad \text{QED} \end{aligned}$$

Alternatively following [2]

$$\begin{aligned} P(X \geq r\,EX) &= \sum_{x \geq r\,EX} P(X = x) \\ &\leq \sum_{x \geq r\,EX} P(X = x)\frac{x}{r\,EX} \quad (\text{from } \frac{x}{r\,EX} \geq 1) \\ &\leq \sum_x P(X = x)\frac{x}{r\,EX} \\ &= E\Big(\frac{x}{r\,EX}\Big) = \frac{1}{r} \quad \text{QED} \end{aligned}$$

Chebyshev's inequality (generalised) *Let X be a random variable and $g(x)$ a non-negative function such that $P(g(X) \geq 0) = 1$ and $P(g(X) = 0) < 1$ and $Eg(X) < \infty$. Then for any $r > 0$*

$$P(g(X) \geq r) \leq \frac{Eg(X)}{r} \quad \textit{and} \quad P(g(X) \geq r\,Eg(X)) \leq \frac{1}{r}$$

Proof: Following [1] Let $f_X(x)$ be the probability density function of the r.v. X. For a given $r > 0$ it holds that

$$\begin{aligned} Eg(X) &= \int g(x) f_X(x) dx \\ &= \int_{x:g(x)\geq r} g(x) f_X(x) dx + \int_{x:g(x)<r} g(x) f_X(x) dx \\ &\geq \int_{x:g(x)\geq r} g(x) f_X(x) dx \\ &\geq \int_{x:g(x)\geq r} r\,f_X(x) dx \quad (\text{from } g(x) \geq r) \\ &= r\,P(g(X) \geq r) \quad \text{QED} \end{aligned}$$

Alternatively

$$\begin{aligned}
P(g(X) \geq r\,Eg(X)) &= \int_{x:g(x)\geq r\,Eg(X)} f_X(x)dx \\
&\leq \int_{x:g(x)\geq r\,Eg(X)} f_X(x)dx\frac{g(x)}{r\,Eg(X)} \quad \text{(from } \frac{g(x)}{r\,Eg(X)} \geq 1) \\
&\leq \int f_X(x)dx\frac{g(x)}{r\,Eg(X)} \\
&= E\Big(\frac{g(x)}{r\,Eg(X)}\Big) = \frac{1}{r} \quad \text{QED}
\end{aligned}$$

Example: Following [1] For X with $EX = \mu$ and $VarX = \sigma^2$ we have

$$P\Big(\frac{(X-\mu)^2}{\sigma^2} \geq r^2\Big) \leq \frac{1}{r^2}E\frac{(X-\mu)^2}{\sigma^2} = \frac{1}{r^2}$$

and therefore we get the classical Chebyshev's bound on deviation from mean

$$P(|X-\mu| \geq r\sigma) \leq \frac{1}{r^2}$$

Chernoff's bound *Let X_i be a random variable and fix $r > 0$. Then for any $t > 0$*

$$P(X_i \geq r) \leq \frac{Ee^{tX_i}}{e^{tr}}$$

Let X be the sum of n independent random variables $X = \sum_i^n X_i$. Then for any $t, r > 0$

$$P(X \geq r) \leq \frac{E\prod_i^n e^{tX_i}}{e^{tr}}$$

The bound is found as

$$P(X_i \geq r) \leq \min_{t>0} \frac{Ee^{tX_i}}{e^{tr}}$$

Proof: Follows trivially from Chebyshev's inquality by setting $g(X_i) = e^{tX_i}$ and observing that $P(X_i \geq r) = P(e^{tX_i} \geq e^{tr})$. Similarly, the second part follows from setting $g(X) = e^{tX} = e^{t\sum X_i} = \prod_i^n e^{tX_i}$.

Hoeffding's lemma *Let X be a random variable wiht $EX = 0$ such that $P(a \leq X \leq b) = 1$. Then the following inequality for the moment generating function $M_X(t) = Ee^{tX}$ holds*

$$Ee^{tX} \leq \exp\frac{t^2(b-a)^2}{8}$$

Proof: Following [2]

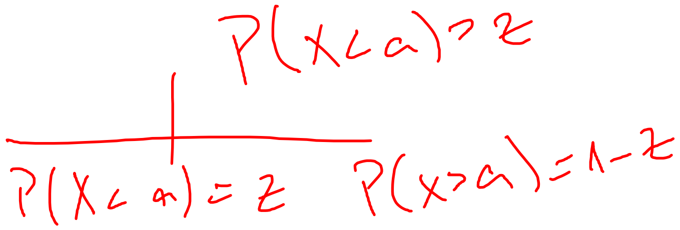
$$
\begin{aligned}
e^{tX} &\leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb} \qquad \text{(by convexity of } e^{tX}\text{)}\\
Ee^{tX} &\leq E\Big(\frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb}\Big) \qquad \text{(by monotonocity of } E\text{)}\\
&= \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} \qquad \text{(from } E(X)=0\text{)}\\
&= e^{ta}\Big(\frac{b}{b-a} + \frac{-a}{b-a}e^{tb-ta}\Big)\\
&= e^{ta}\frac{-a}{b-a}\Big(\frac{-b}{a} + e^{t(b-a)}\Big)\\
&= e^{ta}\frac{-a}{b-a}\Big(\frac{b-a}{-a} + \frac{a}{-a} + e^{t(b-a)}\Big)\\
&= e^{ta}u\Big(\frac{1}{u} - 1 + e^{t(b-a)}\Big), \qquad \text{were } u = \frac{-a}{b-a}\\
&= e^{ta} - e^{ta}u + e^{ta}ue^{t(b-a)}\\
&= (1-u+ue^{t(b-a)})e^{tau/u}\\
&= (1-u+ue^{t(b-a)})e^{-tu(b-a)}\\
&= e^{\phi(t)}, \quad \text{where}\\
\phi(t) &= \log\Big((1-u+ue^{t(b-a)})e^{-tu(b-a)}\Big)\\
&= -tu(b-a) + \log\big(1-u+ue^{t(b-a)}\big)\\
\phi(0) &= -0 + \log\big(1-u+u\big) = 0\\
\phi'(t) &= -u(b-a) + \frac{u(b-a)e^{t(b-a)}}{1-u+ue^{t(b-a)}}\\
\phi'(0) &= a + \frac{-a}{1-u+u} = 0\\
\phi''(t) &= \frac{-a(b-a)e^{t(b-a)}}{1-u+ue^{t(b-a)}} + \frac{ua(b-a)e^{2t(b-a)}}{(1-u+ue^{t(b-a)})^2}\\
\phi''(t) &= \frac{u(b-a)^2e^{t(b-a)}}{1-u+ue^{t(b-a)}}\frac{1-u+ue^{t(b-a)}-ue^{t(b-a)}}{1-u+ue^{t(b-a)}}\\
&= \frac{ue^{t(b-a)}}{1-u+ue^{t(b-a)}}\Big(1 - \frac{ue^{t(b-a)}}{1-u+ue^{t(b-a)}}\Big)(b-a)^2\\
&= z(1-z)(b-a)^2, \qquad \text{where } z = \frac{ue^{t(b-a)}}{1-u+ue^{t(b-a)}} = \frac{-ae^{t(b-a)}}{b-ae^{t(b-a)}}(b-a)^2 > 0\\
&\leq 1/4(b-a)^2 \qquad (\max z(1-z) = 1/4 \text{ at } z = 1/2)
\end{aligned}
$$

Using Taylor's approximation with Lagrange form of remainder we know that for every $t > 0$ there exist $s \in [0, t]$ such that

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2}t^2\phi''(s) \leq \frac{1}{2}t^2 1/4(b-a)^2 = \frac{t^2(b-a)^2}{8}$$

So that finally

$$Ee^{tX} \leq e^{(}\phi(t)) \leq \exp\frac{t^2(b-a)^2}{8} \quad \text{QED}$$



Hoeffding's inequality *Let* $\{X_i\}_{i=1}^n$ *be independent random variable such that* $P(a_i \le X_i \le b_i) = 1$ *and* $S = \sum_{i=1}^n X_i$. *Then for any* $r > 0$

$$P(S - ES \ge r) \le \exp \frac{-2r^2}{\sum_{i=1}^n (b_i - a_i)^2}$$
$$P(S - ES \le -r) \le \exp \frac{-2r^2}{\sum_{i=1}^n (b_i - a_i)^2}$$

Proof: Following [2] Using Chernoff's for any $t, r > 0$

$$\begin{aligned}
P(S - ES \ge r) &\le \frac{Ee^{t(S-ES)}}{e^{tr}} = \frac{Ee^{t(\sum_i^n X_i - E\sum_i^n X_i)}}{e^{tr}} \\
&= \prod_i^n \frac{Ee^{t(X_i - EX_i)}}{e^{tr}} \qquad \text{(by independence)} \\
&= \prod_i^n \frac{Ee^{t(Y_i)}}{e^{tr}} \qquad (Y_i = X_i - EX_i, P(a - EX_i \le Y_i \le b - EX_i) = 1) \\
&\le e^{-tr} \prod_i^n \exp \frac{(t^2(a_i - EX_i - b_i + EX_i)^2}{8} \qquad \text{(by Hoeffding's inequality)} \\
&= \exp\left(\frac{1}{8}t^2 \sum_i^n (a_i - b_i)^2 - tr\right)
\end{aligned}$$

Now find the minimum of this with respect to $t > 0$.

$$\begin{aligned}
L(t) &= \exp\left(\frac{1}{8}t^2 \sum_i^n (a_i - b_i)^2 - tr\right) \\
L'(t) &= \left(\frac{2}{8}t \sum_i^n (a_i - b_i)^2 - r\right) \exp\left(\frac{1}{8}t^2 (a_i - b_i)^2 - tr\right) \\
0 &= \left(\frac{2}{8}t \sum_i^n (a_i - b_i)^2 - r\right) \exp\left(\frac{1}{8}t^2 (a_i - b_i)^2 - tr\right) \\
t &= \frac{4r}{\sum_i^n (a_i - b_i)^2}
\end{aligned}$$

and plug this back

$$\begin{aligned}
P(S - ES \ge r) &\le \min_{t>0} \exp\left(\frac{1}{8}t^2 \sum_i^n (a_i - b_i)^2 - tr\right) \\
&= \exp\left(\frac{16r^2}{8(\sum_i^n (a_i - b_i)^2)^2} \sum_i^n (a_i - b_i)^2 - \frac{4r^2}{\sum_i^n (a_i - b_i)^2}\right) \\
&= \exp \frac{-2r^2}{\sum_i^n (a_i - b_i)^2} \qquad \text{QED}
\end{aligned}$$

Definition Martingale Difference: A sequence of random variables $V_1, V_2, \ldots$ is a martingale difference sequence with respect to $X_1, X_2, \ldots$ if for all $i > 0$ V_i is a function of $X_1, \ldots, X_i$ and

$$E(V_{i+1} | X_1, \ldots, X_i) = 0$$

References

[1] Casella, G., Berger, R. L.: Statistical Inference. Duxbury. 2002

[2] Mohri, M., Rostamizadeh, A., & Talwalkar, A.: Foundations of Machine Learning. MIT Press (2012)

12 Learning theory

12.1 Generalization

Have a sample S of n iid input-output pairs $S = \{(X_i, Y_i) \in (\mathcal{X}, \mathcal{Y})\}_{i=1}^n$ from a probability distribution $\mathcal{D}$ so that $S \sim \mathcal{D}^n$ (both unknown to us). For simplicity we begin with the classification case $\mathcal{Y} \in \{-1, 1\}$.

We want to find a prediction function $h : \mathcal{X} \to \{-1, 1\}$ which has a small probability of error

$$\text{Risk:} \qquad R(h) = P(h(X) \neq Y) = E\Big(I_{h(X) \neq Y}\Big) \tag{12.1}$$

Introduce

$$\text{Bayes hypotheses:} \qquad t = \arg\min_g R(h) \tag{12.2}$$

$$\text{Bayes risk:} \qquad R^* = R(t) = \min_g R(h) \tag{12.3}$$

$$\text{Bayes classifier:} \qquad t(X) = \arg\max_Y P(Y|X) \tag{12.4}$$

$$\text{Noise level:} \qquad s(X) = min\Big(P(Y = 1|X), 1 - P(Y = 1|X)\Big) \tag{12.5}$$

In the *deterministic (noiseless)* setting we have $P(s(X) = 0) = 1$, $P(t(X) \neq Y) = 0$ and $R^* = 0$.

In the *noisy* case the Bayes risk is equal to the expected (*average*) noise $E(s(X)) = R^*$.

The goal is to learn t preferably by minimising $R(h)$. Because we don't know the distribution $\mathcal{D}$, we cannot calculate the expectation needed to get the risk $R(h)$ associated with a hypothesis h. But we can calculate the average over the sample S ($|S| = n$)

$$\text{Empirical risk:} \qquad \widehat{R}_S(h) = \frac{1}{n} \sum_{i=1}^n \Big(I_{h(X_i) \neq Y_i}\Big) \tag{12.6}$$

and get a candidate hypotheses $h_S = \arg\min_g \widehat{R}_S(h)$. The problem is that if we search in an infinite space $\mathcal{H}$ of functions with infinite inputs (or at least equal to n), one can always (under very mild conditions on $\mathcal{D}$) construct a h_S with $\widehat{R}_S(h_S) = 0$ but $R(h_S) = 1$. So we need to control somehow (by prior knowledge) the class $\mathcal{H}$ to avoid over-fitting.

12.2 Bounds

When learning hypotheses in function space $h \in \mathcal{H}$ we define

$$\text{Candidate hypotheses:} \qquad h_S = \arg\min_{g \in \mathcal{H}} \widehat{R}_S(h) \tag{12.7}$$

$$\text{Best-in-class hypotheses:} \qquad h^* = \arg\min_{g \in \mathcal{H}} R(h) \tag{12.8}$$

(We can use $R(h^*) = \inf_{g \in \mathcal{H}} R(h)$ in the below if the minimum does not exist.)

What we we would like to know for our learned candidate h_S is the risk $R(h_S)$ but we don't know $\mathcal{D}$ so cannot. So we at least try to bound $R(h_S)$.

Useful decomposition is

$$R(h_S) - R(t) = \underbrace{\Big(R(h_S) - R(h^*)\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h^*) - R(t)\Big)}_{\text{approximation error}} \tag{12.9}$$

If $t \in \mathcal{H}$ then the approximation error is zero but we can't estimate it cause we don't know t. It approaches zero only if with increasing n we also grow the function class $\mathcal{H}$. Typically the focus is on the estimation error.

Another useful decomposition is

$$R(h_S) = \widehat{R}_(h_S) + \underbrace{\Big(R(h_S) - \widehat{R}_S(h_S)\Big)}_{\text{bound}} \tag{12.10}$$

So we may be generally interested in

$$\begin{aligned} \text{Error bound:} \quad & R(h_S) \leq \widehat{R}_S(h_S) + B(n, \mathcal{H}) \quad \text{(empirical risk instead of risk?)} & (12.11) \\ \text{Error bound relative to best-in-class:} \quad & R(h_S) \leq R(h^*) + B(n, \mathcal{H}) \quad \text{(is our algo optimal?)} & (12.12) \\ \text{Error bound relative to Bayes:} \quad & R(h_S) \leq R(t) + B(n, \mathcal{H}) \quad \text{(convergence to Bayes?)} & (12.13) \end{aligned}$$

12.3 Empirical error bound

Define a loss class as $\mathcal{F} = \{f : (x, y) \to I_{h(X) \neq Y} : h \in \mathcal{H}\}$. So the range of all the functionals $f \in \mathcal{F}$ is $\{0, 1\}$. Using f we can write

$$\begin{aligned} \text{Risk:} \quad & R(h) = P(h(X) \neq Y) = E\big(I_{h(X) \neq Y}\big) = E\big(f(X, Y)\big) = R(f) & (12.14) \\ \text{Empirical risk:} \quad & \widehat{R}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \Big(I_{h(X_i) \neq Y_i}\Big) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i) = \widehat{R}_S(f) & (12.15) \end{aligned}$$

As stated above, we are interested in bounding $R(f_n) - \widehat{R}_S(f_n)$. More specifically, we want to bound $P(R(f_n) - \widehat{R}_S(f_n) > \epsilon)$. Remember, that both risks are random variables (f_n depend on the sample S through h_S). Moreover, $D(f) = \big(R(f) - \widehat{R}_S(f)\big)$ is a random variable such that $D(f) = \sum_{i=1}^n D_i(f)$ where $D_i(f)$ are iid random vars defined as $D_i(f) = \big(R(f)/n - \widehat{R}_i(f)\big)$.

We use the iid random variables $Z = (X, Y)$ or $Z_i = (X_i, Y_i)$ here below

$$R(f) - \widehat{R}_S(f) = E\big(f(Z)\big) - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \tag{12.16}$$

By the strong law of large numbers we have

$$P\Big(\lim_{n \to \infty} \widehat{R}_S(f) - R(f) = 0\Big) = 1 \tag{12.17}$$

From the Hoeffding's inequality and by the fact that f is bounded we have

$$\begin{aligned} P\Big(\widehat{R}_S(f) - R(f) \leq -\epsilon\Big) &\leq \exp \frac{-2n^2\epsilon^2}{n(b-a)^2} & (12.18) \\ P\Big(R(f) - \widehat{R}_S(f) \geq \epsilon\Big) &\leq \exp \frac{-2n\epsilon^2}{(b-a)^2} = \delta & (12.19) \\ \epsilon^2 &= \frac{(b-a)^2 \log(\delta)}{-2n} & (12.20) \\ P\Big(R(f) - \widehat{R}_S(f) \geq (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big) &\leq \delta \qquad (-\log(\delta) = \log(1/\delta)) & (12.21) \\ P\Big(R(f) - \widehat{R}_S(f) \leq (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big) &\geq 1 - \delta & (12.22) \\ & & (12.23) \end{aligned}$$

Now, because $f(Z) \in \{0,1\}(= \{a,b\})$ for all Z, f we have with probability at least $1-\delta$

$$R(f) \leq \widehat{R}_S(f) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{12.24}$$

Now watch out! What this actually says is that for any fixed function $f \in \mathcal{F}$ we can select with probability at least $1-\delta$ a sample S for which the inequality holds. Put it differently, if we have a sample S, the inequality will hold only for some functions f but we don't know for which ones and neither any probability of selecting them. For a given f the empirical risk over different samples S will fluctuate around the risk according to the Hoeffding bound. But for a given S we can find f where the distance will be very large.

12.4 Uniform deviations

In the end, we create and algorithm which after seeing the sample S comes up with a candidate hypotheses h_S. But before seeing the data, we don't know what h_S will be. Nevertheless, for any one function f_n we can bound the difference between the risk and empirical risk by using uniform deviations in the form

$$R(f_n) - \widehat{R}_S(f_n) \leq \sup_{f \in \mathcal{F}} \big(R(f) - \widehat{R}_S(f)\big) \tag{12.25}$$

Consider a *bad* sample S_i for which the Hoeffding's bound for some function f_i does not hold, that is $R(f_i) - \widehat{R}_S(f_i) > \epsilon$. The probability of finding such a sample is $P(S_i) = P\big(R(f_i) - \widehat{R}_S(f_i) > \epsilon\big) \leq \delta$ (from Hoeffding's). For two functions f_1 and f_2 we have

$$P(S_1 \cup S_2) \leq P(S_1) + P(S_2) \leq 2\delta \tag{12.26}$$

extending to N functions in class $\mathcal{F}$ we have

$$P(\cup_i^N S_i) \leq \sum_i^N P(S_i) \leq N\delta \tag{12.27}$$

. In result we have

$$P(\exists f \in \{f_1, \ldots, f_N\} : R(f) - \widehat{R}_S(f) > \epsilon) \leq \sum_{i=1}^N P\big(R(f_i) - \widehat{R}_S(f_i) > \epsilon\big) \leq N \exp(-2n\epsilon^2) \tag{12.28}$$

which comes directly form Hoeffding's.

If I have finite hypotheses set $\mathcal{H}_N = g_1, \ldots, h_S$ we have

$$P\Big(\exists g \in \mathcal{H}_N : R(h) - \widehat{R}_S(h) \leq \sqrt{\frac{\log N + \log(1/\delta)}{2n}}\Big) \geq 1 - \delta \tag{12.29}$$

that is with probability at least $1-\delta$ we will select a sample for which

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{\log N + \log(1/\delta)}{2n}} \quad \text{for all } h \in \mathcal{H}_N \text{ simmultaneously} \tag{12.30}$$

Because this holds for all h it must also hold for the candidate and best-in-class hypotheses $h_S, h^* \in \mathcal{H}_N$.

12.5 Estimation error

From the above we have

$$R(h^*) \leq \widehat{R}_S(h^*) + \sqrt{\frac{\log N + \log(1/\delta)}{2n}} \tag{12.31}$$

we further know (by definition) that $\widehat{R}_S(h_S) \leq \widehat{R}_S(h^*)$ and hence $\widehat{R}_S(h^*) - \widehat{R}_S(h_S) \geq 0$

Hence for the estimation error we have

$$\begin{aligned} R(h_S) &= \underbrace{R(h_S) - R(h^*)}_{\text{estimation error}} + R(h^*) & (12.32) \\ R(h_S) &\leq \widehat{R}_S(h^*) - \widehat{R}_S(h_S) + \underbrace{R(h_S) - R(h^*)}_{\text{estimation error}} + R(h^*) & (12.33) \\ R(h_S) &\leq 2 \sup_{g \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| + R(h^*) & (12.34) \end{aligned}$$

We know the bound for all functions $h \in G_N$ is $\sqrt{\frac{\log N + \log(2/\delta)}{2n}}$ (the $2/\delta$ is there cause this is for the abs value of the difference, trivial to derive following the same steps) and hence with probability at least $1 - \delta$ we will select a sample such that the estimation error will be

$$R(h_S) \leq R(h^*) + 2\sqrt{\frac{\log N + \log(2/\delta)}{2n}} \tag{12.35}$$

12.6 Uncountable hypotheses set

If the set $\mathcal{H}$ is uncountable, we don't know $N = |\mathcal{H}|$ so the above does not work.

Instead we can look at the functions $f \in \mathcal{F}$ projected on the sample, that is

$$\mathcal{F}_S = \{\big(f(Z_1), \ldots, f(Z_n)\big) : f \in \mathcal{F}\} \tag{12.36}$$

In the classification case, the size of $|\mathcal{F}_S|$ is the number of different ways the data in the sample S can be classified which is always countable (for infinite set $\mathcal{F}$ this would be 2^n) and we call it

$$\text{Growth function:} \qquad S_{\mathcal{F}}(n) = S_{\mathcal{H}}(n) = \sup_S |\mathcal{F}_S| \tag{12.37}$$

In the finite case with $|\mathcal{H}| = N$ we have $S_{\mathcal{H}}(n) \leq N$

12.7 Rademacher complexity

In the above we had a random variable $Z = (X, Y) \sim \mathcal{D}$ and the aim was to find a hypotheses h which would have the lowest expected loss $I_{h(X) \neq Y}$. Observe that

$$I_{h(X) \neq Y} = \left\{ \begin{array}{lll} 1, & \text{if } (1,-1),(-1,1), & yh(x) = -1 \\ 0, & \text{if } (1,1),(-1,-1), & yh(x) = 1 \end{array} \right\} = \frac{1}{2} - \frac{Yh(X)}{2} \tag{12.38}$$

So instead of finding h by minimising $I_{h(X) \neq Y}$, we can equivalently maximize the correlation $Yh(X)$.

Instead of the random vars Y with an unknown distribution we can use the Rademacher random variables $\sigma \in \{-1, 1\}$ with a known distribution $P(\sigma = 1) = P(\sigma = -1) = 0.5$ (essentially a random noise as if $t = 0$). We then have the correlation of the hypotheses with the Rademacher variable $\sigma h(X)$.

We can calculate similar correlation for any bounded function ($g : \mathcal{Z} \to [a,b]$) and therefore as well for the functions $f \in \mathcal{F}$. For an iid data sample $S = \{Z_i\}_{i=1}^n \sim \mathcal{D}^n$, iid Rademacher sample $\{\sigma_i\}_{i=1}^n$ and an f, the empirical correlation is $\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)$.

For a sample S and a class of functions $\mathcal{F}$ we define

$$\text{Empirical rademacher complexity:} \qquad \widehat{\mathfrak{R}}_S(\mathcal{F}) = E_\sigma\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\Big] \tag{12.39}$$

In words, this expresses how well the function class $\mathcal{F}$ correlates with Rademacher random noise (any noise on average) over the given sample S. This describes the *richness* of the class $\mathcal{F}$. The richer it is, the better it can correlate with any random noise.

Instead for the richness of the class $\mathcal{F}$ with respect to a single sample S we would prefer the expected richness across all samples of the same size n

$$\text{(Expected) rademacher complexity:} \qquad \mathfrak{R}_n(\mathcal{F}) = E_S\Big[\widehat{\mathfrak{R}}_S(\mathcal{F})\Big] \tag{12.40}$$

Finally, we can use this similarly as the size of the countable function class $|N|$ in deriving the error bounds.

12.8 Rademacher bounds

For function class $\mathcal{F} = \{f : \mathcal{Z} \to [0,1]\}$ for any $\delta > 1$ each of the following holds with probability $1-\delta$ for all $f \in \mathcal{F}$

$$R(f) \leq \widehat{R}_S(f) + 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{12.41}$$

$$R(f) \leq \widehat{R}_S(f) + 2\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{12.42}$$

Proof is based on MdDiarmid's ineuqality.

Note that the 2nd bound depends only on the data!

If instead of $\mathcal{F}$ we would like to express these bounds in terms of the hypotheses set $\mathcal{H} = \{h : \mathcal{X} \to \{-1,+1\}\}$ we can show that $\widehat{\mathfrak{R}}_S(\mathcal{F}) = \frac{1}{2}\widehat{\mathfrak{R}}_S(\mathcal{H})$ and can plug this directly above.

However, the complexity may be still difficult to calculate (it's a maximization problem). Instead we may want to bound it by the Growth function which is essentially a combinatorial problem *Todo: the link for classifications problems.*

12.9 Regression

In regression $\mathcal{Y} = \mathbb{R}$ and the loss is typically $\ell(Z,h) = (Y - h(X))^2$ with $Z = (X,Y)$. As before we have risk $R(h) = E(\ell(Z,h))$ and empirical risk $\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i,h)$. Note that the squared error loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is unbounded from above. This makes the analysis difficult!

Let's assume the loss is non-negative and bounded from above by some $M > 0$ for starters ($\ell \in [0,M]$).

For a finite hypotheses set $|\mathcal{H}_N| = N$ we get directly from Hoeffding's and the union bound: with probability at least $1-\delta$ (we will select a sample for which)

$$R(h) \leq \widehat{R}_S(h) + M\sqrt{\frac{\log N + \log(1/\delta)}{2n}} \quad \text{for all } h \in \mathcal{H}_N \text{ simmultaneously} \tag{12.43}$$

If the loss class $\mathcal{L}\{\ell : (Z, h) \to \mathbb{R}\}$ is a l-lipchitz function of the prediction error $|h(X) - Y|$ and the error is bounded ($|h(X) - Y|_\infty \leq M$) we can calculate $\widehat{\mathfrak{R}}_S(\mathcal{H})$ and than by Talagrand's lemma we have

$$\widehat{\mathfrak{R}}_S(\mathcal{L}) \leq l\,\widehat{\mathfrak{R}}_S(\mathcal{H}) \tag{12.44}$$

References

[1] Bousquet, O., Boucheron, S., & Lugosi, G. : Introduction to Statistical Learning Theory (2004)

[2] Mohri, M., Rostamizadeh, A., & Talwalkar, A.: Foundations of Machine Learning. MIT Press. (2012)

13 Basics of prediction theory

Let $\mathbf{X} \in R^d$ be random vector input variable and $Y \in R$ be the random output variable. The learning task is to learn a function $f(\mathbf{X})$ to predict the values of Y. Any value $\hat{Y} \in R$ is a valid prediction so we need a measure of *goodness* of the prediction - the loss function. In regression, habitual loss function is the squared error. (But others such as absolute error etc. are possible.) So we want to learn a function f for which we can *expect* that the predictions will be as close as possible (in the Euclidean distance sense) to the true values. The expected value of the squared distance (the prediction mean squared error) $PMSE = E(err^2) = E(Y - f(\mathbf{X}))^2$ shall be minimal.

$$f^* = \arg\min_f E(Y - f(\mathbf{X}))^2 \tag{13.1}$$

The optimal such function is the conditional expectation $f^*(\mathbf{X}) = E(Y|\mathbf{X} = \mathbf{x}) = E(Y|\mathbf{X})$ (so it is a function of the r.v. $\mathbf{X}$).

Proof. Let $g(\mathbf{X})$ be an arbitrary function and get its PMSE

$$\begin{aligned}
E(Y - g(\mathbf{X}))^2 &= E\Big(Y - E(Y|\mathbf{X} = \mathbf{x}) + E(Y|\mathbf{X} = \mathbf{x}) - g(\mathbf{X})\Big)^2 \\
&= E\Big(Y - E(Y|\mathbf{X})\Big)^2 + 2E\Big(\big(Y - E(Y|\mathbf{X})\big)\big(E(Y|\mathbf{X}) - g(\mathbf{X})\big)\Big) + E\Big(E(Y|\mathbf{X}) - g(\mathbf{X})\Big)^2 \\
&= E\Big(Y - E(Y|\mathbf{X})\Big)^2 + E\Big(E(Y|\mathbf{X}) - g(\mathbf{X})\Big)^2,
\end{aligned}$$

We have no control over the first term. The 2nd term can be made zero if we put $g(\mathbf{X}) = E(Y|\mathbf{X})$. The middle term dropped out because

$$\begin{aligned}
& E\Big(\big(Y - E(Y|\mathbf{X})\big)\big(E(Y|\mathbf{X}) - g(\mathbf{X})\big)\Big) \\
= \;& E\Big(\big(Y - E(Y|\mathbf{X})\big)h(\mathbf{X})\Big), \quad \text{where } h(\mathbf{X}) = E(Y|\mathbf{X}) - g(\mathbf{X}) \\
= \;& E\Big(Yh(\mathbf{X})\Big) - E\Big(E(Y|\mathbf{X})h(\mathbf{X})\Big) \\
= \;& E\Big(Yh(\mathbf{X})\Big) - E\Big(E(Yh(\mathbf{X})|\mathbf{X})\Big) \qquad \text{because } E(Yh(\mathbf{X})|\mathbf{X}) = E(Y|\mathbf{X})h(\mathbf{X}) \\
= \;& E\Big(Yh(\mathbf{X})\Big) - E\Big(Yh(\mathbf{X})\Big) = 0 \qquad \text{because } E_{\mathbf{X}}(E_Y(Y|\mathbf{X})) = E(Y)
\end{aligned}$$

□

13.1 Prediction as linear projection

Assume we restrict the function f to the class of linear functions $f(\mathbf{X}) = \mathbf{a}'\mathbf{X}$. Function f for which the prediction error is orthogonal with the r.v. $\mathbf{X}$ is called the linear projection of Y on $\mathbf{X}$ and it has the smallest prediction MSE.

Proof. In the probability space we define inner product between two scalar r.v. $\langle Y, X\rangle = E(YX)$ and therefore two scalar r.v. are orthogonal if $E(YX) = 0$. If r.v. Y is orthogonal to all elements of vector r.v. $\mathbf{X}$ it is also orthogonal to any linear combination of those and therefore $E(Y\,\mathbf{h}'\mathbf{X}) = 0$ for arbitrary $\mathbf{h}$. The orthogonal projection of Y on $\mathbf{X}$ thus has $E\big((Y - \mathbf{a}'\mathbf{X})X_j\big) = 0$ for every element X_j of vector r.v. $\mathbf{X}$ and $E\big((Y - \mathbf{a}'\mathbf{X})\mathbf{h}'\mathbf{X}\big) = 0$ for any vector $\mathbf{h}$.

Let $\mathbf{a}'\mathbf{X}$ be the orthogonal projection of Y on $\mathbf{X}$ and $\mathbf{g}'\mathbf{X}$ be an arbitrary linear function. We get its PMSE

$$\begin{aligned}
E(Y-\mathbf{g}'\mathbf{X})^2 &= E\Big(Y-\mathbf{a}'\mathbf{X}+\mathbf{a}'\mathbf{X}-\mathbf{g}'\mathbf{X}\Big)^2 \\
&= E\Big(Y-\mathbf{a}'\mathbf{X}\Big)^2 + 2E\Big((Y-\mathbf{a}'\mathbf{X})(\mathbf{a}'\mathbf{X}-\mathbf{g}'\mathbf{X})\Big) + E\Big(\mathbf{a}'\mathbf{X}-\mathbf{g}'\mathbf{X}\Big)^2 \\
&= E\Big(Y-\mathbf{a}'\mathbf{X}\Big)^2 + E\Big(\mathbf{a}'\mathbf{X}-\mathbf{g}'\mathbf{X}\Big)^2,
\end{aligned}$$

where the middle term

$$\begin{aligned}
E\Big((Y-\mathbf{a}'\mathbf{X})(\mathbf{a}'\mathbf{X}-\mathbf{g}'\mathbf{X})\Big) &= E\Big((Y-\mathbf{a}'\mathbf{X})(\mathbf{a}'-\mathbf{g}')\mathbf{X}\Big) \\
&= 0 \qquad \text{from the orthogonality } E\Big((Y-\mathbf{a}'\mathbf{X})\mathbf{h}'\mathbf{X}\Big)=0
\end{aligned}$$

We have no control over the first term and the 2nd term can be made zero if we put $\mathbf{g}=\mathbf{a}$. □

To get the orthogonal projection parameter

$$\begin{aligned}
0 &= E\Big((Y-\mathbf{a}'\mathbf{X})X_j\Big), \qquad \forall j \in \mathbb{N}_d \\
E(YX_j) &= E(\mathbf{a}'\mathbf{X}X_j) \qquad \forall j \in \mathbb{N}_d \\
\big[E(YX_1),\, E(YX_2),\ldots,E(YX_d)\big] &= \big[E(\mathbf{a}'\mathbf{X}X_j),\, E(\mathbf{a}'\mathbf{X}X_2),\ldots,\mathbf{a}'E(\mathbf{X}X_d)\big] \\
\big[E(Y\mathbf{X}')\big] &= \mathbf{a}'\big[E(\mathbf{X}\mathbf{X}')\big] \\
\mathbf{a}' &= \big[E(Y\mathbf{X}')\big]\big[E(\mathbf{X}\mathbf{X}')\big]^{-1} \\
\mathbf{a} &= \big[E(\mathbf{X}\mathbf{X}')\big]^{-1}\big[E(\mathbf{X}Y)\big], \qquad (13.2)
\end{aligned}$$

where we put square brackets [.] around matrices.

If the inverse does not exist, a is not uniquely determined but $\mathbf{a}'\mathbf{X}$ is uniquely determined.

The prediction MSE of the orthogonal projection is

$$\begin{aligned}
PMSE(Y,\mathbf{a}'\mathbf{X}) = E(Y-\mathbf{a}'\mathbf{X})^2 &= E(YY) - 2E(Y\mathbf{X}'\mathbf{a}) + E(\mathbf{a}'\mathbf{X}\mathbf{X}'\mathbf{a}) \\
&= E(YY) - 2E(Y\mathbf{X}')\mathbf{a} + \mathbf{a}'E(\mathbf{X}\mathbf{X}')\mathbf{a} \\
&= E(YY) - 2E(Y\mathbf{X}')\big[E(\mathbf{X}\mathbf{X}')\big]^{-1}\big[E(\mathbf{X}Y)\big] \\
&\quad +\big[E(Y\mathbf{X}')\big]\big[E(\mathbf{X}\mathbf{X}')\big]^{-1}E(\mathbf{X}\mathbf{X}')\big[E(\mathbf{X}\mathbf{X}')\big]^{-1}\big[E(\mathbf{X}Y)\big] \\
&= E(YY) - \big[E(Y\mathbf{X}')\big]\big[E(\mathbf{X}\mathbf{X}')\big]^{-1}\big[E(\mathbf{X}Y)\big] \qquad (13.3)
\end{aligned}$$

13.2 Link to ordinary least squares

For a standard regression problem, we have got a set of observations $D=\{\mathbf{y}_i,\mathbf{x}_i : \mathbf{y}_i \in \mathcal{Y} = \mathbb{R}, \mathbf{x}_i \in \mathcal{X} = \mathbb{R}^d\}_{i=1}^n$ and we want to learn the linear function $f(\mathbf{x}_i)=\mathbf{b}'\mathbf{x_i}$ to predict $\mathbf{y}_i$. The OLS solution for the function parameters is

$$\mathbf{b}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \big(\sum_{i=1}^n \mathbf{x_i}\mathbf{x_i}'\big)^{-1}\big(\sum_{i=1}^n \mathbf{x_i}y_i\big) = \big(\frac{1}{T}\sum_{i=1}^n \mathbf{x_i}\mathbf{x_i}'\big)^{-1}\big(\frac{1}{T}\sum_{i=1}^n \mathbf{x_i}y_i\big), \qquad (13.4)$$

where $\mathbf{y}$ and $\mathbf{X}$ are the data vector and matrix with observations in rows. We see that OLS is constructed from the sample moments in the same way as linear projections is constructed from the population moments.

The vector of OLS sample residuals

$$\hat{\mathbf{u}}^{OLS} = \mathbf{y} - \mathbf{X}\mathbf{b}^{OLS} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y} - \mathbf{H_x}\mathbf{y} = \mathbf{M_x}\mathbf{y}, \qquad (13.5)$$

where $\mathbf{H_x} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ and $\mathbf{M_x} = (\mathbf{I}_n - \mathbf{H}_x)$ and both H_x and M_x are symmetric and idempotent. Also $\mathbf{M_x X} = 0$ so that $\hat{\mathbf{u}}'^{OLS}\mathbf{X} = \mathbf{y}'\mathbf{M_x X} = 0$ meaning that the residuals are orthogonal to the input variables.

The true errors $\mathbf{u} = \mathbf{y} - \mathbf{Xb}$ so that $\hat{\mathbf{u}}^{OLS} = \mathbf{M_x}(\mathbf{Xb} + \mathbf{u}) = \mathbf{M_x u}$.

For the parameters we have $\mathbf{b}^{OLS} = (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{Xb} + \mathbf{u}) = \mathbf{b} + (\mathbf{X'X})^{-1}\mathbf{X'u}$

13.3 Vector outputs

If the output is a random vector $\mathbf{Y} \in \mathbb{R}^m$ than the linear projection of $\mathbf{Y}$ on $\mathbf{X}$ is the linear map $\mathbf{AX}$ such that each element of the prediction error $\mathbf{Y} - \mathbf{AX}$ is orthogonal to every element of r.v. $\mathbf{X}$.

$$E\big((\mathbf{Y} - \mathbf{AX})\mathbf{X}'\big) = \mathbf{0} \tag{13.6}$$

From which we get for the projection coefficients

$$\begin{aligned} E\big((\mathbf{Y} - \mathbf{AX})\mathbf{X}'\big) &= \mathbf{0} \\ E\big(\mathbf{YX}'\big) &= E\big(\mathbf{AXX}'\big) \\ \mathbf{A} &= [E\big(\mathbf{YX}'\big)]\,[E\big(\mathbf{XX}'\big)]^{-1} \end{aligned} \tag{13.7}$$

And the PMSE is

$$PMSE(\mathbf{Y}, \mathbf{AX}) = E(\mathbf{YY}') - \big[E(\mathbf{YX}')\big]\big[E(\mathbf{XX}')\big]^{-1}\big[E(\mathbf{XY}')\big] \tag{13.8}$$

14 Linear regression basics

Based on [1, 2] and me.

14.1 Ordinary least squares

Classical linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i \in \mathbb{N}_n \ , \tag{14.1}$$

where $\mathbf{x}_i$ is vector of known constants, $\boldsymbol{\beta}$ are unknown coefficients and ϵ_i are i.i.d. from $\mathrm{N}(0, \sigma^2)$. This is equivalent to saying that y_i are i.i.d. from $\mathrm{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$.

Given the Gaussian assumptions the MLE coincides with the least squares problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \tag{14.2}$$

Instead of working across the n individual observations we can concatenate them into random vectors and matrices so that the linear model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{14.3}$$

where $\mathbf{X}$ is the design matrix, $\boldsymbol{\epsilon} \sim \mathrm{MN}(\mathbf{0}, \sigma^2 \mathbf{I})$ and hence $\mathbf{y} \sim \mathrm{MN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$. The corresponding minimisation problem is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \tag{14.4}$$

If $(\mathbf{X}^T\mathbf{X})^{-1}$ exists the minimising solution of (14.4)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{14.5}$$

Proposition 14.1. *Assume* $(\mathbf{X}^T\mathbf{X})^{-1}$ *exists, then the parameter estimates are* $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

Proof.

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$Cov[\hat{\boldsymbol{\beta}}] = Cov[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, \sigma^2\mathbf{I}\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-T} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

The normality follows from the normality of $\mathbf{y}$. □

Proposition 14.2. *Assume* $(\mathbf{X}^T\mathbf{X})^{-1}$ *exists, then the predictions are* $\hat{\mathbf{y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X})$

Proof.

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

$$E[\hat{\mathbf{y}}] = E[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta}$$

$$Cov[\hat{\mathbf{y}}] = \mathbf{X}^T Cov[\hat{\boldsymbol{\beta}}]\mathbf{X} = \sigma^2\mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$$

The normality follows from the normality of $\hat{\boldsymbol{\beta}}$. □

Proposition 14.3. *Assume* $(\mathbf{X}^T\mathbf{X})^{-1}$ *exists, then the predictions errors* $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ *are* $\mathbf{e} \sim MN(\mathbf{0}, \sigma^2\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right))$ *and they are orthogonal to the column space of the design matrix* $\mathbf{X}^T\mathbf{e} = \mathbf{0}$

Proof.

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\,\mathbf{y}$$

$$E[\mathbf{e}] = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)E[\mathbf{y}] = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\begin{aligned}
Cov[\mathbf{e}] &= E[\mathbf{e}\mathbf{e}^T] = E[\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\,\mathbf{y}\mathbf{y}^T\,\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T] \\
&= \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\,\sigma^2\mathbf{I}\,\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) \\
&= \sigma^2\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)
\end{aligned}$$

The normality follows from the normality of $\mathbf{y}$.

$$\langle\mathbf{X},\mathbf{e}\rangle = \mathbf{X}^T\mathbf{e} = \mathbf{X}^T\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\,\mathbf{y} = \left(\mathbf{X}^T - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\,\mathbf{y} = \left(\mathbf{X}^T - \mathbf{X}^T\right)\,\mathbf{y} = \mathbf{0}$$

□

Mean squared estimation error (MSEE)

$$\begin{aligned}
\text{MSEE} &= E\left[||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_2^2\right] = E\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \\
&= E\left[(\boldsymbol{\beta}^T\boldsymbol{\beta} - 2\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}})\right] = \boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\beta} + E[\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}] \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + E[\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + E[\text{Tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\right)] \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + \text{Tr}\left((\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(Var[\mathbf{y}] + E[\mathbf{y}]E[\mathbf{y}^T])\right) \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + \text{Tr}\left((\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T)\right) \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + \sigma^2\,\text{Tr}(\mathbf{X}^T\mathbf{X})^{-1} + \text{Tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right) \\
&= -\boldsymbol{\beta}^T\boldsymbol{\beta} + \sigma^2\,\text{Tr}(\mathbf{X}^T\mathbf{X})^{-1} + \text{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\right) \\
&= \sigma^2\,\text{Tr}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\sum_i^m\frac{1}{\lambda_i} \geq \sigma^2\frac{1}{\lambda_{min}}\quad, \qquad (14.6)
\end{aligned}$$

where λ_i are the eigenvalues of $\mathbf{X}^T\mathbf{X}$. In result, **if the data is such that $\mathbf{X}^T\mathbf{X}$ has some very small eigenvalues, the MSEE will be very large**.

Residual sum of squares (RSS) Residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

$$\begin{aligned}
\text{RSS} &= ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}^T\mathbf{y}
\end{aligned}$$

Total sum of squares (TSS)

$$\text{TSS} = ||\mathbf{y} - \bar{\mathbf{y}}||_2^2 = (\mathbf{y} - \bar{\mathbf{y}})^T(\mathbf{y} - \bar{\mathbf{y}}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}}$$

Explained sum of squares (ESS)

$$\begin{aligned}
\text{ESS} &= ||\hat{\mathbf{y}} - \bar{\mathbf{y}}||_2^2 = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^T(\hat{\mathbf{y}} - \bar{\mathbf{y}}) = \hat{\mathbf{y}}^T\hat{\mathbf{y}} - 2\hat{\mathbf{y}}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \hat{\mathbf{y}}^T\mathbf{y} - 2\hat{\mathbf{y}}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}}
\end{aligned}$$

$$\begin{aligned}
\text{RSS} &= \text{TSS} - \text{ESS} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}} - \hat{\mathbf{y}}^T\mathbf{y} + 2\hat{\mathbf{y}}^T\bar{\mathbf{y}} - \bar{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\bar{\mathbf{y}} - \hat{\mathbf{y}}^T\mathbf{y} + 2\hat{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \text{RSS} - 2\mathbf{y}^T\bar{\mathbf{y}} + 2\hat{\mathbf{y}}^T\bar{\mathbf{y}} \\
&= \text{RSS} - 2\bar{y}\,(\mathbf{y}^T - \hat{\mathbf{y}}^T)\mathbf{1} = \text{RSS} - 2\bar{y}\,\mathbf{e}^T\mathbf{1} \\
&= \text{RSS} \qquad (\text{because } \mathbf{X}_{:,1}^T\mathbf{e} = \mathbf{1}^T\mathbf{e} = \mathbf{0})
\end{aligned}$$

Mean squared prediction error (MSPE) (over the train samples)

$$\begin{aligned}
\text{MSPE} &= E[||\mathbf{y} - \hat{\mathbf{y}}||_2^2] = E[\mathbf{e}^T\mathbf{e}] = E[\mathbf{y}^T\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{y}] \\
&= E[\mathbf{y}^T\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{y}] = E[\text{Tr}\left(\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{y}\mathbf{y}^T\right)] \\
&= \text{Tr}\left(\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)E[\mathbf{y}\mathbf{y}^T]\right) \\
&= \text{Tr}\left(\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)Var[\mathbf{y}] + E[\mathbf{y}]E[\mathbf{y}^T]\right) \\
&= \text{Tr}\left(\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\left(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right)\right) \\
&= \sigma^2\,\text{Tr}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) + \text{Tr}\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right) \\
&= \sigma^2\,\text{Tr}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) + \text{Tr}\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right) \\
&= \sigma^2 n + \sigma^2\,\text{Tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) = \sigma^2 n + \sigma^2\,\text{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\right) \\
&= \sigma^2 n + \sigma^2\,\text{Tr}(\mathbf{I}_m) = \sigma^2(m+n)
\end{aligned}$$

14.2 Ridge regression

In [2] the author proposes to estimate the parameters as

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{W}\mathbf{X}^T\mathbf{y} \;, \tag{14.7}$$

which is a solution to the following optimisation problem

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + k||\boldsymbol{\beta}||_2^2 \tag{14.8}$$

Proposition 14.4.

$$\hat{\boldsymbol{\beta}}_R = \left(I + k(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}} = \mathbf{Z}\hat{\boldsymbol{\beta}} \tag{14.9}$$

Proof.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_R &= \left(\mathbf{I} + k(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}} \\
&= \left(\mathbf{I} + k(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (I + A^{-1})^{-1} = (A+I)^{-1}A \\
&= \left(\mathbf{I} + \mathbf{X}^T\mathbf{X}/k\right)^{-1}\mathbf{X}^T\mathbf{X}/k(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(k\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

□

We indicate by $\lambda_max = \lambda_1 \geq \lambda_2 \ldots \geq \lambda_d = \lambda_min$ the eigenvalues of $\mathbf{X}^T\mathbf{X}$. For the eigenvalues of $\mathbf{W}$ and $\mathbf{Z}$ we have

$$\xi_i(\mathbf{W}) = \frac{1}{\lambda_i + k} \qquad \xi_i(\mathbf{Z}) = \frac{\lambda_i}{\lambda_i + k} \tag{14.10}$$

Relation between $\mathbf{Z}$ and $\mathbf{W}$ is

$$\begin{aligned}\mathbf{Z} &= \left(I + k(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1} = \left(k\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X} \\ &= \mathbf{W}\mathbf{X}^T\mathbf{X} \qquad\qquad (\mathbf{X}^T\mathbf{X} = \mathbf{W}^{-1} - k\mathbf{I}) \\ &= \mathbf{W}(\mathbf{W}^{-1} - k\mathbf{I}) = \mathbf{I} - k\mathbf{W} \qquad (14.11) \\ &= \mathbf{I} - k\left(k\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned}$$

Proposition 14.5. *Ridge estimate is shorter (in ℓ_2) than the OLS estimate $||\hat{\boldsymbol{\beta}}_R||_2^2 \leq ||\hat{\boldsymbol{\beta}}||_2^2$*

Proof.

$$\begin{aligned}||\hat{\boldsymbol{\beta}}_R||_2^2 &= ||\mathbf{Z}\hat{\boldsymbol{\beta}}||_2^2 \qquad \text{(CS - need to use the operator norm)} \\ &\leq ||\mathbf{Z}||_2^2\,||\hat{\boldsymbol{\beta}}||_2^2 = \xi_{max}^2(\mathbf{Z})\,||\hat{\boldsymbol{\beta}}||_2^2 \\ &= \left(\frac{\lambda_{max}}{\lambda_{max} + k}\right)^2 ||\hat{\boldsymbol{\beta}}||_2^2 \leq ||\hat{\boldsymbol{\beta}}||_2^2 \qquad (k, \lambda_{max} \geq 0) \qquad (14.12)\end{aligned}$$

□

As $k \to \infty$ the ridge estimates becomes a lot shorter than the OLS estimate.

Residual sum of squares (RSS) Residuals $\mathbf{e}_R = \mathbf{y} - \hat{\mathbf{y}}_R$

$$\begin{aligned}\mathrm{RSS}_R &= ||\mathbf{y} - \hat{\mathbf{y}}_R||_2^2 = \mathbf{e}_R^T\mathbf{e}_R = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R)^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_R + \hat{\boldsymbol{\beta}}_R^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_R \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_R + \hat{\boldsymbol{\beta}}_R^T(\mathbf{W}^{-1} - k\mathbf{I})\hat{\boldsymbol{\beta}}_R \\ &= \mathbf{y}^T\mathbf{y} - 2\hat{\boldsymbol{\beta}}_R^T\mathbf{X}^T\mathbf{y} + \hat{\boldsymbol{\beta}}_R^T\mathbf{X}^T\mathbf{y} - k\hat{\boldsymbol{\beta}}_R^T\hat{\boldsymbol{\beta}}_R \\ &= \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_R^T\mathbf{X}^T\mathbf{y} - k\hat{\boldsymbol{\beta}}_R^T\hat{\boldsymbol{\beta}}_R \\ &= \mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}_R^T\mathbf{y} - k\hat{\boldsymbol{\beta}}_R^T\hat{\boldsymbol{\beta}}_R\end{aligned}$$

which when compared to OLS has a correction for squared length of $\hat{\boldsymbol{\beta}}_R$.

Ridge trace If $\mathbf{X}^T\mathbf{X} \not\approx \mathbf{I}$ (it has $\lambda_{min} \approx 0$) the MSEE of OLS (14.14) will be large.

For any estimate $\hat{\boldsymbol{\beta}}_A$ of $\boldsymbol{\beta}$ the residual sum of squares is ($\hat{\boldsymbol{\beta}}$ is the OLS estimate)

$$\begin{aligned}\mathrm{RSS}_A &= ||\mathbf{y} - \hat{\mathbf{y}}_A||_2^2 = \mathbf{e}_A^T\mathbf{e}_A = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_A)^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_A) \\ &= (\mathbf{y} - \mathbf{X}(\hat{\boldsymbol{\beta}}_A + \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}))^T(\mathbf{y} - \mathbf{X}(\hat{\boldsymbol{\beta}}_A + \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{X}(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}) \qquad \text{(middle term gone due to proposition (14.3))} \\ &= RSS_{OLS} + ||\mathbf{X}\,\Delta\boldsymbol{\beta}||_2^2 = RSS_{OLS} + \Delta RSS \leq RSS_{OLS} + \lambda_{max}||\Delta\boldsymbol{\beta}||_2^2\end{aligned}$$

There may me multiple estimates $\mathbf{B}$ on the same RSS contour. Ridge searches for $\mathbf{B}$ with minimum length. That is

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}_A} \boldsymbol{\beta}_A^T \boldsymbol{\beta}_A, \quad \text{s.t. } (\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}) = \Delta RSS$$

which is equivalent to

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}_A} \boldsymbol{\beta}_A^T \boldsymbol{\beta}_A + 1/k \left((\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}) - \Delta RSS \right)$$

Taking derivative and equating to zero we get

$$2\hat{\boldsymbol{\beta}}_R + 1/k \left(2\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}_R - 2\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} \right) = 0$$

and

$$\hat{\boldsymbol{\beta}}_R = (k\mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{14.13}$$

Mean squared estimation error of ridge regression (MSEE$_R$)

$$\text{MSEE}_R \quad = \quad E\left[||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_R||_2^2 \right] = E\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_R)^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_R) \right] \tag{14.14}$$

References

[1] Knight, K. (2000). Mathematical statistics. Chapman and Hall/CRC.

[2] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics,

Index

ϵ-insensitive loss, 9
λ_{max}, 24

Canonical hyperplane, 6
Chebyshev's inequality, 33, 34
Chernoff's bound, 34
Complementary slackness, 4, 5
conditional expectation, 44
Constraint qualifications, 4

Dual feasable, 3
Dual problem, 3
Duality gap, 4
Duality gap - optimal, 3

ESS, 49
Explained sum of squares, 49

Gaussian, 15
Gradient descent, 31
Gram matrix, 11

Hinge loss, 8
Hoeffding's inequality, 36

Inverstion identity lemma, 12

Karush-Kuhn-Tucker conditions, 5
KKT conditions, 5

Lagrange dual function, 3
Lagrange dual problem, 3
Lagrangian, 3
Lasso, 24
Linear kernel, 12
Linear projection, 44
Loss function, 44

Markov's inequality, 33
Martingale Difference, 36
Matrix-valued kernel, 12
Mean square estimation error, 48
Mean squared prediction error, 49
MKL, 17
MSEE, 48
MSPE, 49
Multi-output ridge regression, 14

Nuclear norm, 25

OLS, 45
Optimal Lagrange multipliers, 3
Ordinary least squares, 45
Orthogonal random variables, 44

Perturbations, 29
Prediction MSE, 44
Primal problem, 3

Representer theorem, 11, 12
Reproducing property, 11
Residual sum of squares, 48
Ridge regression, 12, 14, 30, 31
RKHS, 11
RSS, 48

Separable kernels, 13
Slater's condition, 4
Soft SVM Lagrangian, 8
Soft-margin SVM, 8
Soft-margin SVM dual function, 8
Soft-margin SVM dual problem, 8, 10
Soft-margin SVM primal problem, 8
Strong duality, 4
subdifferential, 27
subgradient, 27
Support vector regression, 9
SVM dual function, 7
SVM dual problem, 7
SVM Lagrangian, 7
SVM margin, 6
SVM primal problem, 6
SVR dual function, 9
SVR dual problem, 10
SVR Lagrangian, 9
SVR primal problem, 9

Tikhonov regularization, 29
Total sum of squares, 48
Trace norm, 25
TSS, 48

Weak duality, 3