Machine Intelligence 1 **Exercise Sheet 5**
WS 2016/17, Obermayer/Böhmer due: 24.11.2016

Regularization and the bias variance trade-off

Exercise T5.1: Regularization (tutorial)

Regularization is an important aspect of machine learning. Here we will discuss its influence beyond weight decay terms.

(a) We have already discussed what a difference the choice of cost function has on the solution. How will the choice of one of the following regularization terms influence the solution of a quadratic cost function?

$$\begin{aligned} R(\boldsymbol{w}) &= \sum_{i=1}^{d} |w_i|^{\frac{1}{2}} && \text{(square root semi-norm regularization)} \\ R(\boldsymbol{w}) &= \sum_{i=1}^{d} |w_i| && \text{(L_1 norm regularization)} \\ R(\boldsymbol{w}) &= \sum_{i=1}^{d} w_i^2 && \text{(L_2 norm regularization)} \\ R(\boldsymbol{w}) &= \max\Big(|w_1|, \ldots, |w_d|\Big) && \text{(L_∞ norm regularization)} \end{aligned}$$

(b) Give at least two application examples in which one of the above regularization terms is preferable to the others.

(c) Assume a linear neuron with a quadratic cost function. How does the optimization problem change if for each original training sample $\underline{\mathbf{x}}^{(\alpha)}$ an infinite number of samples $\underline{\mathbf{x}}'^{(\alpha)}$ is drawn i.i.d. from a Gaussian with zero mean and variance σ^2, i.e. $\underline{\mathbf{x}}'^{(\alpha)} = \underline{\mathbf{x}}^{(\alpha)} + \epsilon, \epsilon \sim \mathcal{N}(\underline{\mathbf{0}}, \sigma^2 \underline{\mathbf{I}})$?

(d) Explain the concept of *nested n-fold cross-validation* and describe the difference to normal *n-fold cross-validation*.

Exercise T5.2: Nonlinear basis functions (tutorial)

Instead of dealing with deep neural networks, many machine learning approaches use a linear neuron on input samples $\underline{\mathbf{x}}$, which are "expanded" by non-linear basis functions $\phi_i(\underline{\mathbf{x}})$. In the lecture, these are parameterizable *radial basis functions*, but here we want to discuss the set of all monomials up to some degree/order.

(a) What are monomials and how is a linear combination thereof called?

(b) Which monomials would be sufficient to solve the XOR problem?

(c) Monomials can grow very large for bigger inputs. To standardize the input space, one often *spheres* the data before expansion. How is "whitening" or "sphering" performed?

(d) Monomial basis functions can be regularized by weight decay. However, even small weights can lead to very rough functions. How can the regularization term be modified to enforce smooth functions?

Exercise H5.1: Biased and unbiased estimators (homework, 4 points)

Let $\{x_1, \dots, x_n\} \subset \mathbb{R}$ be a sequence of n *iid* drawn samples, i.e. $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle \, , \forall i \neq j \, ,$ and let $\langle x_i \rangle = \mu \, , \langle (x_i - \mu)^2 \rangle = \sigma^2 \, , \forall i \, ,$ denote the common mean and variance. We will discuss the following *empirical estimates*:

$$\begin{aligned} E_n[x] &= \frac{1}{n} \sum_{t=1}^{n} x_t && \text{(empirical mean)} \\ V_n[x] &= \frac{1}{n} \sum_{t=1}^{n} (x_t - E_n[x])^2 && \text{(empirical variance)} \end{aligned}$$

(a) (1 point) Prove that $E_n[x]$ is an unbiased estimator of μ, i.e. $\langle E_n[x] \rangle = \mu \, .$

(b) (1 point) Show that $\langle x_i^2 \rangle = \sigma^2 + \mu^2, \forall i.$

(c) (1 point) Show that the variance of $E_n[x]$ is larger than zero, i.e. $\langle (E_n[x] - \mu)^2 \rangle = \frac{\sigma^2}{n} \, .$

(d) (1 point) Prove that $V_n[x]$ is a biased estimator of σ^2, i.e. $\langle V_n[x] \rangle \neq \sigma^2 \, .$ For which normalization is the estimator $V_n[x]$ unbiased?

Solution

(a) The samples are drawn i.i.d., and thus holds $\langle x_t \rangle = \mu$, i.e., $\langle E_n[x] \rangle = \frac{1}{n} \sum_{t=1}^{n} \langle x_t \rangle = \mu \; .$

(b) $\langle x_i^2 \rangle = \langle (x_i - \mu)^2 \rangle + 2 \langle x_i \mu \rangle - \mu^2 = \sigma^2 + \mu^2 \; .$

(c) Note that $\langle x_i x_j \rangle = \mu^2, \forall i \neq j.$

$$\begin{aligned} \langle (E_n[x] - \mu)^2 \rangle &= \langle E_n^2[x] \rangle - 2\mu \langle E_n[x] \rangle + \mu^2 &&= \tfrac{1}{n^2} \sum_{i,j=1}^{n} \langle x_i x_j \rangle - \mu^2 \\ &= \tfrac{1}{n^2} \sum_{i=1}^{n} \Big(\langle x_i^2 \rangle + \sum_{j \neq i}^{n} \langle x_i x_j \rangle \Big) - \mu^2 &&= \tfrac{1}{n} \Big(\sigma + n\mu^2 \Big) - \mu^2 = \frac{\sigma^2}{n} \, . \end{aligned}$$

(d) We use in the following $\frac{1}{n^2} \sum_{i,j=1}^{n} \langle x_i x_j \rangle = \frac{\sigma^2}{n} + \mu^2$ from (b).

$$\begin{aligned} \langle V_n[x] \rangle &= \tfrac{1}{n} \sum_{i=1}^{n} \langle x_i^2 \rangle - \tfrac{1}{n^2} \sum_{i,j=1}^{n} \langle x_i x_j \rangle \\ &= \sigma^2 + \mu^2 - \tfrac{\sigma^2}{n} - \mu^2 = \tfrac{n-1}{n} \sigma^2 \end{aligned}$$

The variance is therefore *biased*. The normalization $\frac{n}{n-1}$ yields an unbiased estimate.

Exercise H5.2: Cross-validation (homework, 6 points)

This exercise asks you to assess the impact of a quadratic regularization penalty on the parameter estimates for a linear connectionist neuron to solve a regression task.

Data: The file `TrainingRidge.csv` contains the *training set*, which are 200 observations $\{\mathbf{x}_n, y_n\}$. The two input variables for each observation $\mathbf{x}_i = [x_{i,1}, x_{i,2}]$ are contained in the first 2 columns. The target values (labels) y_i are contained in the last column. The second file `ValidationRidge-Y.csv` contains 1476 combinations $[x_{i,1}, x_{i,2}]$ (a 36×41 grid) as a *validation set* in the same format.

(a) (1 point) The observations are from a large space. Monomials (see below) can grow very large for bigger inputs. Perform *whitening* (sphering) of the training data, such that the resulting input samples are decorrelated, have zero mean and unit variance (i.e. $\sigma = 1$). Plot the sphered training and validation sets in two scatter-plots, where the color of the markers represents the associated label.

(b) (1 point) A single linear neuron is not able to predict the target labels very well. To increase the representational power of the model class, *expand* the sphered 2-dimensional input samples to all possible *monomials* up to degree 9. Here, a monomial of order k correspond to a term $x_{i,1}^a x_{i,2}^b$ with $a + b = k$, and the model should contain all 55 terms $x_1^a x_2^b$ with $k = 0, 1, ..., 9$. Plot the resulting validation set for the first 10 monomials ($0 \leq k \leq 3$) either as a scatter plot or as a 36×41 image, where the colors indicate the labels.

(c) (2 points) To avoid over-fitting of a linear neuron with the polynomial expansion of (b), the solution must be regularized with a weight-decay term, i.e., for a regularization constant λ, an input matrix $\underline{\mathbf{X}} \in \mathbb{R}^{55 \times N}$ and a label vector $\underline{\mathbf{y}} \in \mathbb{R}^{N \times 1}$, the prediction function is

$$y(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \;=\; \underline{\mathbf{w}}^\top \underline{\mathbf{x}}, \qquad \text{with} \qquad \underline{\mathbf{w}} \;=\; \big(\mathbf{X}\mathbf{X}^\top + \lambda \underline{\mathbf{I}}\big)^{-1} \underline{\mathbf{X}}\underline{\mathbf{y}}\,.$$

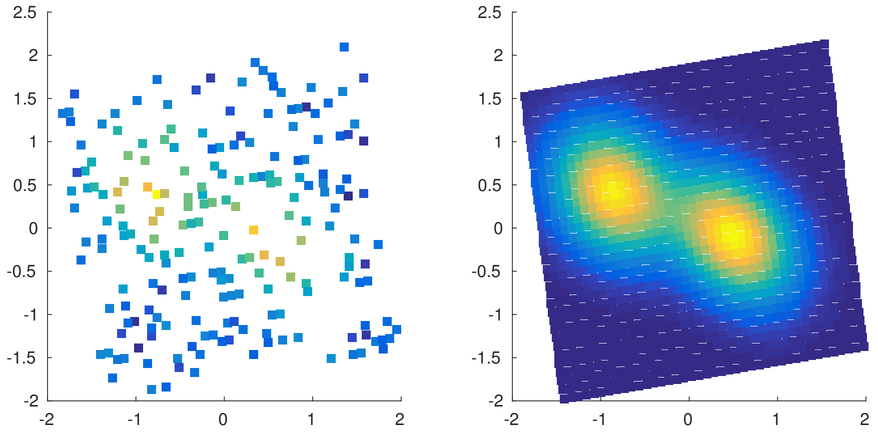
To find the best regularization constant, perform a 10-fold cross-validation with the training set for all $\lambda \in \{10^z \,|\, z \in \{-4, -3.9, -3.8, \ldots, 3.9, 4\}\}$. Plot the resulting mean and standard deviation of the MSE over all folds against λ (as an error-bar plot with a logarithmic x-axis). Also plot (similarly to (b)) the true labels of the validation set next to the predicted labels for the best regularization parameter (λ_T), which minimizes the mean MSE over all folds .

(d) (2 points) To compare these empirical estimates of bias and variance with the true generalization error, repeat (c) with the polynomial expansion of the *validation set*. Is the best lambda (λ_G) different from λ_T? Compare the learned function in (c) with functions that are learned with λ_G on (i) the training set and (ii) the validation set.

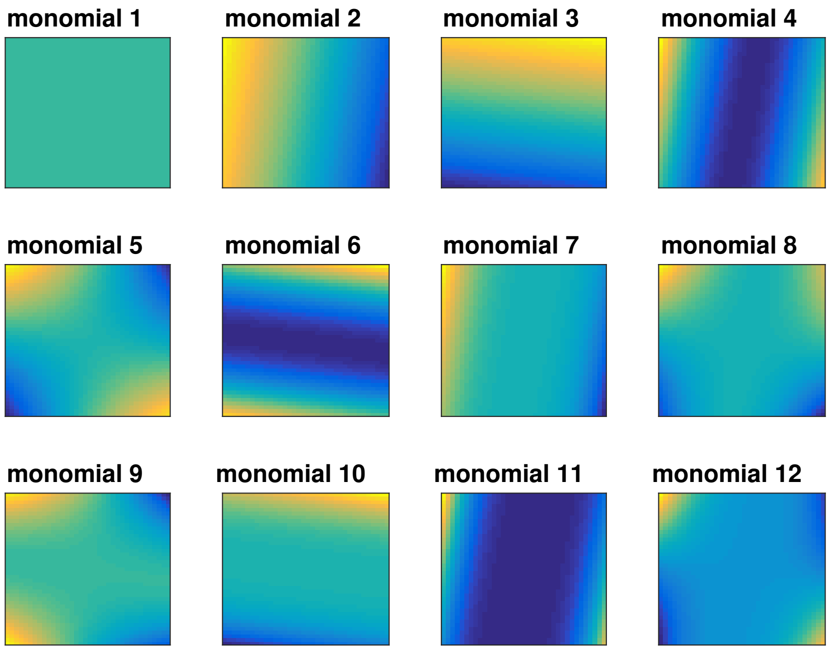
(e) (optional) Perform (c) with one of the possible regularization schemes discussed in (T5.2d), e.g. $R(\underline{\mathbf{w}}) = \sum_{a=1}^{d} r_i \, w_i^2$, where the basis-specific regularization of the i'th monomial $\phi_i(\underline{\mathbf{x}}) = x_1^a x_2^b$ is $r_i = (a-1)^2(b-1)^2$.

Solution

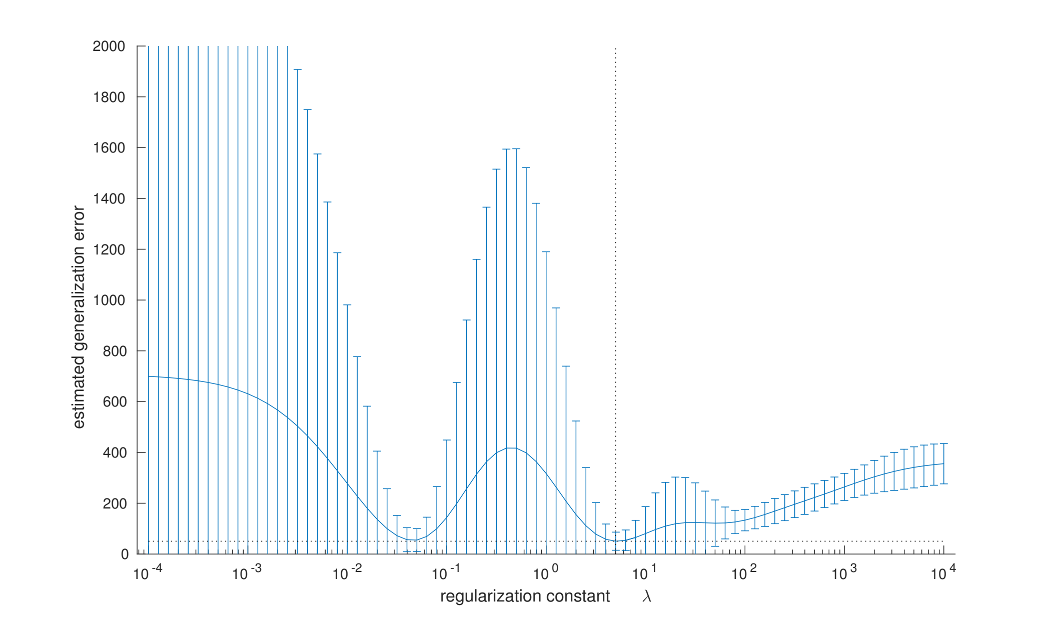
(a) The sphered data are centered can be slightly rotated.



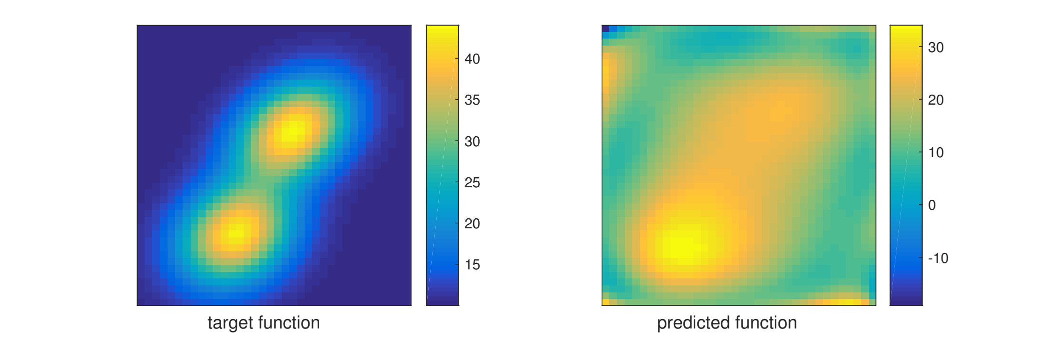
(b) The result are monomials in the sphered space.



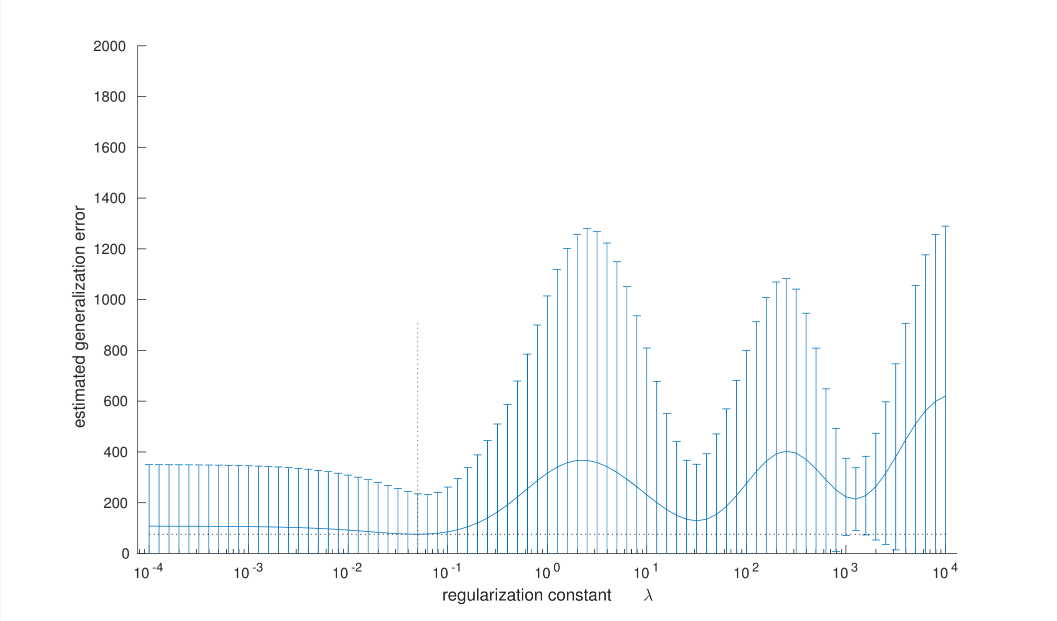
(c) The mean and standard deviation of the MSE over 10 folds for the training set with all tested regularization constants λ:



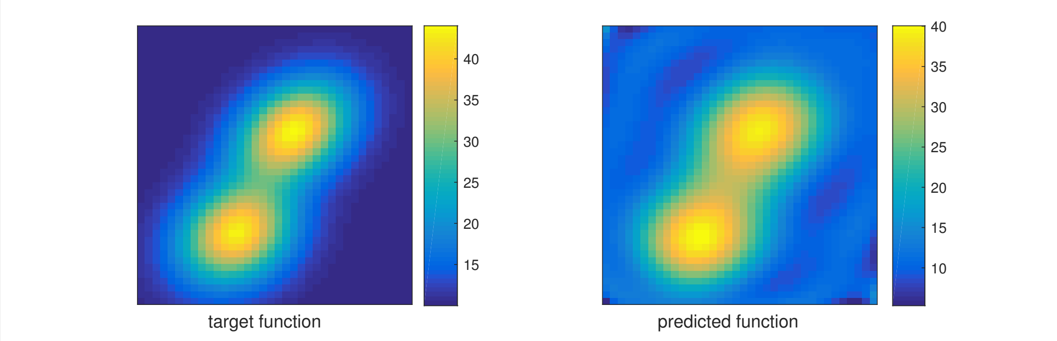
The predicted function for the training set and $\lambda_T \approx 5$:



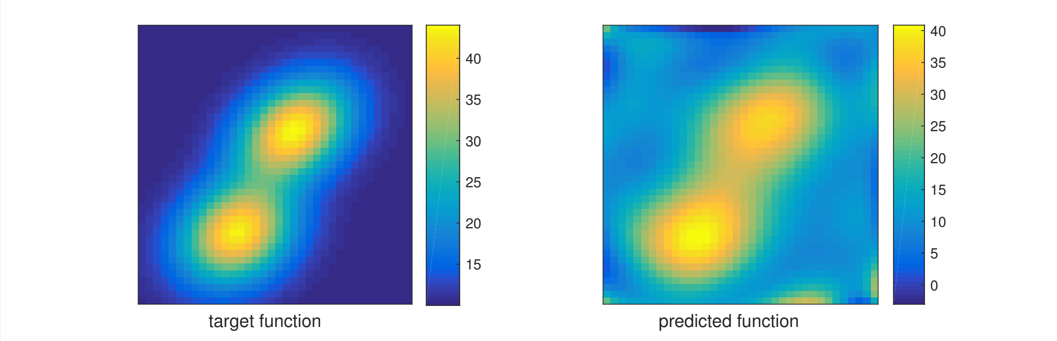
(d) The mean and standard deviation of the MSE over 10 folds for the validation set with all tested regularization constants λ:



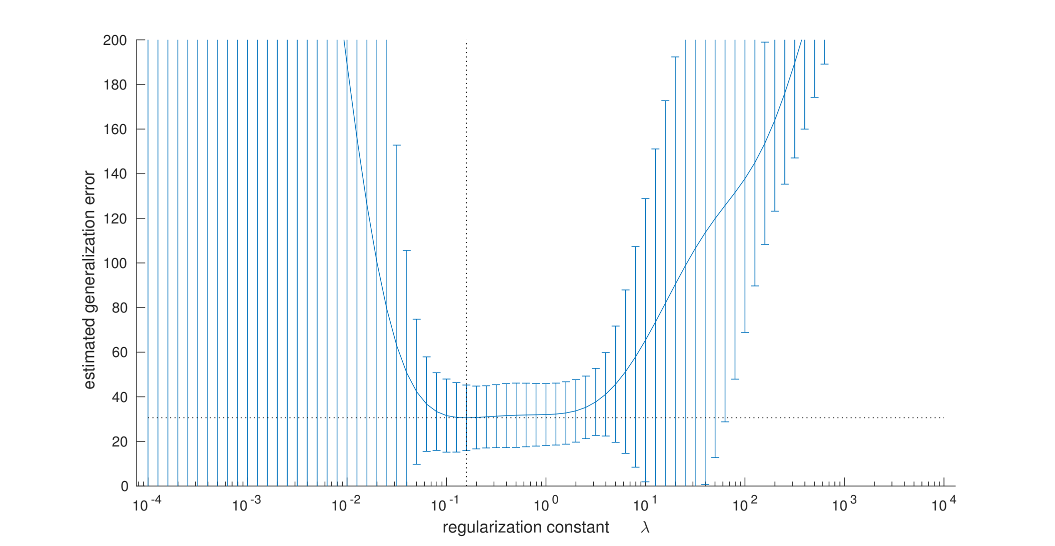
The predicted function for the validation set and $\lambda_G \approx 0.05$:



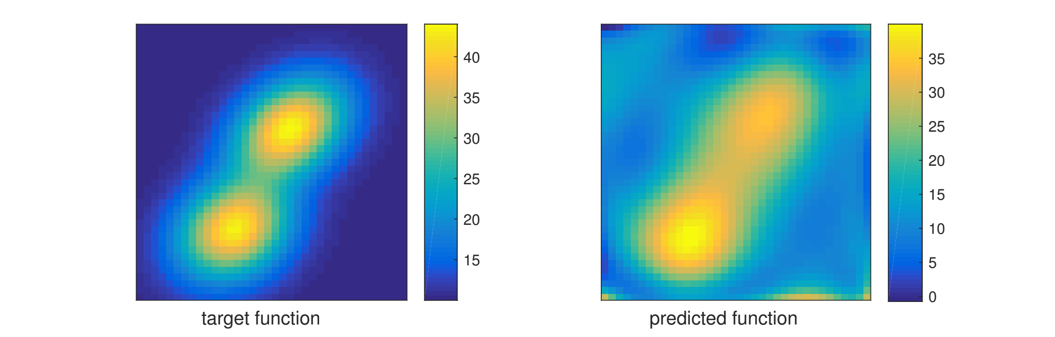
The predicted function for the training set and $\lambda_G \approx 0.05$:



(e) The mean and standard deviation of the MSE over 10 folds for the training set with all tested regularization constants λ and basis-specific regularization:



The predicted function for the training set and $\lambda_T' \approx 0.15$ (not comparable with λ_T):



Total 10 points.