Elements of Machine Learning, WS 2022/2023
Aleksandar Bojchevski and Jilles Vreeken
Exam, February 24th, 2023, Solution Sheet

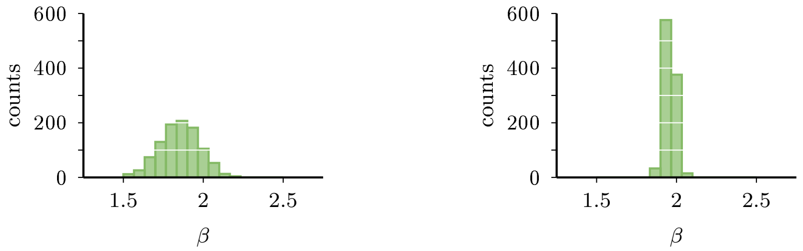
Problem 1 (Lots of Data) **(10 points)**

1. We want to predict target Y given a single predictor X. We collected two datasets from the same distribution, one of $n = 10$ samples, and a second of $n = 1000$ samples. (10 pts)

   We first fit a simple linear regression model.

   (a) What is a good approach to compare the least-squares estimate of β we get for the one dataset to the least-squares estimate of β we get for the other dataset? Explain in your own words. (1 pt)

   (b) We use bagging to obtain a thousand estimates of the parameter β for each dataset. We show the results in Fig. 1. Which of the two figures corresponds to the small dataset ($n = 10$) and which to the large ($n = 1000$) dataset? Explain your choice. (1 pt)



(a) Estimates for dataset A.

(b) Estimates for dataset B.

Figure 1: Estimates of the linear coefficient β for two datasets.

   (c) We now consider polynomial regression with degree d. Compare the bias and variance of a model with degree $d = 3$ to a model with degree $d = 10$. (2 pts)

   (d) Explain how we can control the flexibility of

      i. a spline regression model, and (1 pt)

      ii. a regression tree. (1 pt)

   (e) Will fitting a more flexible model on the large dataset ($n = 1000$) *always* achieve a lower test error than fitting a less flexible model on the small dataset ($n = 10$)? If yes, explain your reasoning; if no, explain what modification we can make to the learning procedure to address this aspect. (2 pts)

   (f) To decide the flexibility (e.g., the degree d) of the above models, we consider using k-fold cross validation or leave-one-out cross validation (LOOCV).

      (a) Which do you recommend for the small, respectively the large dataset? Why? (1 pt)

      (b) Which do you recommend in general and why? (1 pt)

Solution.

1. The uncertainty of the coefficients depends on the sample size n.

   (a) Two different ways to compare the estimates for different datasets are confidence intervals and bagging. For confidence intervals, we have seen in the lecture that the 95% confidence interval for β_i where $i \in \{0,1\}$ is given by

   $$\left[\widehat{\beta}_i - 2 \cdot \mathrm{SE}(\widehat{\beta}_i), \widehat{\beta}_i + 2 \cdot \mathrm{SE}(\widehat{\beta}_i)\right]$$

   where $\mathrm{SE}^2(\widehat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$ and $\mathrm{SE}^2(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$. In particular, these confidence intervals for the parameters generally become smaller as n increases and if the $\widehat{\beta}_i$ for different datasets have overlapping confidence intervals, their results are generally considered not to be significantly different. For boosting, we resample data X^k, Y^k from the original dataset X, Y and compute $\widehat{\beta}^k$ for each of these datasets. We can then look at the empirical confidence intervals for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by sorting the $\widehat{\beta}_i^k$ and looking at the 2.5% and 97.5% quantiles, and again check whether the confidence intervals for different datasets overlap.

   (b) Smaller datasets will lead to large variance of estimates of $\widehat{\beta}$, whereas larger datasets will lead to much smaller variance of estimates of $\widehat{\beta}$. In fact, in a difference of $n = 1000$ to $n = 10$, we would expect roughly a $\sqrt{100}\times$ decrease in the variance, which looks about right for the plots shown here.

   (c) Compared to a polynomial regression model with $d = 3$, a model with $d = 10$ has many more free parameters and is therefore more flexible. The model with $d = 10$ therefore has lower bias and higher variance than the model with $d = 3$.

   (d) i. For a regression spline, we can change the number of knots or the degrees of its local polynomials, both of which would change the number of free parameters, $d+K+1$ (assuming $d-1$-times differentiability). Alternatively, we could change the number of continuity/differentiability constraints at each knot.

      ii. For regression trees, we can change the depth of the tree, which constrains its flexibility by allowing it to perform fewer splits and therefore containing more data points per leaf node.

   (e) No, even if $n = 1000$, if we try to fit an arbitrarily complex model we still overfit. To avoid this issue, e.g., for polynomial regression, we can regularize the parameters. For splines, we could do the same, or impose smoothness constraints at the knots.

   (f) Cross Validation approaches and diminished returns for larger n.

      (a) For the smaller dataset LOOCV may be better, since using k-fold CV may end up throwing away too much data so that the model cannot be trained well enough. For the larger dataset, however, k-fold CV is definitely preferable since the loss of some data is less impactful the more data we have, but the improvement on the estimate of the generalization error of k-fold CV due to less correlated datasets compared to LOOCV is significant.

(b) In general, there is a trade-off in terms of bias and variance between K-fold CV and LOOCV. When the goal is to estimate the generalization error after our data containing n points has been observed, then both of them *over*estimate the generalization error by using estimates based on fewer than n points. Since LOOCV uses n-1 points while K-fold CV uses $(K-1)/K \cdot n$ points, the bias for LOOCV is generally lower than for K-fold CV. However, the datasets used for training LOOCV have much larger overlap and are therefore more correlated than the ones in K-fold CV, leading to much larger variance for LOOCV. When the number of points n is large, the bias of using, say, 5-fold CV is usually not very large since the amount of information lost by losing 20% of the data is usually not very significant. Furthermore, using LOOCV requires us to fit the model n times, increasing the computation load dramatically, while K-fold CV only increases it by a constant factor K independent of n. This is another reason why K-fold CV with some moderate K is often preferred.

Problem 2 (Linear Regression) (10 points)

1. Determine if the below statements are true or false. For every false statement, either provide a counter example *or* correct it by replacing a *single* term (noun or adjective). (3 pts)

   (i) The Gauss-Markov theorem states there exists no estimator for the coefficients of the linear model that achieves lower variance than the least squares estimator.

   (ii) We should use the t-test to determine if a set of more than one predictor is significantly correlated with the outcome.

   (iii) Assume that we fit a linear model on a single predictor; if its coefficient is 0 then the outcome must be statistically independent with this predictor.

   (iv) Assume a dataset that comes from an underlying linear model $y = \beta x + \epsilon$, where ϵ is arbitrary noise. Then $t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$ follows a student t distribution, where $\hat{\beta}$ is the least squares estimate of β and $\text{SE}(\hat{\beta})$ its standard error.

2. One of our physicist friends is studying a phenomenon between a single predictor X and a target variable Y that can be described as a linear model satisfying the least squares assumptions. We have 50 datapoints shown on Fig 2.

   (a) Explain which out of x_a, x_b, x_c are *outliers*, which of these are *high leverage points*, and which are both? (1 pt)

   (b) Give a short explanation why our friend should be concerned about *outliers*, respectively about *high leverage points*? (1 pt)

   (c) Suppose we may remove *one* datapoint before we fit the least squares estimate. Which out of x_a, x_b or x_c would you remove? Why? (1 pt)

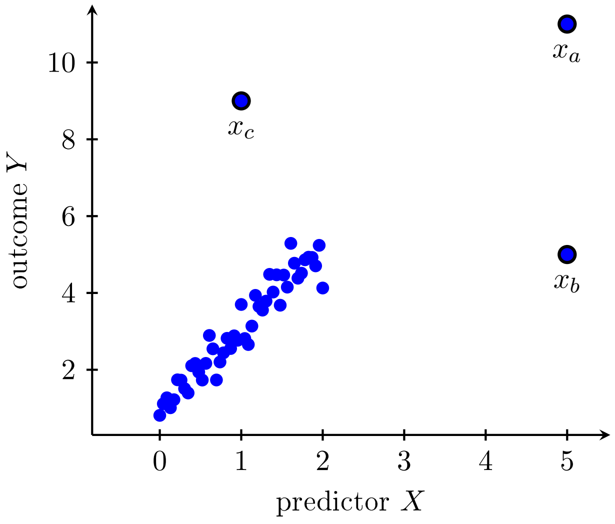


Figure 2: A dateset with 50 datapoints, where three points are annotated as x_a, x_b, x_c.

3. Consider linear models with a single predictor.

   (a) Give a counterexample to the following statement. (1 pt)

   Heteroskedastic noise leads to higher prediction error than homoskedastic noise.

   (b) When should we check for heteroskedasticity? How do we do so? (1 pt)

4. We are analyzing how sales are affected by advertising on three different media. By fitting a multiple linear regression model we get the following coefficients.

	intercept	YouToob	Bacefook	Twutter
coefficient	3.010	0.040	0.190	-0.010

   Based on these results, co-worker A suggests that the company should stop advertising on Twutter as it hurts sales. Co-worker B suggests the opposite.

   (a) Give a brief explanation or a counter example why colleague A might be wrong. (1 pt)

   (b) Explain how we can test whether colleague B might be right. (1 pt)

Solution.

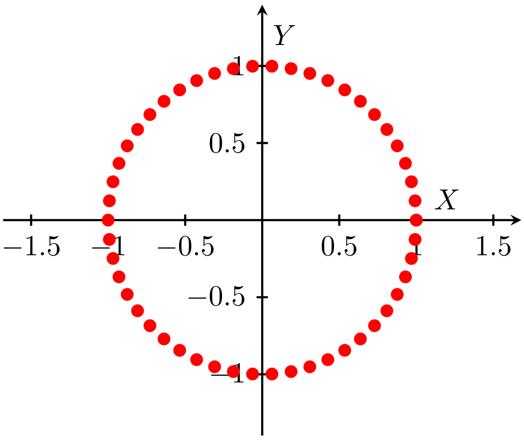
1. All of these statements are false. The notes are not a required part of the answer.

   (a) False.
   The Gauss-Markov theorem refers only to the *unbiased* estimators. A simple *counter-example* is an estimator which gives a constant estimate for the weights. Its variance would therefore be zero, and its bias, of course, arbitrarily high.

   (b) False.
   We should *replace* the word 't-test' with 'F-test'.

   (c) False.
   A zero coefficient only implies that the predictor is *uncorrelated* with the outcome, which is a weaker statement than statistical independence.

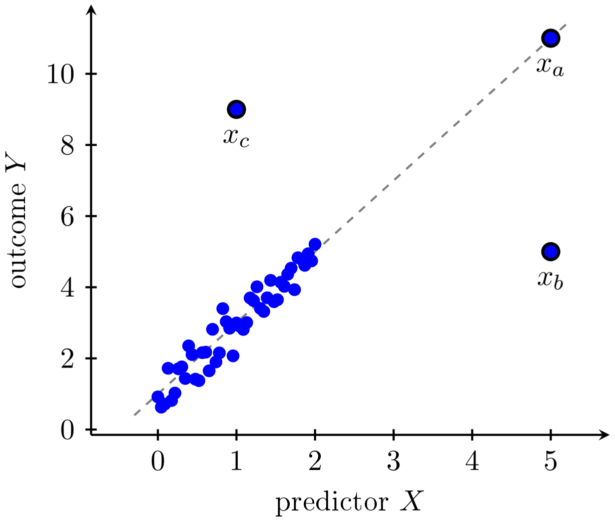
   As a simple *counter-example* consider a dataset of a predictor X and an outcome Y, represented below as points in $\mathbb{R}^2$.



   Although these are uncorrelated, i.e., $\mathbb{E}[XY] = 0$ (at least to the best of our empirical evidence), they are not statistically independent.

   (d) False.
   We can correct this statement by *replacing* 'arbitrary' noise with 'Gaussian' noise.

2. To answer this question it is helpful to quickly draw the model prediction line. The regression line for the data in the lower end looks like the gray dashed line shown below.



(a) The samples x_a and x_b are *high leverage* points, because they lie far from the mean of all predictors.
The samples x_b and x_c are *outliers*, because they lie far from the regression line.

(b) The *outliers* are samples whose outcome is far from the prediction of the model, based on the rest of the predictors, and therefore pull the regression line away from the correct estimate due to the quadratic term in the loss.

The *high leverage* points are samples whose predictors are far from the "bulk" of the remaining predictors and have a higher capacity to affect the regression line. They might or might not introduce errors depending on whether or not they are also outliers.

(c) The problem arises from outliers, so we would have to choose between points x_c and x_b. Based on our estimated regression line, however, both of these points seem to have the same residual. Therefore, we would remove the point x_b, as it is not only an outlier, but also a high leverage point, and therefore has a greater potential to affect the regression line compared to x_c; in fact, the leverage of outlier x_c is almost zero.

3. (a) Generally, in linear models heteroskedasticity does not lead to worse accuracy than homoskedastic noise of the same variance. One counter-example can easily be constructed by starting with data with homoskedastic noise and lowering the variance of it in one side while increasing it in the other, so that the overall noise is similar.

(b) The issue with heteroskadasticity is that it changes the model assumptions to the point where the statistical analysis becomes inaccurate, also in the case that the noise is Gaussian; this includes, for instance, t-tests, confidence intervals and standardised residual estimates, among others.
We should therefore always test for heteroskedasticity whenever we use the linear model for statistical analysis and not just prediction.

An easy way to test for heteroskedasticity visually is through the residual plot. For the above datasets this would look like below. Homoskedastic noise should follow the same distribution for each sample, whereas heteroskedastic one look like a funnel.

(c) Based on the course material, this question can admit more than one answers. One could be the following.

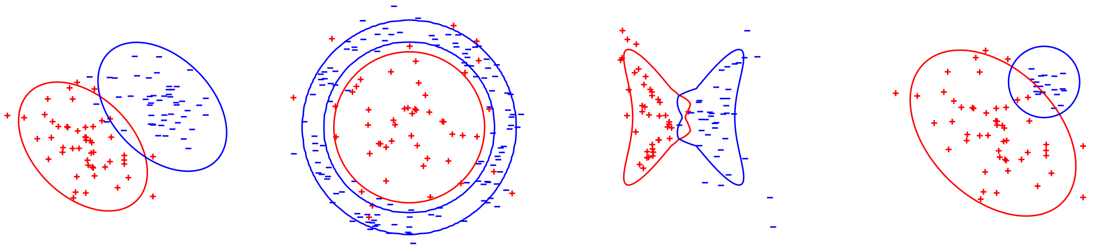
  i. You suspect that the quite small negative coefficient appears solely because of noise, whereas the true coefficient should have been zero. This would mean that the revenue is uncorrelated with the predictor for `Twutter` and is therefore not harmful for advertising.

  ii. We can test for the significance of this coefficient with a t-test. Your colleague is only confidently correct if this test manages to reject the hypothesis that the respective coefficient is zero.

PROBLEM 3 (CLASSIFICATION) (10 points)

1. Assign each of the following classification methods to one dataset shown in Figure 3 such that they achieve the best possible accuracy. Briefly explain your choices. (3 pts)

   (a) logistic regression
   (b) linear discriminant analysis
   (c) quadratic discriminant analysis
   (d) nearest neighbours with $k = 1$



(a) Dataset A (b) Dataset B (c) Dataset C (d) Dataset D

Figure 3: Likelihood contour plots for four binary classification datasets. Positive points in red (+), negative points in blue (−).

2. Consider the following statement. Is it correct? Why (not)? (2 pts)

$$\boxed{\text{Logistic Regression is a linear model.}}$$

3. One of the most commonly used kernels in SVMs is the Gaussian RBF kernel $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$. Suppose we have three points z_1, z_2 and x. Assume $\sigma = 1$, z_1 is close to x and $||z_1 - x|| \ll \sigma$, and z_2 is far from x and $||z_2 - x|| \gg \sigma$. Here the symbols $\ll$ and $\gg$ mean "much smaller" and "much greater" respectively. (2 pts)

   What is the value of $k(z_1, x)$ and $k(z_2, x)$? Choose one of the following. Explain why.

   (i) $k(z_1, x)$ will be close to 1, and
   $k(z_2, x)$ will be close to 0.

   (ii) $k(z_1, x)$ will be close to 0, and
   $k(z_2, x)$ will be close to 1.

   (iii) $k(z_1, x)$ will be close to c_1 such that $c_1 \gg 1$, and
   $k(z_2, x)$ will be close to c_2 such that $c_2 \ll 0$ and $c_1, c_2 \in \mathbb{R}$.

   (iv) $k(z_1, x)$ will be close to c_1 such that $c_1 \ll 0$, and
   $k(z_2, x)$ will be close to c_2 such that $c_2 \gg 1$ and $c_1, c_2 \in \mathbb{R}$.

4. You are training a hard-margin SVM on the dataset shown in Fig. 4.

   (a) Find the optimal weight vector $\mathbf{w}$ and bias b. What is the equation corresponding to the decision boundary? (2 pts)

   (b) Circle the support vectors and draw the decision boundary. Note: Do *not* provide your answer here, but rather on Fig. 1 on page 10 of the *answer sheet*. (1 pt)

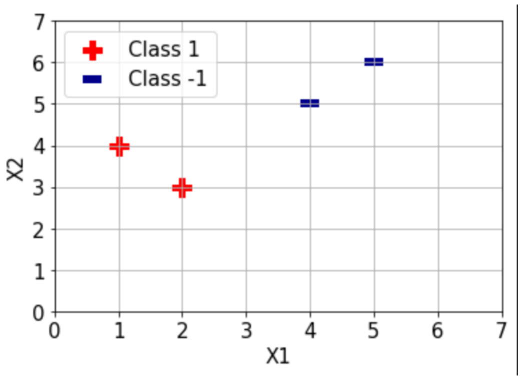


Figure 4: Dataset of two positive (+) and two negative (−) datapoints.

Solution.

1. We should make the following choices.
   - Dataset A: linear discriminant analysis
     The LDA method is optimal when the class likelihoods are Gaussian with the same variance for both classes, as is clearly the case in this dataset.
   - Dataset B: nearest neighbours with $k = 1$
     This dataset is far from linearly separable, and we therefore need a method that is flexible enough to adapt to the peculiarities of each class likelihood. The k-NN method is such a case, and even more so when $k = 1$.
   - Dataset C: logistic regression
     Although this dataset is close to linearly separable, its likelihoods are far from Gaussian and other methods cannot be used, or are sub optimal. Instead, logistic regression makes no particular assumptions on the shape of the likelihoods, apart from the linearity of the log-odds.
   - Dataset D: quadratic discriminant analysis
     The QDA method assumes the class likelihood to be the Gaussian with the different variance for both classes, as is clearly the case in this dataset.

2. Logistic regression is indeed a linear model, and more specifically belongs to the class of *generalised linear models*; in these methods we model a derived quantity with a linear model.

   In the case of logistic regression we use a linear function to model the log odds, which gives rise to the linear relationship

$$\log \text{odd}(X) := \log\left(\frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=0|X)}\right) = X\beta + \beta_0 \,. \tag{3.1}$$

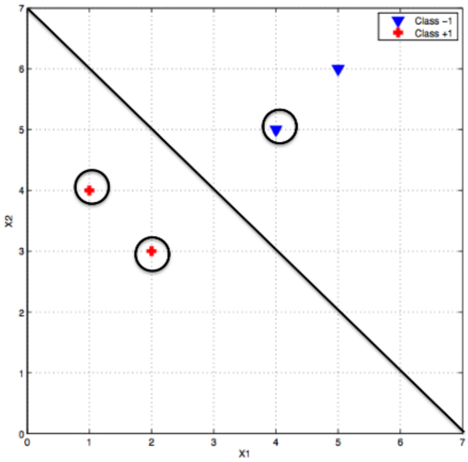
3. By simple mathematical operations, we can see that the correct answer is (i).

4. - The SVMs try to maximise the margin between the two classes. Therefore, the optimal decision boundary must be a diagonal line that crosses the point $(x_1, x_2) = (3, 4)$. It is perpendicular to the line beetween the support vectors $(4, 5)$ and $(2, 3)$, hence it has slope $m = -1$. Thus the line equation is $x_2 - 4 = -1(x_1 - 3)$ which is $x_1 + x_2 = 7$. From this equation, we can deduce that the weight vector has to be of the form (w_1, w_2) where $w_1 = w_2$. It also has to satisfy the following equations:

$$2w_1 + 3w_2 + b = 1 \quad \text{and}$$
$$4w_1 + 5w_2 + b = -1$$

     Hence, $w_1 = w_2 = -1/2$ and $b = 7/2$.

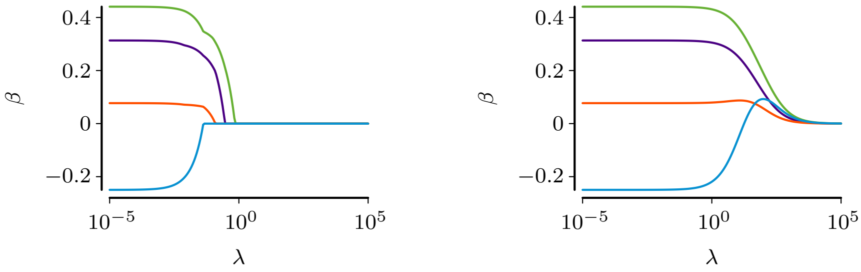
- The decision boundary and support vectors are as shown below.



5. RBF kernel generates a "bump" around the center x. For points z_1 close to the center of the bump, $K(z_1, x)$ will be close to 1, for points away from the center of the bump $K(z_2, x)$ will be close to 0.

Problem 4 (Model Selection) **(10 points)**

1. We have a dataset where we want to predict Y given four predictors, and consider a simple linear regression model with coefficients β.



(a) Regularization method A. (b) Regularization method B.

Figure 5: Linear coefficients β per predictor for varying regularization strength λ.

   (a) Which of the plots in Fig. 5a, 5b corresponds to Ridge and which to Lasso? Explain your reasoning. (2 pts)

   (b) Assume that we know that Y is influenced by only few predictors. Which of the two methods would you then prefer? Why?
   Explain how you would interpret the plots in Fig. 5 in this case. (2 pts)

   (c) Compare the behavior, in terms of bias and variance, of linear models with ridge regularization strengths $\lambda = 0.1, \lambda = 1$, and $\lambda = 10$. (1 pt)

   (d) Which approach can you use to select an appropriate tuning parameter λ? Explain this approach in your own words. (1 pt)

2. Consider the following regression objective for n datapoints and p predictors, (2 pts)

$$\hat{\beta} = \arg\min_\beta \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 \right) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \, .$$

Explain how this is different from the linear regression objectives you have encountered in the lecture, and what behavior you expect for this model.

3. Finally, we consider standard regression splines (i.e. not smoothing splines). Give two ways to constrain the flexibility of the model. Explain how this is similar or different to linear and polynomial regression. (2 pts)

Solution.

1. Comparing Ridge and Lasso,

   (a) Left: Lasso. This is because Lasso will make parameters equal to exactly zero already for moderate amount of regularization.

   Right: Ridge. This does not make parameters *exactly* zero.

   (b) If only few parameters are truly important, then the feature selection performed by Lasso would make it the obvious choice.

   (c) The larger the regularization parameter, the lower the variance of the fit models since regularization makes all model parameters closer to zero and therefore closer to each other for different datasets. In contrast, the bias is increased for the same reason of model parameters being closer to zero and therefore being unable to capture models requiring large parameter values.

   (d) We can select the tuning parameter λ by using K-fold cross validation. Here, we split the data into K different batches and train K different models with one batch left out on the $K-1$ remaining batches. By averaging the losses of the different models over their left out batches, we can estimate the generalization error of the model.

   We can use this to estimate the effect of λ on the generalization error, and use whichever λ leads to the lowest estimate of the generalization error.

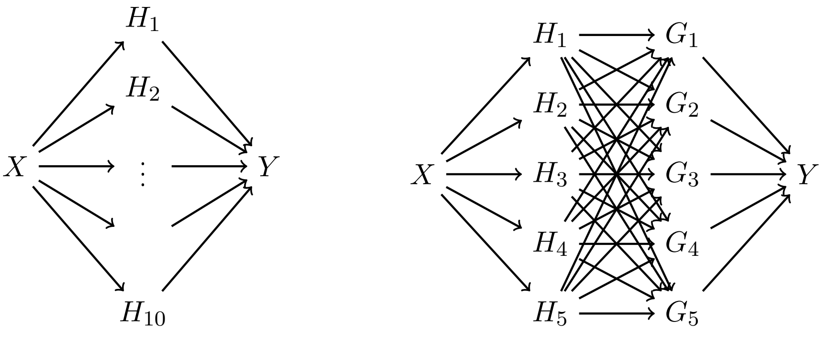
2. This adds a regularization term which is a combination of Ridge and Lasso regularization and therefore should perform regularization intermediate between these two. In particular, when $\lambda_1 \gg \lambda_2$ we expect the regularization to perform mostly like Lasso, whereas for $\lambda_2 \gg \lambda_1$ we expect it to perform mostly like Ridge. This holds in particular when one of the $\lambda_i = 0$.

3. There are (at least) three distinct ways of constraining the flexibility of a regression spline. First, we can control the number of knots, K, which affects the flexibility by introducing additional local models. This has no direct correspondence to linear or polynomial regression, although the effect on the free parameter count of the model is similar to a change in the degree d of a polynomial regression model.

   Second, we can control the degree d of the local polynomials. This directly affects the flexibility of the model at every point, and is similar to a change in the degree of a polynomial regression model.

   Third, we can change the continuity requirements at each knot. That is, instead of requiring $d-1$-times differentiability at each knot with degree d polynomials, we could require $c < d-1$-times differentiability. This does not really have any direct correspondence to linear or polynomial regression models which are already infinitely differentiable. However, the effect on the free parameter count is again similar to a change in the degree of a polynomial regression model.

PROBLEM 5 (ALL THOSE PARAMETERS) **(10 points)**



(a) Network with 1 hidden layer. (b) Network with 2 hidden layers.

Figure 6: Two neural networks with 10 hidden neurons (bias neurons not shown).

1. Consider the neural networks in Figure 6.

   (a) How many free parameters does the network in Fig. 6a have, if X and Y are univariate, and biases are *non-zero*. Explain your reasoning. (1 pts)

   (b) Explain which of the two networks is more expressive when using the sigmoid activation function $\sigma(t) = \frac{e^t}{1+e^t}$. (2 pts)

2. Yunn LeCann says that deeper networks are better, and proposes to use a neural network with 3 hidden layers and a total of 50 free parameters.

   (a) How many knots (K) would we have to pick for a cubic spline to have 50 free parameters? How many for a linear spline? Explain your reasoning. (2 pts)

   (b) The linear spline, the cubic spline, and the neural network that Yunn LeCann proposes all have 50 free parameters. Does this mean they will fit a given data set equally well? If so, explain why. If not, give a counter example on which one of them performs better than *at least* one of the other two models. (2 pts)

3. Let M_1 and M_2 be two model classes such that the ratio of number of free parameters $\frac{\text{FP}(M_2)}{\text{FP}(M_1)} > C$ is very large. Is it possible for M_1 to perform better in terms of generalization than M_2 even for arbitrarily large C? Why (not)? (1 pt)

4. We have two models that obtain exactly the same error on a held out test set. Give three reasons why one model may nevertheless be preferable to the other and explain your reasoning. (2 pts)

Solution.

1. (a) For regression, the model depicted by the network in Fig. 6a can be written as

$$Y = W_2 h_1 + b_2$$
$$h_1 = \sigma(W_1 x + b_1)$$

where $W_1 \in \mathbb{R}^{10\times 1}, b_1 \in \mathbb{R}^{10}$ and $W_2 \in \mathbb{R}^{1\times 10}, b_2 \in \mathbb{R}$. That is, there are ten parameters for W_1 corresponding to edges from X to H, ten for W_2 for the edges H to Y, ten for the biases b_1 of H and one parameter for the bias b_2 of Y. The network in Fig. 6a thus has 31 parameters total.

   (b) The model in Fig. 6b is more expressive for two reasons. First, the addition of another introduces another source of nonlinearity and therefore allows the model to capture more highly nonlinear functions. Second, the model also contains clearly more parameters, with the connections between the two hidden layers already making up for 25 parameters, which combined with its eleven biases is already larger than that of Fig. 6a.

2. (a) For 50 free parameters, we can use the formula that a spline of degree d with K knots contains $d + 1 + K$ free parameters. A spline of degree d with K knots that is $d - 1$ times differentiable has

$$(d + 1) \cdot (K + 1) - d \cdot K = dK + d + K + 1 - dK = d + K + 1$$

free parameters as we have seen both in the slides and in Assignment 4.

Thus, for the cubic spline we obtain $50 - 3 - 1 = 46$ knots and for the linear spline we obtain $50 - 1 - 1 = 48$ knots.

   (b) They do not. The simplest example is the noiseless relationship $y = x^3$. Clearly, this function would be fit by a cubic spline. However, since it is not piecewise linear, the linear spline will not perform as well on this kind of data.

3. Yes this is possible, and we can use the same example from above. Even with $K = 0$ knots, the cubic spline will still fit the data perfectly, while the linear spline will not fit the data perfectly for any finite number of knots.

4. Three possible reasons for preferring one model over another might be: a) privacy, because one model might leak more information about its data than another when shared; b) fairness, because one model might be trained on sensitive attributes such as race or gender; c) interpretability, because one model could be much more interpretable than the other.