Quiz: ISRL (chapter-6) **Student name**:

Let $\mathcal{M}_A$ be a polynomial regression model fitted to some dataset $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^n$ (where each observation is just a single scalar feature x_i and a single scalar label y_i), with resulting parameter values $\widehat{\boldsymbol{\beta}}_A = (0, 0.5, 3, 0)$. These parameter values produce the following prediction function:

$$\widehat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3 = 0.5X + 3X^2$$

Let $\mathcal{M}_B$ be another polynomial regression model fitted to the same dataset, with resulting parameter values $\widehat{\boldsymbol{\beta}}_B = (0.5, 0.5, 0.5, 0.5)$, producing the following prediction function:

$$\widehat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3 = 0.5 + 0.5X + 0.5X^2 + 0.5X^3$$

Question: (2.5 points) **Compute the following norms (L1 and L2) for** $\mathcal{M}_A$ and $\mathcal{M}_B$: (where $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j| \quad and \quad \|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2$) ()

a. $\|\boldsymbol{\beta}_A\|_1$ (b) $\|\boldsymbol{\beta}_A\|_2^2$

c. $\|\boldsymbol{\beta}_B\|_1$ (d) $\|\boldsymbol{\beta}_B\|_2^2$

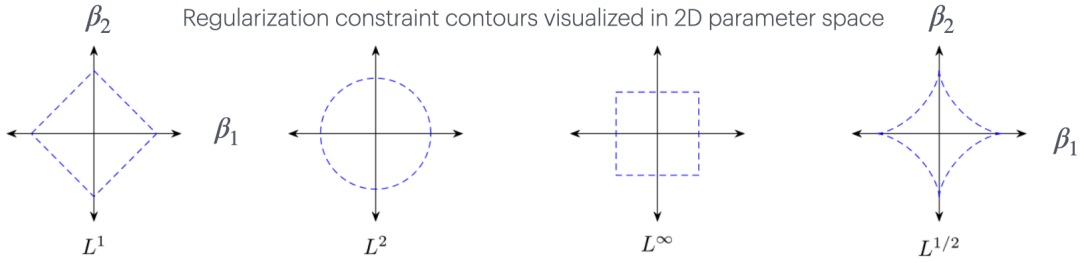
Question: (2.5 points) Assume that both models result in the exact same non-regularized loss value (for example, the exact same Mean Squared Error value). Based on your answers in the previous question.

a. Which model is "more optimal" for fixed Ridge regularization penalty $\lambda > 0$? ○ $\mathcal{M}_A$ ○ $\mathcal{M}_B$
b. Which model is "more optimal" for a fixed Lasso regularization penalty $\lambda > 0$? ○ $\mathcal{M}_A$ ○ $\mathcal{M}_B$

Question: (2.5 points) In one or two sentences, describe what model sparsity is in the context of a linear model, and why it is useful for model selection? Use equations as needed to aid your description.

Question: (2.5 point) Which regularization method is more likely to inject sparsity? (L1) or (L2)

Question: (2.5 points) Model sparsity occurs when the regularization constraint contours have sharp, axis-aligned corners. This is because the optimization contour is most likely to touch the feasible region at a point where one or more coefficients are exactly zero. The sharper the corner, the stronger the sparsity. The following plots display the 2 -dimensional unit circles produced by four different norms- L^1, L^2, L^∞, and $L^{1/2}$-which are commonly used to quantify the complexity of a model based on its estimated parameter values $\widehat{\boldsymbol{\beta}}$ (since the plots here are 2 -dimensional, $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \widehat{\beta}_2\right)$, but the norms can be computed just as well for higher-dimensional $\boldsymbol{\beta}$ vectors!) [Hint] Another way to phrase this question is: if model complexity is constrained such that the parameter vector $\boldsymbol{\beta}$ must lie within the unit circle, which of the four unit circles is most likely to produce an optimized-under-constraints $\boldsymbol{\beta}$ such that $\beta_1 = 0$ and / or $\beta_2 = 0$?



Based on these figures, which of the norms would most "aggressively" shrink parameters $\beta_j \in \boldsymbol{\beta}$ down to 0 (thus removing feature j from the model), as the complexity penalty λ is increased? (Select one)

a. ○ L^1 (b) ○ L^2
b. ○ L^∞ (d) ○ $L^{1/2}$

Question: You are building a linear regression model using **forward stepwise selection** (FSS-algorithm) (where the FSS-algorithm uses RSS to select the optimal model for each case) . We have predictors: $X_1,\ X_2,\ X_3,\ X_4$

The algorithm uses OLS to obtain the following values from the training set. Note that true Forward selection would not have evaluated all of these options, although best subset selection would.

1-variable models	2-variable models	3-variable models	4-variable models
• $Y \sim X_1$: RSS = 110, BIC = 220 • $Y \sim X_2$: RSS = 95, BIC = 210 • $Y \sim X_3$: RSS = 105, BIC = 218 • $Y \sim X_4$: RSS = 100, BIC = 215	• $Y \sim X_1 + X_2$: RSS = 85, BIC = 205 • $Y \sim X_1 + X_3$: RSS = 90, BIC = 210 • $Y \sim X_1 + X_4$: RSS = 88, BIC = 208 • $Y \sim X_2 + X_3$: RSS = 75, BIC = 195 • $Y \sim X_2 + X_4$: RSS = 80, BIC = 200 • $Y \sim X_3 + X_4$: RSS = 92, BIC = 212	• $Y \sim X_1 + X_2 + X_3$: RSS = 70, BIC = 200 • $Y \sim X_1 + X_2 + X_4$: RSS = 72, BIC = 201 • $Y \sim X_1 + X_3 + X_4$: RSS = 74, BIC = 203 • $Y \sim X_2 + X_3 + X_4$: RSS = 65, BIC = 198	• Full model: $Y \sim X_1 + X_2 + X_3 + X_4$: RSS = 63, BIC = 202

Which model will the FSS-algorithm select as the best model for 1-variable, 2-variable, 3-variable, and 4-variable cases? Question: (3.5 points) Right your answer in order that the FSS-algorithm would select them. Specify your choices using the same notation as in the table (e.g. $Y \approx X_1$ if would want to select that model choice)

Question: (2.5 points) At the end we want to select the final model, of the models you selected in the previous question, which model is the final optimal model using to BIC as or selection criteria metric?

Question: (2.5 points) In true forward subset selection, certain models in the 2-variable model case would never have been explored, which models would never have been evaluated by the algorithm, **and why?** i.e. which 2-variable models would be excluded from the search?

Question: (4 points) Match the following with the relevant definition, also circle whether it should be maximized or minimized.

Options: AIC (Akaike Information Criterion) , CIC (Celeriac Information Criterion) , Cp, Bp, Qp, Adjusted-R^2, R^2, (EIC) Euclidean Information Criterion, BIC (Bayesian Information Criterion)

_____ $-2\log(\hat{L}) + 2p$: A likelihood-based metric that estimates expected prediction error and penalizes models for having more parameters. **(maximized or minimize)**

_____ $-2\log(\hat{L}) + p\log(n)$: A metric derived from a famous theorem which relates various conditional probabilities, this quantity approximates the posterior probability of a model and includes a sample-size-dependent penalty. **(maximized or minimize)**

_____ $\frac{1}{n}\left(\mathrm{RSS} + 2p\hat{\sigma}^2\right)$: A statistic that compares a model's residual variance to an estimate of the true error variance to assess bias and variance trade-off. **(maximized or minimize)**

_____ $1 - \frac{(1-R^2)(n-1)}{n-p-1}$: A modified version of the standard OLS metric which quantifies correlation between targets and predictors, but also accounts for the number of predictors and sample size when measuring explained variability. **(maximized or minimize)**