Deadline: Th. Nov. 22, 10:00 am

On this sheet you can earn up to 4 bonus points.

Exercise 1 (Hyperparameters - 5+2 points)**.**

1. (3p) Why can hyperparameters in general not be trained directly via gradient descent?

2. (4p) Consider the regularized loss

$$\ell_\lambda(\beta) = \lambda_0 \,\mathrm{MSE}(\hat{y}(\beta)) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

   Why is a grid-search with $\lambda_i \in \{1, 2, 4\}$ inefficient? How many of the HP combinations are redundant? How does random search avoid this problem?

Exercise 2 (Ridge Regression - 5+2 points)**.**

1. (3p) Write down the gradient descent update for Ridge-Regression in vectorized form!

2. (4p) The **condition number** κ of a symmetric matrix is defined as

$$\kappa(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \tag{1}$$

   where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest (in terms of absolute value) eigenvalue of A. The condition number measures how sensitive the solution x of the linear system $y = Ax$ is with respect to changes in y.

   A matrix is said to be **well/ill-conditioned** if its condition number is **small/large**.[1] Ill conditioned matrices can lead to numerical instabilities during computations! Show that the modified matrix $\tilde{A} = X^\intercal X + 2\lambda I$ of Ridge-Regression with $\lambda > 0$ is always better conditioned than $A = X^\intercal X$.

   Hint: If $\lambda_{\max} \geq \ldots \geq \lambda_{\min} \geq 0$ are the eigenvalues of A (note that A is positive semi-definite!), then what are the eigenvalues of $\tilde{A}$?

Exercise 3 (Model Selection - 10 points)**.** The most common data model in statistics is

$$y = f(x) + \epsilon \tag{2}$$

where ϵ is typically random white noise $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. In particular, one can rewrite y as a random variable given by

$$p(y|x) = \mathcal{N}(y|f(x), \sigma^2) \tag{3}$$

[1] It holds that $1 \leq \kappa(A) \leq \infty$. Note that A is invertible if and only if $\kappa(A) < \infty$

In linear regression one assumes that $\hat{f}(x) = w^\intercal x$, or, more generally[2], $\hat{f}(x) = w^\intercal \phi(x)$ is a useful estimator for f. This parametrization yields, with $\theta = (w, \sigma)$, the linear regression model

$$p(y|x, \theta) = \mathcal{N}(y|w^\intercal x, \sigma^2) \tag{4}$$

1. (3p) Show that the **log-likelihood** $\ell(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(y_i|x_i, \theta)$ is equal to

$$-\frac{1}{2} N \Big(\frac{1}{\sigma^2} \operatorname{MSE}(\hat{f}) + \log\big(2\pi\sigma^2\big) \Big) \tag{5}$$

2. (2p) In Tutorial 3.2 we have already seen that the $\hat{w}$ which maximizes $\ell(\theta)$ can be found by solving $\frac{\partial}{\partial w} \operatorname{MSE}(\hat{f}) = 0$, i.e. the normal equation $X^\intercal X w = X^\intercal y$.

   Show that the $\hat{\sigma}$ which maximizes $\ell(\theta)$ is given by

$$\hat{\sigma}^2 = \operatorname{MSE}(\hat{f})$$

   In particular the maximum likelihood estimator (MLE) for the linear regression model is given by

$$\ell(\hat{\theta}) = \ell(\hat{w}, \hat{\sigma}) = -\frac{1}{2} N \Big(1 + \log\big(2\pi \operatorname{MSE}(\hat{f})\big) \Big) \tag{6}$$

3. (2p) Show that the BIC of the linear regression model is equal to

$$-\frac{1}{2} N \log\big(\operatorname{MSE}(\hat{f})\big) - \frac{1}{2} D \log(N) + cN \tag{7}$$

   where c is some constant.

   Note: The constant term is typically dropped because it plays no role when comparing the BIC of different models on the same dataset.

4. (3p) Given data $(x_i, y_i)_{i=1\ldots10}$, 3 models have been fitted to the data. A linear model $\hat{y}_1(x) = a_1 x + a_0$, a quadratic model $\hat{y}_2(x) = b_2 x^2 + b_1 x + b_0$ and a polynomial of degree 5: $\hat{y}_3(x)$. Which one has the best BIC ?

[2] For example we have already seen $\phi(x) = [x, 1]^\intercal$, $\phi(x) = [x^2, x^1, x^0]^\intercal$ and $\phi(x, y) = [x^2, 2xy, y^2]^\intercal$

x	y	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.00	1.90	0.98	2.02	1.90
0.28	2.04	1.41	1.76	2.07
0.56	1.83	1.84	1.67	1.72
0.83	1.56	2.27	1.75	1.53
1.11	1.29	2.70	2.01	1.76
1.39	2.84	3.13	2.44	2.39
1.67	3.49	3.56	3.04	3.22
1.94	3.33	3.99	3.82	4.05
2.22	5.26	4.42	4.76	4.81
2.50	5.61	4.85	5.88	5.70