CSE 446/546 Winter 2025 Midterm Exam

Feburary 7, 2025

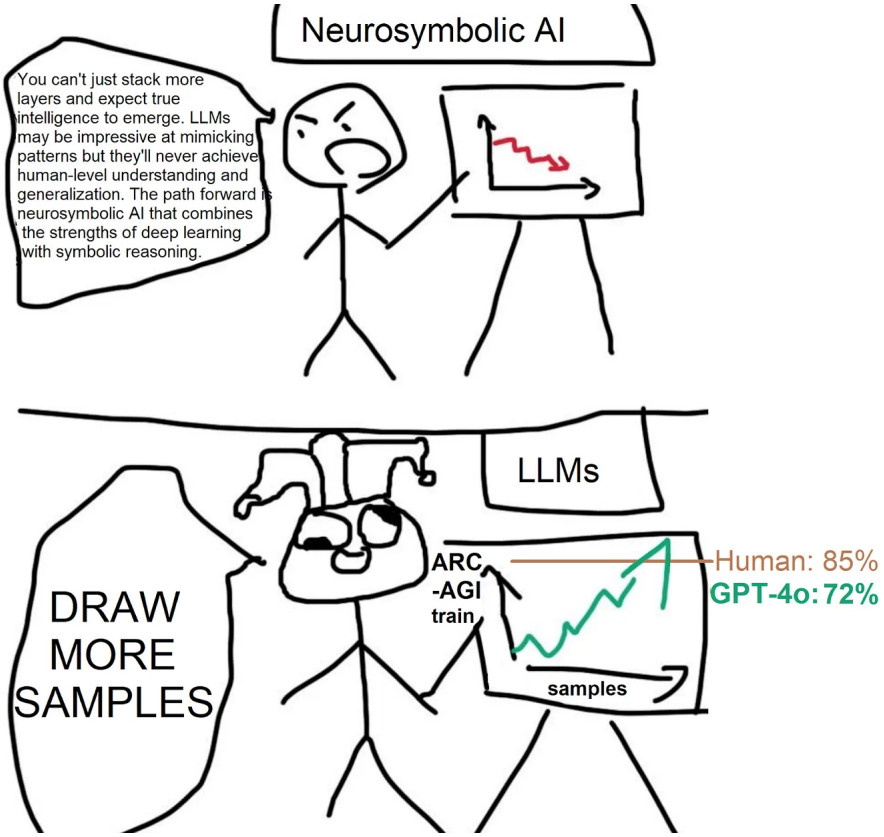
Name _____ **UW NetID** _____

Please **wait** to open the exam until you are instructed to begin, and please take out your Husky Card and have it accessible when you turn in your exam.

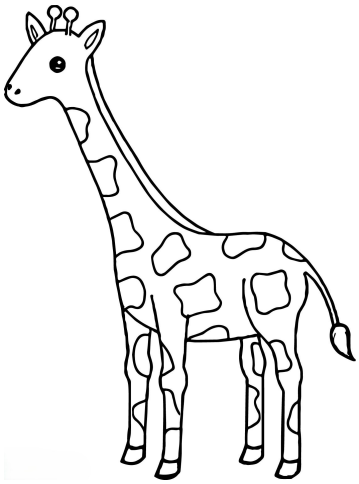
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with One Answer should only be marked with one answer. All other multiple choice questions are Select All That Apply, in which case any number of answers may be selected (**including none, one, or more**).
- For Select All That Apply questions, you will receive proportional credit for each option based on whether you get each "option" correct/incorrect. For example if there are 4 options, you will receive 0.25 points for each option that matches the solution.
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, please very clearly indicate what your final answer is and what you would like graded. Responses where we cannot determine the selected option will be marked as incorrect.
- Please remain in your seats for the last 10 minutes of the exam. If you complete the exam before the last 10 minutes, you may turn in your exam to a TA.

These images are included only to cover the back of this page. They have no relation to the exam.
Comic from Natasha:



Giraffe for good luck :)



1. 1 points | One Answer

In a popular gacha game, the probability of pulling an SSR character on a single pull is 0.6% ($P = 0.006$). Assume that each pull is independent and follows a Bernoulli distribution. In such games, players often perform a 10-pull, which means making 10 pulls. Each of these 10 pulls is still independent, meaning the probability of getting an SSR in each pull remains 0.006. If you perform a 10-pull, what is the probability of pulling exactly 2 SSRs? You do not need to know what gacha game is to solve this problem.

(a) 0

(b) $1 - (1 - P)^{10} = 5.84\%$

(c) $(1 - P)^9 \times P \times 10 = 5.68\%$

(d) $(1 - P)^8 \times P^2 \times \frac{10 \times 9}{2} = 0.15\%$

Correct answers: (d)

Explanation: We model each pull as a Bernoulli trial, where the probability of success (pulling an SSR) is $P = 0.006$. Since a 10-pull consists of 10 independent trials, the number of SSRs obtained follows a Binomial distribution:

$$X \sim \text{Binomial}(n = 10, p = 0.006)$$

We want to find the probability of pulling exactly 2 SSRs, which is given by the binomial probability mass function (PMF):

$$P(X = 2) = \binom{10}{2} P^2 (1 - P)^8$$

Computing the combination:

$$\binom{10}{2} = \frac{10!}{2!(10 - 2)!} = \frac{10 \times 9}{2} = 45$$

Thus, the probability is:

$$P(X = 2) = (1 - P)^8 \times P^2 \times 45 = 0.15\%$$

Therefore, the correct answer is **D**.

A: 0

This option would be correct if it were impossible to pull an SSR. However, since the probability of obtaining an SSR is nonzero, this option is incorrect.

B: $1 - (1 - P)^{10} = 5.84\%$

This formula calculates the probability of pulling **at least 1 SSR**, computed as:

$$P(\text{at least 1 SSR}) = 1 - P(0 \text{ SSRs}) = 1 - (1 - P)^{10}.$$

This probability includes cases where the player gets 1, 2, 3, ..., or even 10 SSRs.

C: $(1 - P)^9 \times P \times 10 = 5.68\%$

This formula calculates the probability of pulling **exactly 1 SSR**, given by:

$$P(X = 1) = \binom{10}{1} P^1 (1 - P)^9 = 10 \times P \times (1 - P)^9.$$

Hence, the correct answer remains **D**.

2. 1 points | Select All That Apply

Below are several statements about Gradient Descent (GD) and Stochastic Gradient Descent (SGD). Which of the following are correct?

(a) For GD, each step aims to move along the gradient descent direction at the current point to reduce the value of the objective function.

(b) In SGD, each step computes an estimated gradient based on a single sample, introducing randomness, which may not guarantee that the objective function decreases in every step.

(c) Suppose you have model w_t at the t-th iteration of SGD. The expectation of the direction of the model update for SGD at step t is different from the negative direction of the gradient $-\nabla_w f(w)|_{w=w_t}$.

(d) GD requires the full gradient information of the objective function, while SGD only needs the gradient information on a single sample at each step.

Correct answers: (a), (b), (d)

Explanation: The correct answers are (A), (B), and (D). Below is an explanation of each option.

(A) Correct: In Gradient Descent (GD), each step moves in the direction of the negative gradient of the objective function at the current point. This ensures that the function value decreases (assuming an appropriate step size). Mathematically, the update rule for GD is:

$$w_{t+1} = w_t - \eta \nabla_w f(w_t)$$

where η is the learning rate, and $\nabla_w f(w_t)$ is the gradient of the objective function over the entire dataset at iteration t. This step guarantees movement in the direction that minimizes the objective function.

(B) Correct: In Stochastic Gradient Descent (SGD), instead of computing the exact gradient using the full dataset, an estimated gradient is computed based on a single sample (or a mini-batch). This introduces randomness in the updates, meaning that the objective function might not necessarily decrease in every step. The update rule for SGD can be rewritten as:

$$w_{t+1} = w_t - \eta(\nabla_w f(w_t) + \xi_t)$$

where $\nabla_w f(w_t)$ is the true full-batch gradient, and ξ_t is a noise term introduced due to the stochastic nature of SGD, representing the deviation from the full gradient when using only one sample. This noise term makes each individual update potentially non-optimal, but in expectation, the updates still align with the true gradient direction over multiple iterations.

(C) Incorrect: The expectation of the SGD update direction is equal to the negative full gradient:

$$\mathbb{E}[\nabla_w f(w_t) + \xi_t] = \nabla_w f(w_t)$$

Since SGD approximates the full gradient using randomly sampled data points, its update direction is an unbiased estimate of the true gradient. That is, while each step introduces noise, on average, the update follows the same direction as GD. Therefore, the claim that the expectation is different from $-\nabla_w f(w_t)$ is false.

(D) Correct: GD requires the full gradient of the objective function, meaning it computes $\nabla_w f(w)$ over the entire dataset at each step. In contrast, SGD only uses the gradient of a single sample (or a mini-batch), significantly reducing the computational cost per step, especially in large-scale datasets.

Therefore, the correct answers are **(A), (B), and (D)**.

3. 1 points

In a gacha game, the probability of obtaining an SSR character per pull is p, but p is unknown. To estimate p, Bob performed 100 pulls and obtained SSRs $k = 3$ times (i.e., 3 successes). Assume that each pull is independent and follows a Bernoulli distribution.

What is the likelihood of this this scenario occurring?

Likelihood function: $L(p) =$_____

What is the Maximum Likelihood Estimate (MLE) of p as a fractional number?

MLE: $\hat{p} =$_____

Explanation: Since each pull is independent and follows a Bernoulli distribution, the total number of SSRs obtained follows a Binomial distribution:

$$X \sim \text{Binomial}(n = 100, p)$$

The likelihood function is given by the binomial probability mass function:

$$L(p) = \binom{100}{3} p^3 (1-p)^{97} = \frac{100 \times 99 \times 98}{6} p^3 (1-p)^{97}$$

Maximum Likelihood Estimate (MLE) of p

To find the MLE $\hat{p}$, we maximize $\ln L(p)$:

$$\ln L(p) = \text{constant} + 3 \ln p + 97 \ln(1-p)$$

Taking the derivative and setting it to zero:

$$\frac{3}{p} - \frac{97}{1-p} = 0$$

Solving for p:

$$\hat{p} = \frac{3}{100}$$

Final Answers:

$$L(p) = \frac{100 \times 99 \times 98}{6} p^3 (1-p)^{97}$$

$$\hat{p} = \frac{3}{100}$$

4. 1 points | One Answer

Suppose you train a linear regression model (without doing feature expansion), i.e., $f_w(x) = wx + b$, to approximate the cubic function $g(x) = 2x^3 + 7x^2 + 4x + 3$. What's the most likely outcome?

(a) The model will have low bias and low variance

(b) The model will have low bias and high variance

(c) The model will have high bias and low variance

(d) The model will have high bias and high variance

Correct answers: (c)

Explanation: The model being linear means the variance is likely to be low.
The linear model trying to approximate a cubic function, which is of a higher degree, means that the bias is likely to be high.

5. 1 points | One Answer

Adding more basis functions to a linear regression model always leads to better prediction accuracy on new, unseen data.

(a) True

(b) False

Correct answers: (b)

Explanation: As the complexity of the model increases, the prediction accuracy on new, unseen data (test data) doesn't always get better as the model may overfit.

6. [1 points] [One Answer]

What datasets from the training/validation/test data split should you utilize during hyperparameter tuning?

(a) Training Data

(b) Training Data, Validation Data

(c) Training Data, Validation Data, Test Data

(d) Training Data, Test Data

Correct answers: (b)

Explanation: You never want to use your Test Data for hyperparameter tuning, as it will bias the model on the test data. The test data should only be used for final evaluation of the model.
Instead, for hyperparameter tuning, you want to train your model with different hyperparameters using the Training Data, then evaluate those models on the Validation Data to select the best hyperparameters for the model. (Methods like K-fold Cross Validation)

7. [1 points] [One Answer]

Consider $u = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$, $v = \begin{bmatrix} -4 \\ 5 \\ 1 \end{bmatrix}$, $w = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$. Let $x \in \mathbb{R}^3$. Does there exist unique $a, b, c \in \mathbb{R}$ such that $a \cdot u + b \cdot v + c \cdot w = x$?

(a) Yes

(b) No

(c) Not enough information to determine

Correct answers: (a)

Explanation: u, v, w are linearly independent, so the function $f(a, b, c) = a \cdot u + b \cdot v + c \cdot w$ is onto $\mathbb{R}^3$ as well as one-to-one. This can be verified by using row reduction on $\begin{bmatrix} u & v & w \end{bmatrix}$ or by noticing the three vectors are orthogonal.

8. 1 points

Consider data matrix $X \in \mathbb{R}^{n \times d}$, label vector $y \in \mathbb{R}^n$, and regularization parameter $\lambda > 0$. Write the closed form solution for ridge regression.

Answer: _____

Explanation: $(X^\top X + \lambda I)^{-1} X^\top y$

9. 1 points | One Answer

Consider a dataset containing three observations for a simple linear regression problem, where y is the dependent variable and x is the independent variable. The dataset is given as follows:

x	y
1	7
2	8
3	9

Find the coefficient β_1 of the linear regression (without bias) $y = \beta_1 x$ using the least squares as loss.

a) $\frac{46}{14}$

b) $\frac{14}{46}$

c) $\frac{50}{14}$

d) $\frac{14}{50}$

Correct answers: (c)

Explanation: $\beta_1 = \frac{50}{14}$

$\beta_1 = (X^\top X)^{-1}(X^\top Y) = \frac{1}{14} \times 50$

10. 1 points | One Answer

We can find the solution for LASSO by setting the gradient of the loss to 0 and solving for weight parameter w.

(a) True

(b) False

Correct answers: (b)

Explanation: LASSO has no closed form solution, which is why we use gradient descent.

11. 1 points | One Answer

You are building a model to detect fraudulent transactions from a dataset of 100K samples. What would be the most effective way to split and utilize your data?

(a) Randomly take an 80-20 data split. Use 80% of the data for training, and 20% for both validation and evaluation.

(b) Use the first 80% of the data for training, the next 10% for validation, and the last 10% for evaluation.

(c) Randomly make a 80-10-10 data split. Use 80% of the data for training, 10% for validation, and 10% for evaluation.

(d) Select a random 80% of the data for training, use the remaining 20% for validation. Evaluate on the training set.

Correct answers: (c)

Explanation: A is incorrect since the validation and test set should be separate. B is incorrect since the data splits should be randomized. C is the correct as it is the standard data split method. D is incorrect since evaluating on the known train set induces bias.

12. 1 points | One Answer

You are implementing a model to predict house prices. Your dataset contains 15 features (e.g., location, acres, proximity to city, etc.). However, you believe that many of these features are irrelevant to the house prices. Which method would be most suitable for your model?

(a) Logistic regression with L1 regularization.

(b) Logistic regression with L2 regularization.

(c) Linear regression with L1 regularization.

(d) Linear regression with L2 regularization.

Correct answers: (c)

Explanation: Since this is a regression not classification task, we use linear regression. Additionally, since we are interested in some kind of feature selection, L1 regularization would be more effective in setting unimportant features to 0.

13. 1 points | Select All That Apply

While training a model, you notice that it has a small bias but a high variance on the training data. Which of the following are valid strategies to address the high variance?

(a) Increase regularization constant.

(b) Train on a model class that is simpler.

(c) Increase the size of the training dataset.

(d) Use higher-degree features to capture more complex patterns in the data.

Correct answers: (a), (b), (c)

Explanation: A and B are correct because increasing regularization and simplifying model complexity help decrease the impact of less important features, improving generalization and reducing variance. C is also correct because increasing the training set size allows the model to generalize better, which can reduce variance. D is incorrect since using higher-degree features increases model complexity and often leads to overfitting, increasing variance.

14. 1 points | Select All That Apply

After a student trains and evaluates a Logistic Regression model, you notice their test accuracy is 99.99%. You know that this was supposed to be a difficult dataset to model, so you investigate. Which of the following are **reasonable explanations** for this excessively high accuracy? Note that if you select multiple answers, not all of them have to be true at the same time.

(a) There was some form of train/test leakage, resulting in the model over-performing on the test set

(b) The data was not linearly separable, making it very easy for the model to classify things correctly

(c) The dataset was incredibly imbalanced, with most of the data points being labeled as positives

(d) The dataset was incredibly imbalanced, with most of the data points being labeled as negatives

Correct answers: (a), (c), (d)

Explanation: A is true. Train/test leakage can result in incredibly high performance on the evaluation data. C and D are also true. A very imbalanced dataset can make it so that the model only predicts the majority class yet scores very high. B is false as having data that is not linearly separable does not make it easier for a linear model to separate the classes.

15. 1 points | One Answer

$W \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{n \times n}$, $Y \in \mathbb{R}^{p \times m}$, $Z \in \mathbb{R}^{m \times m}$, and $a \in \mathbb{R}^n$. If m, n, p are distinct, which one of the following expressions is valid?

(a) $(X^{-1}aa^\top W^\top)^{-1}(X^\top a)$

(b) $Xa^\top aW^\top(Z^{-1}Y^\top)$

(c) $WXaa^\top XZY^\top$

(d) None of the above

Correct answers: (b)

Explanation: A is incorrect because $X^{-1}aa^\top W^\top \in \mathbb{R}^{n \times m}$, and you cannot take the inverse of a non-square matrix. B is correct because even though $Xa^\top$ is not possible ($n \times n$, $1 \times n$), $a^\top a$ becomes a scalar and allows for the expression to be valid. C is incorrect because XZ is not possible ($n \times n$, $m \times m$). D is incorrect because B is correct.

16. [1 points]

For what value of k will k-fold cross-validation create cross-validation splits equivalent to Leave-one-out cross-validation (LOOCV)? Assume you have n data points.

$$k = \_\_\_\_\_$$

Explanation: If $k = n$, then there will be n folds, each one only leaving 1 datapoint out. This is equivalent to LOOCV

17. [1 points] [One Answer]

We can decrease the variance of a model by increasing the model complexity.

(a) True

(b) False

Correct answers: (b)

Explanation: As model complexity increases, this increases the variance error due to higher degree of freedom

18. [1 points] [Select All That Apply]

Which of the following statements are true about logistic regression? Recall that the sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ for $x \in \mathbb{R}$

(a) L2 regularization is often used to reduce overfitting in logistic regression by adding a penalty for large coefficient values

(b) The logistic sigmoid function is used to model the probability of the positive class in binary logistic regression

(c) The maximum likelihood estimates for the logistic regression coefficients can be found in closed-form

(d) For any finite input $x \in \mathbb{R}$, $\sigma(x)$ is strictly greater than 0 and strictly less than 1. Thus, a binary logistic regression model with finite input and weights can never output a probability of exactly 0 or 1, and can never achieve a training loss of exactly 0.

Correct answers: (a), (b), (d)

Explanation: The MLE for logistic regression coefficients cannot be found in closed-form which is why an iterative approach (eg. SGD) is used to find the coefficients.

(d) True. $\sigma(x)$ has horizontal asymptotes at 0 and 1 and therefore is strictly bounded between those values. Because the output probability is the output of , this implies that the output probability is also strictly contained in $(0, 1)$. As it cannot output positive or negative labels with probability 1, it is therefore unable to reduce the training loss to exactly 0, though it can get arbitrarily close.

19. 1 points | Select All That Apply

Below are several statements about the train/test/validation sets and cross-validation. Which of the following are correct?

(a) k-fold cross validation (where $k > 1$) is faster but more biased than leave-one-out (LOO) cross validation.

(b) k-fold cross validation (where $k > 1$) is faster and more accurate than leave-one-out (LOO) cross validation.

(c) The test set can be used to evaluate models during training and for hyperparameter tuning.

(d) The test error gives us an assessment of how our model does on unseen data.

Correct answers: (a), (d)

Explanation: A is correct since k-fold is faster but generally more biased. D is correct since the purpose of the test set is to test the model on unseen data and assess its performance.

20. 1 points | Select All That Apply

Consider the principle of Maximum Likelihood Estimation (MLE), which is a method to estimate the parameters of a statistical model. Which of the following statements is correct?

(a) For MLE, samples must be drawn i.i.d. (independent and identically distributed).

(b) Once we have a log-likelihood function, we maximize it with respect to the parameter θ to find the parameter estimate $\hat{\theta}_{MLE}$.

(c) MLE always provides an unbiased estimator of the true parameter.

(d) MLE identifies the model parameters that maximize the likelihood of the observed data.

Correct answers: (b), (d)

Explanation: While i.i.d. is commonly assumed when doing MLE, it is not strictly necessary. Additionally, although it can sometimes be unbiased, MLE is generally a biased estimator.

21. 1 points | One Answer

If we run gradient descent on $f(x)$, gradient descent guarantees that we will converge to the global minimum even if $\nabla^2 f(x) \succeq 0$ does not hold some x, i.e., the Hessian of the objective function is not positive semi-definite for some x.

(a) True

(b) False

Correct answers: (b)

Explanation: Gradient descent only guarantees a global minimum if $f(x)$ is convex.

22. 1 points | One Answer

Let $A_1, A_2, ..., A_n \sim \mathcal{N}(\mu, \sigma^2)$. What is $\mathbb{E}[A_1 + A_2 + A_3]$?

(a) 3μ

(b) 6μ

(c) 9μ

(d) Cannot be determined from the given information.

Correct answers: (a)

Explanation: By linearity of expectation, $\mathbb{E}[A_1 + A_2 + A_3] = \mathbb{E}[A_1] + \mathbb{E}[A_2] + \mathbb{E}[A_3]$. Since $A_1, A_2, ..., A_n \sim \mathcal{N}(\mu, \sigma^2), \mathbb{E}[A_1] = \mathbb{E}[A_2] = \mathbb{E}[A_3] = \mu$. Thus, $\mathbb{E}[A_1] + \mathbb{E}[A_2] + \mathbb{E}[A_3] = \mu + \mu + \mu = 3\mu$.

23. 1 points

Consider a function f(x,y) representing a loss function in a 2-dimensional space, where gradient descent is used to minimize f. Given the function: $f(x, y) = x^2 + 2y^2 + 4xy$ where the initial point is $(x_0, y_0) = (1, 1)$ and the learning rate is 0.1, write down the (x_1, y_1) you get after one step of gradient descent.

Answer: _____

Explanation: From the gradient descent algorithm: $x_1 = x_0 - \eta \cdot \frac{\partial f(x_0, y_0)}{\partial x}$ and $y_1 = y_0 - \eta \cdot \frac{\partial f(x_0, y_0)}{\partial y}$. It is given that $\eta = 0.1$. $\frac{\partial f(x,y)}{\partial x} = 2x + 4y$. $\frac{\partial f(x,y)}{\partial y} = 4y + 4x$. So, $x_1 = 1 - 0.1 \cdot 6 = 0.4$ and $y_1 = 1 - 0.1 \cdot 8 = 0.2$.

Final answer: $(x_1, y_1) = (0.4, 0.2)$

24. 1 points | One Answer

For machine learning models and datasets in general, as the number of training data points grows, the prediction error of the model on unseen data (data not found in the training set) eventually reaches 0.

(a) True

(b) False

Correct answers: (b)

Explanation: There is irreducible error that leads to non-zero error.

25. 1 points | One Answer

Which of the following statements about ridge regression are true?

(a) When there are correlated features, ridge regression typically sets the weights of all but one of the correlated features to 0.

(b) Compared to unregularized linear regression, the additional computational cost of ridge regression increases proportional to the number of data points in the dataset.

(c) Ridge regression reduces variance at the expense of increasing bias.

(d) Using ridge and lasso regularization together (e.g., minimizing a training objective of the form $f(w) = \sum_{i=1}^{n}(y^{(i)} - {x^{(i)}}^{\top} w)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$) makes the training loss no longer convex.

Correct answers: (c)

Explanation: Ridge regression typically shrinks the weights of correlated features about evenly. This means it probably won't set the weights of all but one of the correlated features to 0. That would be more akin to LASSO regression. The additional computational cost increases proportional to the number of weights in the dataset. Ridge-regression is an example of the bias-variance trade off. The sum of convex functions is convex so ridge and LASSO regression combined is still convex.

26. 1 points | Select All That Apply

Let $n \in \mathbb{N}$ such that $n > 1$. Which of the following functions are convex (with respect to x) over its entire domain?

(a) $f(x) = 5 + \sum_{i=1}^{n} x^{2i}$

(b) $f(x) = 5 + \sum_{i=1}^{n} x^{2i+1}$

(c) $f(x) = 3 \cdot e^{-\frac{x^2}{n}}$

(d) $f(x) = x - \log_\pi(x^n)$ on $(0, \infty)$

Correct answers: (a), (d)

Explanation: Below is an explanation for each option:

a) Even (positive) monomials are convex, and a sum over convex functions is convex.

b) Similarly, odd monomials are strictly concave on $(-\infty, 0)$, so their sum will be strictly concave on $(-\infty, 0)$.

c) $3 \cdot e^{-\frac{x^2}{n}}$ is not convex (recall the shape of the Gaussian).

d) $\log x$ is concave, and x is convex, so $x - \log_\pi(x^n) = x - n \cdot \log_\pi(x)$ is convex on $(0, \infty)$.

27. 1 points | Assume $n \neq d$. Suppose $x_1, x_2, ..., x_n$ span $\mathbb{R}^d$. What is the rank of $\sum_{i=1}^{n} x_i x_i^\top$? Write your answer in terms of n and d. Hint: for any matrix A, $\text{rank}(A^\top A) = \text{rank}(A)$.

Answer: _____

Explanation: The answer is d. Let $X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}$. Note that since $x_1, x_2, ..., x_n$ span $\mathbb{R}^d$, $\text{rank}(X) = d$. Also notice $X^\top X = \sum_{i=1}^{n} x_i x_i^\top$. So

$$\begin{aligned} \text{rank}(\sum_{i=1}^{n} x_i x_i^\top) &= \text{rank}(X^\top X) \\ &= \text{rank}(X) \\ &= d \end{aligned}$$

The takeaway here is that our design matrix $X^\top X$ is full rank (invertible) if and only if our data spans $\mathbb{R}^d$.

28. [1 points]
Describe one advantage of full-batch gradient descent over mini-batch gradient descent.

Answer: _____

_____

_____

Explanation: Full batch gradient descent will be more accurate when calculating the gradient as it uses the whole dataset while mini-batch gradient descent uses a subset of the dataset to calculate gradient. This result in a more stable convergence.

29. [1 points] Describe one advantage of mini-batch stochastic gradient descent ($1 < B < n$) over stochastic gradient descent with batch size $B = 1$ (e.g., updating the parameters at each iteration based only on one randomly sampled training point).

Answer: _____

_____

_____

Explanation: Mini-batch stochastic gradient descent will be more stable as it uses a subset of the training data for gradient calculation while stochastic gradient descent with batch-size 1 calculates gradient based on only one training data point making it more susceptible to noise.

END OF EXAM