HOMEWORK 6
GRADIENT DESCENT, OPTIMIZATION, & PROBABILITY *

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy in the syllabus.
- **Late Submission Policy:** See the late submission policy in the syllabus.
- **Submitting your work:** You will use Gradescope to submit answers to all questions.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. To receive full credit, you are responsible for ensuring that your submission contains exactly the same number of pages and the same alignment as our PDF template.
  - **Latex Template:** https://www.overleaf.com/read/zmpwpzqqpjmc#e83343

Question	Points
Gradient Descent	16
Optimization	4
Probability	10
Total:	30

*Compiled on Thursday 2nd October, 2025 at 21:32

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Henry Chai

○ Marie Curie

● Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

| 10-606 |

| 10-60~~6~~7 |

1 Gradient Descent (16 points)

In this problem, you will seek minima of a nonconvex function, the Styblinski-Tang function. The form of the function is:

$$f(\mathbf{x}) = f(x_1, x_2, ..., x_D) = \frac{1}{2}\sum_{d=1}^{D}(x_d^4 - 16x_d^2 + 5x_d)$$

where $\mathbf{x} \in \mathbb{R}^D$ are the parameters of the function to be optimized. You will optimize this function using gradient descent.

1. (3 points) **Derivation:** What is the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$?

2. (7 points) The gradient descent algorithm computes the gradient at each step and steps opposite the gradient multiplied by a step size γ. Implement gradient descent for the function $f(\mathbf{x})$. Your implementation should be a Python function that accepts three parameters:

```
gd(x_initial, step_size, max_iterations)
```

where `x_initial` is a numpy vector of length D representing the initial point for gradient descent, `step_size` is the step size γ, and `max_iterations` is the total number of iterations that gradient descent should run before stopping. Your implementation should work for arbitrary D. At the end, it should print out *both* the value of $\mathbf{x}$ after running for the specified number of iterations and the value of $f(\mathbf{x})$ at that point. You should only use the `numpy` library.

3. (2 points) **Short answer:** What are the values of $f(\mathbf{x})$ and $\mathbf{x}$ found by your implementation when it is called as follows?

```
gd(np.array([5,5]), 0.01, 100)
```

4. (2 points) **Short answer:** What are the values of $f(\mathbf{x})$ and $\mathbf{x}$ found by your implementation when it is called as follows?

```
gd(np.array([-5,-5]), 0.01, 100)
```

5. (2 points) **Short answer:** Suppose you only knew that the Python function `gd(...)` was running gradient descent on a non-convex function (i.e. assume you don't know the exact form of the function $f(\mathbf{x})$ that it's optimizing). Provide **two** reasons why `gd(np.array([5,5]), 0.01, 100)` and `gd(np.array([-5,-5]), 0.01, 100)` could return different values of $\mathbf{x}$.

Reason 1

Reason 2

2 Optimization (4 points)

1. (1 point) A function $f : \mathbb{R}^n \to \mathbb{R}$ has multiple local minima x_1, x_2 for which $f(x_1) \neq f(x_2)$. Which of the following best describes this function?

   ◯ Strictly convex

   ◯ Convex but not strictly convex

   ◯ Nonconvex

   ◯ Cannot be determined

2. (1 point) Consider the function $f(x, y) = xy$ where $x, y \in \mathbb{R}$. Is f convex or nonconvex?

   ◯ Convex

   ◯ Nonconvex

3. (1 point) For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the first-order condition $\nabla f(x^*) = 0$ is which of the following for x^* to be a local minimum?

   ◯ Necessary and sufficient

   ◯ Necessary but not sufficient

   ◯ Sufficient but not necessary

   ◯ Neither necessary nor sufficient

4. (1 point) Consider two objective functions $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ where f_1 is convex and f_2 is non-convex. Gradient descent with appropriately chosen step sizes is applied to both functions, converging to stationary points x_1^* and x_2^* satisfying $\nabla f_1(x_1^*) = 0$ and $\nabla f_2(x_2^*) = 0$. Which of the following describes the optimality of these solutions?

   ◯ x_1^* is a global minimum of f_1 and x_2^* is a global minimum of f_2.

   ◯ x_1^* is guaranteed to be a global minimum of f_1; x_2^* is a local (but not necessarily global) minimum of f_2

   ◯ x_1^* and x_2^** may be either a minimum or a saddle point of f_1 and f_2 respectively

   ◯ x_1^* is guaranteed to be a global minimum of f_1; x_2^* may be a local minimum, local maximum, or saddle point of f_2.

3 Probability (10 points)

1. (2 points) Suppose we flip a fair coin 20 times, and record the sequence of heads (H) and tails (T) that is observed. Which of the following subsequences is most likely to appear within the full sequence?

   ◯ HHHH

   ◯ HHHHH

   ◯ HTHHH

   ◯ HTHTH

   ◯ HTHTHT

2. (2 points) Suppose we have a six-sided die (numbered 1 to 6) that is not evenly weighted (i.e. it is not fair). Define A to be the event of rolling a number less than 4. Define B to be the event of rolling an even number. Assume we know that $P(A) > 0.5$. Which of the following is definitely true? **Select all that apply.**

   ◯ $P(B) < 0.5$

   ◯ $P(A \wedge B) < 0.5$

   ◯ $P(A) > P(B)$

   ◯ $P(A \vee B) > P(A)$

   ◯ None of the above

3. (2 points) Assume there are three events, A, B, and C. Also, assume that $P(A \vee B) = 0.5$, and $P(C) = 0.5$. Which of the following is true? **Select all that apply.**

   ◯ A and B are disjoint

   ◯ A, B, and C are disjoint

   ◯ $P(C) > P(A \wedge B)$

   ◯ $P(A \wedge B \wedge C) < 0.5$

   ◯ None of the above

4. (4 points) Consider the following experiment: you stand on a street corner in Pittsburgh and survey people. You ask two questions: "Do you like computer science?" and "Do you like the Steelers?" You find that 67% of passers by respond "yes" to the first question, and 90% of passers by respond "yes" to the second question. Furthermore, 65% of passers by respond "yes" to both questions.

   What is the probability of finding a person who responds "no" to both questions?

4 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found in the syllabus.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer