Exam 1 Prep(Stats102B)

Derivative Rules:

Basic Derivative Rules

- **Constant Rule**

$$\frac{d}{dx}[c] = 0$$

- **Power Rule**

$$\frac{d}{dx}[x^n] = nx^{n-1} \quad \text{for any real } n$$

- **Constant Multiple Rule**

$$\frac{d}{dx}[c \cdot f(x)] = c \cdot \frac{d}{dx}[f(x)]$$

- **Sum and Difference Rule**

$$\frac{d}{dx}[f(x) \pm g(x)] = \frac{d}{dx}[f(x)] \pm \frac{d}{dx}[g(x)]$$

- **Product Rule**

$$\frac{d}{dx}[f(x) \cdot g(x)] = f'(x)g(x) + f(x)g'(x)$$

- **Quotient Rule**

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$$

Chain Rule

If $y = f(g(x))$, then:

$$\frac{dy}{dx} = f'(g(x)) \cdot g'(x)$$

Common Function Derivatives

- **Exponential Functions**

$$\frac{d}{dx}[e^x] = e^x$$

$$\frac{d}{dx}[e^{u(x)}] = e^{u(x)} \cdot u'(x)$$

- **Logarithmic Functions**

$$\frac{d}{dx}[\ln x] = \frac{1}{x}$$

$$\frac{d}{dx}[\ln(u(x))] = \frac{u'(x)}{u(x)}$$

Note: Trigonometric derivatives like $\sin$ *and* $\cos$ *are not required.*

Example: Chain Rule

Let $f(x) = \ln(3x^2 + 1)$

Then:

$$f'(x) = \frac{d}{dx}[\ln(3x^2 + 1)] = \frac{6x}{3x^2 + 1}$$

Practice Exam 1 - FRQs

Problem 1

Let $f(x) = x - log(x)$,
where log(.) donates the natural log base algorithm

1. Show that f(x) has a unique global min. Justify your answer

Find $f(x), f'(x), f''(x)$, set $f'(x) = 0$ to get the critical points, if we can show that $f'(x) = 0$ has only one critical point, and that $f''(x)$ is concave up at that point, then that will be the unique global min

$$\begin{aligned}
f(x) &= x - \ln(x) \\
f'(x) &= 1 - \frac{1}{x} \\
0 &= 1 - \frac{1}{x} \\
\frac{1}{x} &= 1 \\
x &= 1 \\
f''(x) &= \frac{1}{x^2} \\
f''(1) &= 1
\end{aligned}$$

Because $f(x)$ has one critical point at $x = 1$, and the second derivative is positive, the function is concave up, meaning that $f(x)$ has a unique global minimum at that point.

2. Let $x_0 = 2$ be the initial point used in the gradient descent algorithm. What will x_1 be based on the gradient descent algorithm, if the step size is set to $\eta = 0.5$?
3. Derive the range of values for the step size parameter η, so that gradient descent is convergent? Justify your answer.
4. Suppose that somebody that does not know how to derive the range of eligible step sizes η, decided to use an initial $\eta = 2$ for initial point $x_0 = 2$. Explain what calculations the backtracking line search algorithm will check to select and appropriate step size η_0 to proceed calculating the next update x_1.

Practice Exam 1 - MCQs

Question 1: Which direction does the gradient descent algorithm move in each iter-ation?

- Random direction
- Direction of the gradient
- **Opposite to the gradient**
- Along the eigenvectors of the Hessian

Question 2: If the step size (learning rate) for the optimization problem of a statistical model is too large, gradient descent can:

- Converge slowly
- **Not converge**
- Overfit the data
- Always converge faster

Question 3: In the context of gradient descent, a ”step size schedule” is used to:

- Randomly choose learning rates
- Ensure faster convergence
- **Decrease learning rate over time**
- Increase gradient magnitude

Question 4: In the heavy ball momentum method, the new direction is a combination of:

- Current gradient and noise
- **Previous parameter and current value**
- Previous update and current gradient
- Gradient norm and function value

Question 5: In AdaGrad, the learning rate is

- Constant for all parameters
- Increasing over time
- Decreasing for frequently updated parameters
- Randomly sampled at each step

Question 6: Why is bias correction used in ADAM?

• To make convergence faster • To compensate for initialization at zero • To prevent exploding gradients • To add noise to the gradient update

Question 7: In ADAM, the step size is:

• Fixed • Increasing • Scaled by gradient history • Decreased linearly

Question 8: Mini-batch SGD uses:

• The entire data set 5 • Only one sample • A subset of data points • Data sorted by the objective function value

Question 9: What is a common downside of very small batch size SGD?

• Too slow • Too stable • High variance in updates • Uses entire data set per iteration