Partial Derivative.

$f(x,y) := x^3 + 2y^2 + 3x^2y^4$

$f_x := \frac{d}{dx} f(x,y)$

Quadratic Forms.

Every quadratic form can be expressed as x^TAx where A is symmetric matrix.

What quadratic form corresponds to the matrix $\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$

Multiply x^T with A.

$\{x, y\} \cdot \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix} =$

$4x+y \quad x+2y$

Multiply this result by x.

$\{4x+y \quad x+2y\} \cdot \{x \quad y\}$

$= (4x+y)\cdot x + (x+2y)\cdot y$

$= 4x^2 + xy + xy + 2y^2$

Find the symmetric matrix A that correspond to the quadratic form $4x_1^2 + 2x_1x_2 + x_2^2$

Vector $x = \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}$

Rewritten as: $x_1^2 \cdot 4 + x_1x_2 \cdot 2 + x_2^2 \cdot 1$

A will be symmetric. $A = \begin{Bmatrix} a & b \\ b & c \end{Bmatrix}$

a: coefficient of $x_1^2 = 4$. c correspond to the coefficient of $x_2^2 = 1$

b correspond to half the coefficient x_1x_2

$b = \frac{2}{2} = 1 \rightarrow$ Matrix A: $\begin{Bmatrix} 4 & 1 \\ 1 & 1 \end{Bmatrix}$

Gradient Descent.

First two steps of GD. $f(x) = x^4 - 6x^3 + 12x^2 - 8x$, initial guess of $x = 2$, $\alpha = 0.01$

$f'(x) = 4x^3 - 18x^2 + 24x - 8$

$x_{new} = x_{old} - \alpha \cdot f'(x_{old})$.

$f'(2) = 0$. $x_1 = 2 - 0.01 \cdot 0 = 2$

$x_2 = x_1 = 0 \rightarrow$ Local minimum, maximum, or saddle point.

$f(x) = 3x^2 + 5y^2$

$\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) = (6x, 10y)$

Initial guess at $x_0 = 1$ and $y_0 = 1$

First iteration: $\nabla f(1,1) = (6, 10)$

Update x and y:

$x_1 = 1 - 0.05 \cdot 6 = 0.7$

$y_1 = 1 - 0.05 \cdot 10 = 0.5$

Second iteration: Compute gradient at $(0.7, 0.5)$

$\nabla f(0.7, 0.5) = (4.2, 5)$

Update x and y.

$x_2 = 0.7 - 0.05 \cdot 4.2 = 0.49$

$y_2 = 0.5 - 0.05 \cdot 5 = 0.25$

Linear Regression by GD.

Loss function: $L(w) = ||y - XB||^2$

Initial guess for the slope B_1 is 1 and the intercept B_0 is 0.

Three data points, $x = (10, 15, 20)$, with $y = (10, 20, 30)$.

$X = \begin{Bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \end{Bmatrix} \quad Y = \begin{Bmatrix} 10 \\ 20 \\ 30 \end{Bmatrix}$

$B = \begin{Bmatrix} B_0 \\ B_1 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$.

Compute the predictions:

$\hat{y} = X \cdot B = \begin{Bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \end{Bmatrix} \cdot \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 10 \\ 15 \\ 20 \end{Bmatrix}$

Compute the loss gradient.

The gradient of the loss with respect to B is $\nabla_B L = -2X^T(y - \hat{y})$

Compute $y - \hat{y}$:

$\begin{Bmatrix} 10 \\ 20 \\ 30 \end{Bmatrix} - \begin{Bmatrix} 10 \\ 15 \\ 20 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 5 \\ 10 \end{Bmatrix}$

Compute $\nabla_B L$:

$-2 \cdot \begin{Bmatrix} 1 & 1 & 1 \\ 10 & 15 & 20 \end{Bmatrix} \cdot \begin{Bmatrix} 0 \\ 5 \\ 10 \end{Bmatrix}$

$= -2 \cdot \begin{Bmatrix} 15 \\ 250 \end{Bmatrix}$

$B_0 = B_0 - \alpha \cdot \nabla_{B_0} L = 0 - 0.05 \cdot (-2 \cdot 15)$

$= 1.5$

$B_1 = B_1 - \alpha \cdot \nabla_{B_1} L = 1 - 0.05 \cdot (-2 \cdot 250)$

$= 26$

Gradient Descent with Momentum.

$f(x) = x^3 - 3x$, initial guess of $x = 1$, a learning rate $\alpha = 0.05$, momentum factor of 0.1.

The gradient of $f(x) = f'(x) = 3x^2 - 3$

$x_0 = f'(1) = 3 \cdot (1)^2 - 3 = 0$

Update the velocity:

$V_1 = \beta V_0 - \alpha f'(1) = 0.1 \cdot 0 - 0.05 \cdot 0 = 0$

$x_1 = x_0 + V_1 = 1 + 0 = 1$

Second Iteration:

x_1: as before, $= 0$

Update the velocity:

$V_2 = \beta V_1 - \alpha f'(1) = 0.1 - 0.05 \cdot 0$

$= 0$

Update the position: $x_2 = x_1 + V_2$

$= 1 + 0 = 1$

Might be a stationary point.

HMM and Viterbi Algorithm.

For simplicity, we assume that there are two categories of mood: good and bad. We also assume that there are three possible facial expressions: smiling, neutral, and frowning.

When we first meet a person, there is a 70% chance that the person is in a good mood and a 30% chance that a person is in a bad mood.

Assume the mood change on a daily basis.

Part 1: Initialization Matrix, transition matrix, and emission matrix.

$\pi = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix} \begin{matrix} \text{Good} \\ \text{Bad} \end{matrix}$

$A = \begin{array}{c|cc} & G & B \\ \hline G & 0.8 & 0.4 \\ B & 0.2 & 0.6 \end{array}$

$B = \begin{array}{c|cc} & G & B \\ \hline S & 0.5 & 0.2 \\ N & 0.4 & 0.35 \\ F & 0.1 & 0.45 \end{array}$

Read the columns name first!

G - good mood, B - Bad mood.

$P(\text{Frowning} | G) \cdot P(G) = 0.1 \times 0.7 = 0.07$

$P(\text{Frowning} | B) \cdot P(B) = 0.45 \cdot 0.3 = 0.135$

$C = \begin{bmatrix} 0.07 & * & * \\ 0.135 & * & * \end{bmatrix}$

$D = \begin{bmatrix} 0 & * & * \\ 0 & * & * \end{bmatrix}$

$P(\text{Smiling} / G) \cdot P(G/G)$

$= 0.5 \times 0.8 = 0.4$

$P(\text{Smiling} / B) \cdot P(B/G)$

$= 0.2 \times 0.2 = 0.04$

$P(\text{Smiling} / B) \cdot P(B/B)$

$= 0.2 \times 0.6 = 0.12$

$P(\text{Smiling} / G) \cdot P(G/B)$

$= 0.5 \times 0.4 = 0.2$

$C_{12} = \max(0.4 \times 0.07,$

$0.2 \times 0.135) = 0.0162$

$C_{22} = \max(0.04 \times 0.07,$

$0.12 \times 0.135) = 0.028$

$P(\text{Neutral} | G) \cdot P(G|G)$

$= 0.4 \times 0.8 = 0.32$

$P(\text{Neutral} | B) \cdot P(B|G)$

$= 0.35 \times 0.2 = 0.07$

$P(\text{Neutral} | B) \cdot P(B|B)$

$= 0.35 \times 0.6 = 0.21$

$P(\text{Neutral} | G) \cdot P(G|B)$

$= 0.4 \times 0.4 = 0.16$

$C_{13} = \max(0.32 \times 0.028,$

$0.16 \times 0.0162)$

$= 0.00896$

$C_{23} = \max(0.07 \times 0.028,$

$0.21 \times 0.0162)$

$= 0.003402$

$C = \begin{bmatrix} 0.07 & 0.028 & 0.00896 \\ 0.135 & 0.0162 & 0.003402 \end{bmatrix}$

$D = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}$

Backward Pass

Since $0.00896 > 0.003402$,

$S = 1$

$D = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}$

Answer: (Good mood, Good mood, Good mood).

Another Example for Viterbi

1. Mucking 2. Bumbling.

$\pi = \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} \quad A = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}$

$B = \begin{bmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{bmatrix}$

Initialization and Forward Pass.

$P(M|E) \cdot P(E) = 0.8 \times 0.3 = 0.24$

$P(M|R) \cdot P(R) = 0.1 \times 0.7 = 0.07$

$C = \begin{bmatrix} 0.24 & * \\ 0.07 & * \end{bmatrix}$

$D = \begin{bmatrix} 0 & * \\ 0 & * \end{bmatrix}$.

Forward Pass:

$P(B|E) \times P(E|E) = 0.2 \cdot 0.6$

$= 0.12$

$P(B|R) \times P(R|E) = 0.9 \times 0.4$

$= 0.36$

$P(B|R) \times P(R|R) = 0.9 \times 0.8$

$= 0.72$

$P(B|E) \times P(E|R) = 0.2 \times 0.2$

$C_{12} = \max(0.12 \times 0.24,$

$0.04 \times 0.07) = 0.0288$

$C_{22} = \max(0.36 \times 0.24,$

$0.72 \times 0.07) = 0.0864$.

$C = \begin{bmatrix} 0.24 & 0.0288 \\ 0.07 & 0.0864 \end{bmatrix}$

$D = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

Part 3: Find which primary states correspond to the first two readings received by the scientists by performing the backward pass.

MCMC with Graph

The acceptance probability, $A(x, x')$, for a transition from a current state x to a candidate state x' is given by:

$$A(x, x') = \min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right)$$

$p(x)$ is the target distribution

$q(x'|x)$ is the proposal distribution for moving from state x to state x'.

$q(x'|x)$ is a normal distribution with a standard deviation of 1.

Sometimes, the proposal distribution is symmetric, (i.e. $q(x'|x) = q(x|x')$).

$$A(3, 0) = \min\left(1, \frac{p(0)}{p(3)}\right) = \min\left(1, \frac{0.05}{0.3}\right)$$

MLE #2.

Suppose x is a discrete random variable with the probability mass function:

x	0	1	2	3
$p(x;\theta)$	$\frac{2}{3}\theta$	$\frac{1}{3}\theta$	$\frac{2}{3}(1-\theta)$	$\frac{1}{3}(1-\theta)$

Ten independent observations:

$$x = \{3, 0, 2, 1, 3, 2, 1, 0, 2, 1\}$$

Likelihood function:

$$p(x;\theta) = \prod_{i=1}^{10} p(x^{(i)};\theta) = \left(\frac{2}{3}\theta\right)^2\left(\frac{1}{3}\theta\right)^3$$

$$\left(\frac{2}{3}(1-\theta)\right)^3\left(\frac{1}{3}(1-\theta)\right)^2$$

Log-Likelihood function:

$$\log p(x;\theta) = \sum_{i=1}^{10} \log p(x^{(i)};\theta) = 5\log\theta +$$

$$5\log(1-\theta) + 5\log\left(\frac{2}{9}\right)$$

The derivative of the log-likelihood function:

$$\frac{\partial}{\partial\theta}\log p(x;\theta) = \frac{5}{\theta} - \frac{5}{1-\theta}$$

Setting the derivative to zero:

$$5(1-\theta) - 5\theta = 0$$

$$1 - 2\theta = 0$$

MLE for θ is potentially:

$$\hat{\theta} = \frac{1}{2}$$

Central Limit Theorem

The mean and standard deviation

$\mu_{sum} = n \times \mu \quad \sigma^2_{sum} = n \times \sigma^2$

the variance of the sum is the sum of the variance. $\sigma_{sum} = \sqrt{n} \cdot \sigma$

Bayes Table

Three hypotheses:

1. H_1: The coin is fair ($P(heads) = 0.5$)
2. H_2: The coin is biased towards heads ($P(heads) = 0.6$)
3. H_3: The coin is biased away from heads ($P(heads) = 0.4$)

Our priors ($P(H)$):

- $P(H_1) = 0.5$ (fair), $P(H_2) = 0.25$ (Biased towards heads), $P(H_3) = 0.25$ (Biased away from heads).

The likelihood ($P(D|H)$) of observing data (two heads in a row):

- $P(D|H_1) = 0.5^2 = 0.25$
- $P(D|H_2) = 0.6^2 = 0.36$
- $P(D|H_3) = 0.4^2 = 0.16$

Calculate the numerator of Bayes Theorem, $P(H) \times P(D|H)$. Sum these to get the total probability $P(D)$.

Posterior probability: $P(H|D) = \dfrac{P(D|H) \cdot P(H)}{P(D)}$

Beta distribution

Have a prior distribution that is a Beta distribution and you are looking to update this with evidence in the form of binary outcomes (like correct or incorrect answers), the posterior distribution is also a Beta distribution.

Beta distribution is a conjugate prior for the binomial distribution.

Prior distribution: $\alpha = 9$, $\beta = 1$

- # of successes: 6, - # of failures: Total - # of S.

Posterior distribution: α_{post}: α + # of successes = $9 + 6$.

β_{post}: β + # of failures = $1 + 4$

MAP and Minimize MSE.

To minimize the MSE, we can use the expected value (mean) of the posterior distribution.

$$E[X] = \frac{\alpha}{\alpha+\beta}$$

Stochastic Gradient Descent for Simple Linear Regression.

Initialization: Slope (β_1) = 4. Intercept (β_0) = 0, Learning Rate $\alpha = 0.01$

Data point: (2, 5)

5 is the actual observed value.

The predicted y-value: $\hat{y} = \beta_1 x + \beta_0$
$= (-4) \cdot 2 + 0 = -8$

Compute the gradient of the loss function with respect to each parameter.

The loss function for a single data point is the squared error: $(y - \hat{y})^2$.

The gradient for β_1: $-2x(y - \hat{y})$

The gradient for β_0: $-2(y - \hat{y})$

Calculate gradient for the data point:

β_1: $-2 \cdot 2 \cdot (5 - (-8)) = -52$

β_0: $-2 \cdot (5 - (-8)) = -26$

Update:

β_1 = old β_1 − $\alpha \cdot$ Grad β_1
$= -4 - 0.1 \times (-52)$
$= 1.2$

β_0 = old β_0 − $\alpha \cdot$ Grad β_0
$= 0 - 0.1 \times (-26)$
$= 2.6$

$$y = 1.2x + 2.6$$

Another Method for SGD

$$\beta^{n+1} = \beta^n - \alpha(x_i^T \cdot \beta^n - y_i) \cdot x_i$$

Initial parameters: $\beta^n = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -4 \end{bmatrix}$

Feature vector (x_i): $x_i = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$\begin{bmatrix} 1 \\ x_{value} \end{bmatrix}$ Actual x-value.

Calculate the predicted value

$x_i^T \cdot \beta^n = \begin{bmatrix} 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ -4 \end{bmatrix} = -8$

Calculate the error ($x_i^T \cdot \beta^n - y_i$)

$x_i^T \cdot \beta^n - y_i = -8 - 5 = -13$

y_i is the actual y-value.

$(-\alpha \cdot (x_i^T \cdot \beta^n - y_i) \cdot x_i)$:

$-0.1 \times -13 \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1.3 \\ 2.6 \end{bmatrix}$

Update the parameter vector (β^{n+1}):

$$\beta^{n+1} = \begin{bmatrix} 0 \\ 4 \end{bmatrix} - \begin{bmatrix} 1.3 \\ 2.6 \end{bmatrix} = \begin{bmatrix} -1.3 \\ -6.6 \end{bmatrix}$$

Continue Here!

From the last column of C follows that S = 2, because 2nd row > 1st row. This leads to (R). Resting, being the second hidden state.

Since $D_{22} = 1$, the first hidden state is E, Energy.

Answer: (Energy, Resting).

Initialization: $P(A|1) \times P(1)$

$P(A|2) \times P(2)$

Forward Pass: $P(B|1) \times P(1|1)$

$P(B|2) \times P(2|1)$

$P(B|2) \times P(2|2)$

$P(B|1) \times P(1|2)$

C_{12} : $1 \times C_1$, $4 \times C_2$

C_{22} : $2 \times C_1$, $3 \times C_2$

Bias, Variance, MSE.

Low Variance: Precise

Low Bias: Accurate.

$$\text{bias}(\hat{\theta}_m) = E[\hat{\theta}_m] - \theta$$

Unbiased: $\text{bias}(\hat{\theta}_m) = 0$

OR In other words: $E[\hat{\theta}] = \theta$

MSE: Measures the "average distance squared" between the estimator and the true value.

$$MSE(\hat{\theta}_m) = E[(\hat{\theta}_m - \theta)^2]$$
$$= \text{bias}(\hat{\theta}_m)^2 + Var(\hat{\theta}_m)$$