COMP350 Assignment 2:

Q1)

1. Program:

Text

Description automatically generated

Issue: X(0.1, 20, 2000) = infinity instead of 0.3

This is because, in the first for loop, at the beginning of the 11th loop (when j = 11), X = 1.992351e+36. When you multiply X by (N+1-j), meaning X = X\*1989 = X \*1.989e3, the resulting X will be to the power of 39 (meaning magnitude of e+39). However, this value is larger than the largest normal FPN that can be stored in single precision (is of magnitude e+38, as seen in lecture), so the number is not a FPN anymore, so it will need to be rounded. This is called overflow. It is rounded to infinity according to the “to nearest” rounding rule. As a result, all following computations will have X equal to infinity. Therefore, the result of X(0.1, 20, 2000) = infinity.

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generated

* This code gives the result X(0.1, 200, 2000) = 2.972153e-02 = 0.0297, which is close to 0.03.

Q2)

Text

Description automatically generated

Results:

p\_3 = 3.061467458920719e+00, diff = 8.012519466907442e-02

p\_4 = 3.121445152258053e+00, diff = 2.014750133174026e-02

p\_5 = 3.136548490545941e+00, diff = 5.044163043852468e-03

p\_6 = 3.140331156954739e+00, diff = 1.261496635053927e-03

p\_7 = 3.141277250932757e+00, diff = 3.154026570362234e-04

p\_8 = 3.141513801144145e+00, diff = 7.885244564764804e-05

p\_9 = 3.141572940367883e+00, diff = 1.971322191041125e-05

p\_10 = 3.141587725279961e+00, diff = 4.928309832230582e-06

p\_11 = 3.141591421504635e+00, diff = 1.232085157898410e-06

p\_12 = 3.141592345611077e+00, diff = 3.079787163073888e-07

p\_13 = 3.141592576545004e+00, diff = 7.704478877101906e-08

p\_14 = 3.141592633463248e+00, diff = 2.012654487515420e-08

p\_15 = 3.141592654807589e+00, diff = -1.217796086194767e-09

p\_16 = 3.141592645321215e+00, diff = 8.268577822434509e-09

p\_17 = 3.141592607375720e+00, diff = 4.621407345695161e-08

p\_18 = 3.141592910939673e+00, diff = -2.573498796287765e-07

p\_19 = 3.141594125195191e+00, diff = -1.471605397984632e-06

p\_20 = 3.141596553704820e+00, diff = -3.900115026489459e-06

p\_21 = 3.141596553704820e+00, diff = -3.900115026489459e-06

p\_22 = 3.141674265021758e+00, diff = -8.161143196439014e-05

p\_23 = 3.141829681889202e+00, diff = -2.370282994084150e-04

p\_24 = 3.142451272494134e+00, diff = -8.586189043406911e-04

p\_25 = 3.142451272494134e+00, diff = -8.586189043406911e-04

p\_26 = 3.162277660168380e+00, diff = -2.068500657858641e-02

p\_27 = 3.162277660168380e+00, diff = -2.068500657858641e-02

p\_28 = 3.464101615137754e+00, diff = -3.225089615479613e-01

p\_29 = 4.000000000000000e+00, diff = -8.584073464102069e-01

p\_30 = 0.000000000000000e+00, diff = 3.141592653589793e+00

p\_31 = 0.000000000000000e+00, diff = 3.141592653589793e+00

p\_32 = 0.000000000000000e+00, diff = 3.141592653589793e+00

p\_33 = 0.000000000000000e+00, diff = 3.141592653589793e+00

p\_34 = 0.000000000000000e+00, diff = 3.141592653589793e+00

p\_35 = 0.000000000000000e+00, diff = 3.141592653589793e+00

The calculated value gets closer and closer to pi until p\_15, where it becomes smaller than pi. Afterwards, the difference becomes larger and larger, until p\_30, where the difference becomes pi because p\_30 = 0.

The difference gets smaller after p\_14 because (p/2n)2 can no longer be represented as a FPN for p\_15. This is because, for p\_14, (p/2n)2 = 1.470685588866872e-07, which is of the power 2^-23. This number subtracted from 1 (ie. 1 - \_\_\_\*2^-23) is still a FPN and will not be rounded. However, for p\_15, (p/2n)2 = 3.676714105393941e-08, which is of the power 2^-25, which is no longer a FPN and will be rounded. For p\_16, similarly, (p/2n)2  < 2-23, so the value is once again rounded. These rounding errors lead to loss of accuracy.

For p\_30, (p/2n)2 becomes so small that, after rounding, 1 - (p/2n)2 = 1. Then, when we subtract that from 1, we get that (1 - sqrt(1 – (p/2n)2)) = 0 due to numerical cancellation. As a result, p = 2n \* sqrt(2 \* (1 - sqrt(1 – (p/2n)2))) = 2^n \* sqrt( 2 \* 1 – 1) = 2^n \* sqrt(0) = 0. Therefore, all later p’s also equal 0.

Text

Description automatically generatedTo avoid getting p\_n = 0 for large n, we can multiply the formula by its conjugate. This means:

Since we are no longer doing 1 - we will avoid numerical cancelation. As a result, we see that the difference decreases as n increases, with p\_35 equalling pi.

p\_3 = 3.061467458920718e+00, diff = 8.012519466907486e-02

p\_4 = 3.121445152258052e+00, diff = 2.014750133174070e-02

p\_5 = 3.136548490545939e+00, diff = 5.044163043853800e-03

p\_6 = 3.140331156954753e+00, diff = 1.261496635039716e-03

p\_7 = 3.141277250932773e+00, diff = 3.154026570197921e-04

p\_8 = 3.141513801144301e+00, diff = 7.885244549177273e-05

p\_9 = 3.141572940367092e+00, diff = 1.971322270133413e-05

p\_10 = 3.141587725277160e+00, diff = 4.928312632657139e-06

p\_11 = 3.141591421511201e+00, diff = 1.232078592483532e-06

p\_12 = 3.141592345570118e+00, diff = 3.080196746552133e-07

p\_13 = 3.141592576584873e+00, diff = 7.700491977402635e-08

p\_14 = 3.141592634338564e+00, diff = 1.925122949941738e-08

p\_15 = 3.141592648776986e+00, diff = 4.812807041787437e-09

p\_16 = 3.141592652386592e+00, diff = 1.203201094313044e-09

p\_17 = 3.141592653288994e+00, diff = 3.007993853998414e-10

p\_18 = 3.141592653514594e+00, diff = 7.519940226075050e-11

p\_19 = 3.141592653570994e+00, diff = 1.879918443137285e-11

p\_20 = 3.141592653585094e+00, diff = 4.698907929423513e-12

p\_21 = 3.141592653588619e+00, diff = 1.173727781633715e-12

p\_22 = 3.141592653589500e+00, diff = 2.926547892911913e-13

p\_23 = 3.141592653589721e+00, diff = 7.238654120556021e-14

p\_24 = 3.141592653589776e+00, diff = 1.731947918415244e-14

p\_25 = 3.141592653589790e+00, diff = 3.552713678800501e-15

p\_26 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_27 = 3.141592653589794e+00, diff = -8.881784197001252e-16

p\_28 = 3.141592653589794e+00, diff = -8.881784197001252e-16

p\_29 = 3.141592653589794e+00, diff = -4.440892098500626e-16

p\_30 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_31 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_32 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_33 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_34 = 3.141592653589793e+00, diff = 0.000000000000000e+00

p\_35 = 3.141592653589793e+00, diff = 0.000000000000000e+00