AMSC 660 Assignment 2

Jiaqi Leng

September 12, 2019

1 Problem 1

Define $S(k) = \sum_{i=0}^{k-1} x_i$ for $1 \leq k \leq n$. The for-loop computes $S(k)$ by iteration:

$$\hat{S}(k) = \hat{S}(k-1) + x_{k-1}, \text{ with initial data } \hat{S}(0) = 0.$$

We will prove $|\hat{S}(n) - S(n)| \leq (n-1)\epsilon_M \sum_{i=0}^{n-1} |x_i|$ by induction on n, i.e., we want to show $|\hat{S}(k) - S(k)| \leq (k-1)\epsilon_M \sum_{i=0}^{k-1} |x_i|$ is true for all $1 \leq k \leq n$.

When $k = 1$, $\hat{S}(1) = \mathtt{Fl}(0 + x_0) = x_0$ (adding 0 is exact), so

$$|\hat{S}(1) - S(1)| = |x_0 - x_0| = 0.$$

Now, if the statement is true for $n = k$,

$$|\hat{S}(k+1) - S(k+1)| = |\mathtt{Fl}(\hat{S}(k) + x_k) - S(k+1)| = |(1+\epsilon)(\hat{S}(k) + x_k) - S(k+1)|$$

$$= |\hat{S}(k) + x_k + \epsilon(\hat{S}(k) + x_k) - (S(k) + x_k)| \leq |\hat{S}(k) - S(k)| + |\epsilon||\hat{S}(k) + x_k|. \tag{1}$$

We claim that $|\epsilon||\hat{S}(k) + x_k| \leq \epsilon_M \sum_{i=0}^{k} |x_i|$ for all $1 \leq k \leq n-1$. This can be easily verified inductively: ($k = 1$ is a trivial case)

$$|\epsilon||\hat{S}(k+1) + x_{k+1}| = |\epsilon||\mathtt{Fl}(\hat{S}(k) + x_k) + x_{k+1}| = |\epsilon||(1+\epsilon_1)(\hat{S}(k) + x_k) + x_{k+1}|$$

$$\leq |\epsilon||\hat{S}(k) + x_k + x_{k+1}| + |\epsilon\epsilon_1||\hat{S}(k) + x_k| \leq \epsilon_M \sum_{i=0}^{k+1} |x_i|,$$

in the last inequality we omit the second-order error term.

By our inductive assumption and the claim, Eq. (1) yields

$$|\hat{S}(k+1) - S(k+1)| \leq (k-1)\epsilon_M \sum_{i=0}^{k-1} |x_i| + \epsilon_M \sum_{i=0}^{k} |x_i| \leq k\epsilon_M \sum_{i=0}^{k} |x_i|,$$

which concludes the induction.

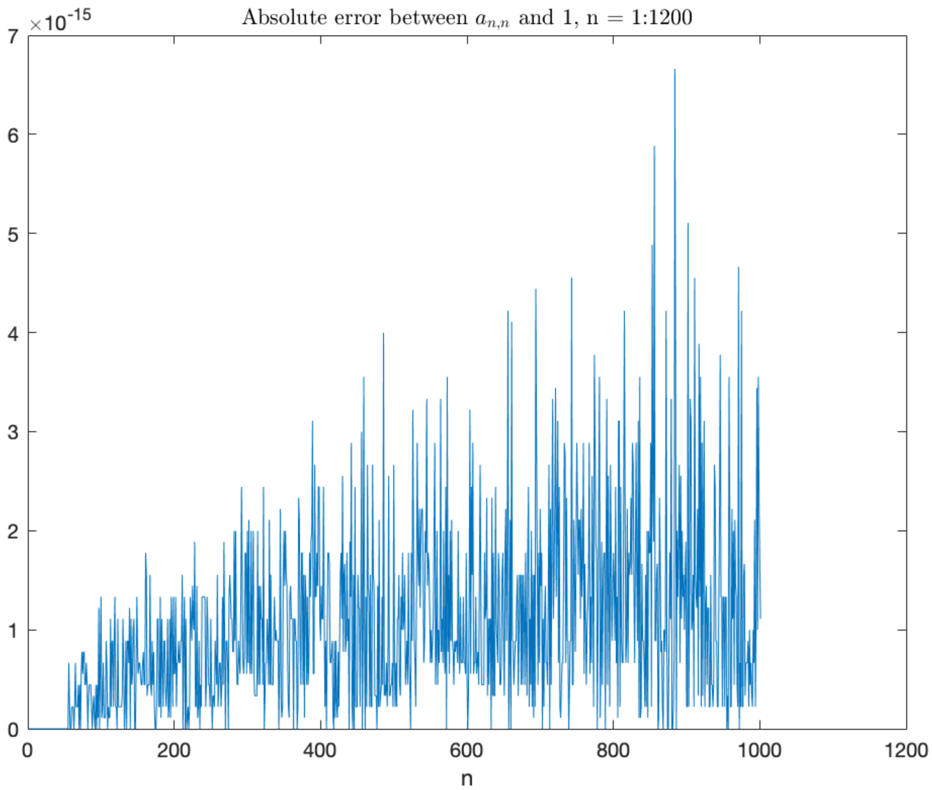

Figure 1: Absolute error between $a_{n,n}$ and 1, for $1 \leq n \leq 1200$

2 Problem 2

(a) We use the iterative relation $a_{n,k+1} = \frac{n-k}{k+1} a_{n,k}$ to calculate $a_{n,k}$ for $1 \leq n \leq 1200$. And we plot the absolute error between $\hat{a}_{n,n}$ and 1, see Figure 1. It turns out an overflow occurs when $n \geq 1021$, as all $a_{n,n} = \texttt{Inf}$ after 1021.

As we can see from the plot 1, if there is no overflow in the computation, the errors between $\hat{a}_{n,n}$ and 1 are extremely small (less than 10^{-14}). Given this numerical fact, we are certain that the roundoff is not a problem in this iterative scheme.

(b)

For E_n and M_n, see Figure 2 and 3.

The absolute errors between $E(n)$ and $n/2$ is shown in Figure 4, and we can see they are extremely small (less than 2×10^{-12}) in magnitude.

The reason why the computed results are of high accuracy even the intermediate values are in a large range (from 1 to 10^n for very large n) is that terms in $a_{n,k}$ that are close to $M(n)$ dominate the summation $\sum_k ka_{n,k}$. As we can see from Figure 3, $M(n)$ is almost the same as 2^n in log scale. These dominating terms yield significant result when divided by 2^n, while other small terms are almost negligible. In other words, small combinatoric numbers $a_{n,k}$ may be rounded off when adding to large $ka_{n,k}$, but they actually do not matter compared to 2^n. As a result, the cancellation caused by small $a_{n,k}$ is not significant compared to M_n, so the final result is still very accurate.

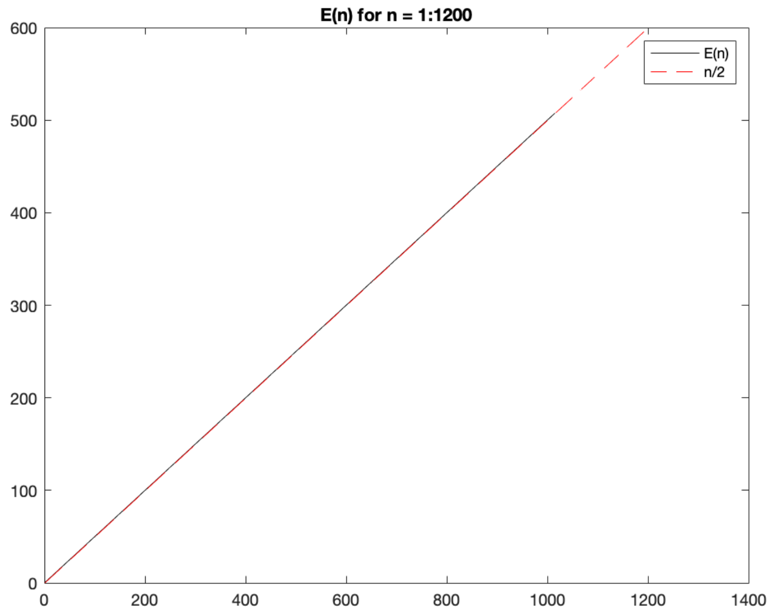

Figure 2: $E(n)$, $(1 \leq n \leq 1200)$, compared to $n/2$

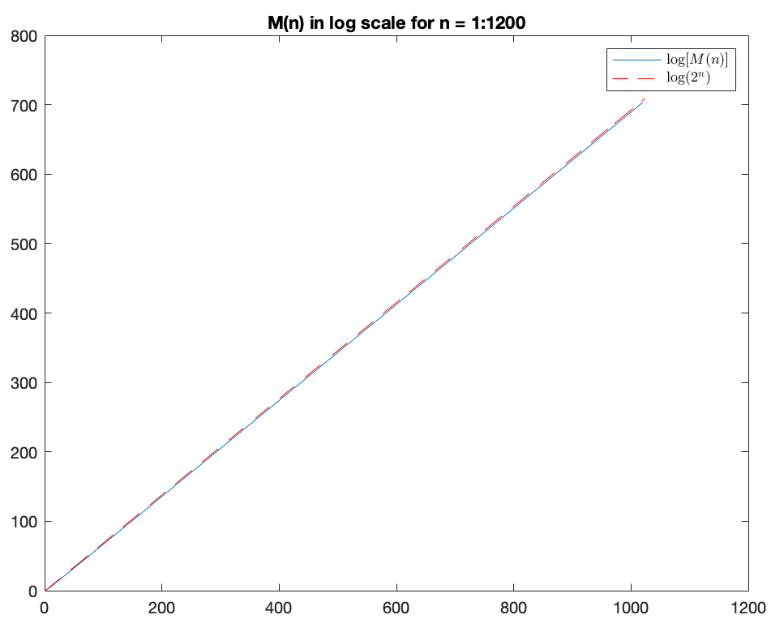

Figure 3: $M(n)$ $(1 \leq n \leq 1200)$, compared to 2^n

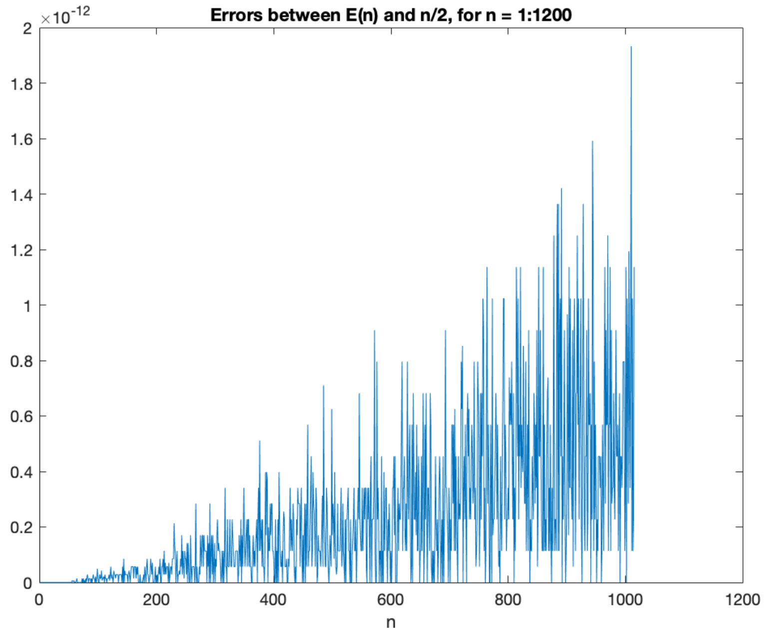

Figure 4: Errors between $E(n)$ and $n/2$ ($1 \leq n \leq 1200$)

For large n (in my computed, when $n \geq 1021$), an overflow occurs in M_n, so the outcome of M_n all becomes `Inf`. Then for n ranges from 1021 to 1023, 2^n is sill meaningful, so the output for E_n is also `Inf` for $1021 \leq n \leq 1023$. When $n \geq 1023$, 2^n overflows, hence $E(n) = \frac{\texttt{Inf}}{\texttt{Inf}}$ =NaN. This analysis coincides with the computed result for $E(n)$ and $M(n)$.

3 Problem 3

(a) When $n = 0$,

$$E_0 = \int_0^1 e^{x-1}\mathrm{d}x = e^{x-1}\Big|_0^1 = 1 - e^{-1}.$$

(b) Integral by parts: (for $n \in \mathbb{N}^*$)

$$E_n = \int_0^1 x^n e^{x-1}\mathrm{d}x = x^n e^{x-1}\Big|_0^1 - \int_0^1 nx^{n-1}e^{x-1}\mathrm{d}x = 1 - nE_{n-1}.$$

(c) For $x \in [0, 1]$, $e^{-1} \leq e^{x-1} \leq 1$. Hence,

$$\int_0^1 x^n e^{-1}\mathrm{d}x \leq \int_0^1 x^n e^{x-1}\mathrm{d}x \leq \int_0^1 x^n\mathrm{d}x,$$

$$\frac{1}{e(n+1)} \leq E_n \leq \frac{1}{n+1}.$$

(d) The error in the first term E_0 will be amplified through the iteration. If $\hat{E}_0 = (1+\epsilon)E_0$, we claim the absolute error for $\hat{E}_n$ will be $n!\epsilon E_0$.

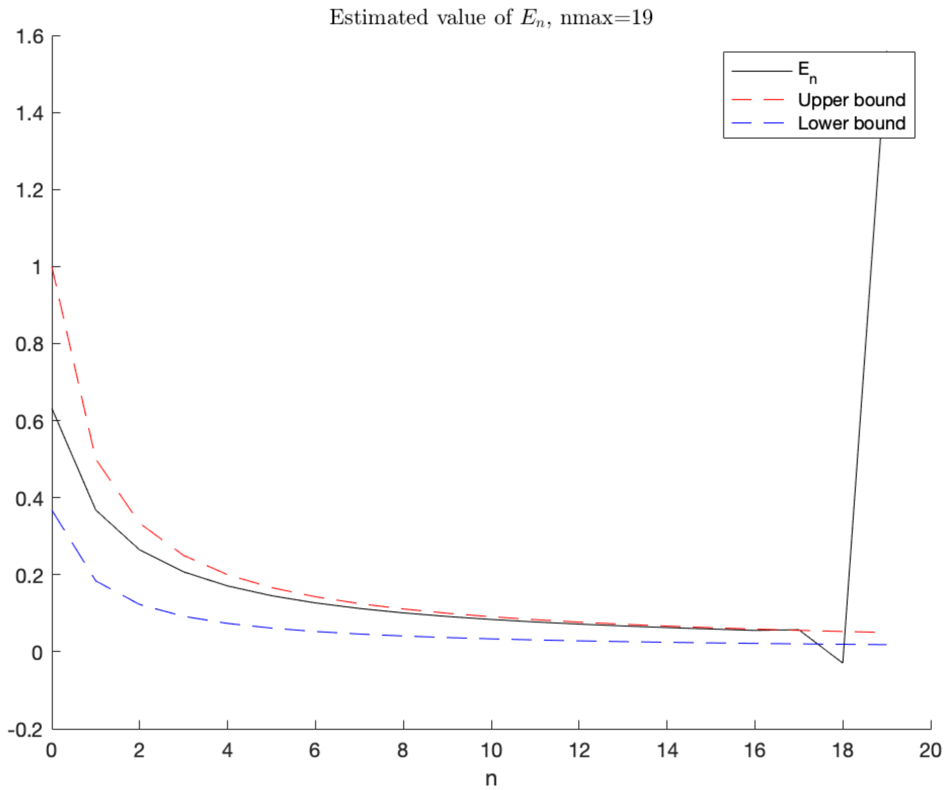

Figure 5: Approximated value of E_n, for $0 \leq n \leq 19$

This can be proven by induction. When $n = 0$, $\hat{E}_0 - E_0 = \epsilon E_0$. Suppose $\hat{E}_k - E_k = k!\epsilon E_0$, we have

$$\hat{E}_{k+1} - E_{k+1} = [1 - (k+1)\hat{E}_k] - [1 - (k+1)E_k] = (k+1)(E_k - \hat{E}_k) = (k+1)!\epsilon E_0.$$

It turns out that after n iterations, the absolute error between $\hat{E}_n$ and E_n will be $n!\epsilon E_0$. Moreover, as $E_n \in [1/e(n+1), 1/(n+1)]$, we can even compute the relative error for $\hat{E}_n$:

$$|\epsilon|(n+1)!E_0 \leq \left|\frac{\hat{E}_n - E_n}{E_n}\right| \leq e|\epsilon|(n+1)!E_0.$$

It is now clear that both the absolute error and relative error in E_n will amplify when n gets large.

(d) The figure is as in Figure 5. We choose nmax = 19 to evidence something wrong.