Question 1

Part a  Error bound when summing from left to right

Define $a_k = \frac{1}{k^2}, s_n = \sum_{k=1}^n a_k, s_n^*$ be the result for $s_n$ in floating point arithmetic (summing from left to right in $s_n$), and $s_n^* - s_n = e_n \varepsilon$, where $\varepsilon$ is the machine precision.

Under these notations, we have $e_1 = 0$, and that

$$s_{n+1}^* = fl(s_n^* + fl(a_{n+1})) \quad (*)$$

$$= (s_n^* + a_{n+1}(1 + \varepsilon_1))(1 + \varepsilon_2), \text{ where } |\varepsilon_1| \leq \varepsilon, |\varepsilon_2| \leq \varepsilon$$

$$= s_n^* + a_{n+1} + a_{n+1}\varepsilon_1 + s_n^*\varepsilon_2 + a_{n+1}\varepsilon_2 + a_{n+1}\varepsilon_1\varepsilon_2$$

$$= s_{n+1} + e_n \varepsilon + a_{n+1}\varepsilon_1 + s_n \varepsilon_2 + e_n \varepsilon\varepsilon_2 + a_{n+1}\varepsilon_2 + a_{n+1}\varepsilon_1\varepsilon_2.$$  

Disregarding terms of order higher than 1 in $\varepsilon$'s gives

$$s^*_{n+1} = s_{n+1} + e_n \varepsilon + a_{n+1}\varepsilon_1 + s_n \varepsilon_2 + a_{n+1}\varepsilon_2,$$

which indicates

$$|s^*_{n+1} - s_{n+1}| \leq |e_n \varepsilon + a_{n+1}\varepsilon_1 + s_n \varepsilon_2 + a_{n+1}\varepsilon_2|$$

$$\leq (|e_n| + a_{n+1} + s_{n+1})\varepsilon;$$

that is,

$$|e_{n+1}| \leq |e_n| + a_{n+1} + s_{n+1}.$$

Applying this inequality repeatedly and the estimate that $\sum_{k=1}^{n+1} a_k \leq \sum_{k=1}^{\infty} a_k = \frac{\pi^2}{6} < 2$ to get

$$|e_{n+1}| \leq |e_1| + \sum_{k=2}^{n+1} a_k + \sum_{k=2}^{n+1} s_k$$

$$\leq 2 + 2n$$

$$= 2(n + 1).$$

Therefore we have

$$s^*_n = s_n + O(n)\varepsilon.$$  

□
Part b  Error bound when summing from right to left

For 1 \leq k \leq n, define \( S_k = \sum_{i=k}^{n} \frac{1}{i^2} \), \( S^*_k \) be the result for \( S_k \) in floating point arithmetic (summing from left to right in \( S_k \)), and \( S_k^* - S_k = e_k \varepsilon \), where \( \varepsilon \) is the machine precision.

Additionally, let \( b_k = \frac{1}{(n+1-k)^2} \) for 1 \leq k \leq n.

Under these notations, we have \( S_1 = \frac{1}{n^2}, |e_1| \leq b_1 \), and
\[
S_{k+1}^* = fl(S_k^* + fl(b_{k+1})) \quad (**) 
\]

Notice that (**) is similar to (**), so by the same argument in part a, we get
\[
|e_n| \leq |e_1| + \sum_{k=1}^{n-1} b_{k+1} + \sum_{k=1}^{n-1} S_{k+1} = (b_1 + \sum_{k=1}^{n-1} b_{k+1}) + \sum_{k=1}^{n-1} S_{k+1} = S_n + (n-1)b_1 + \sum_{k=2}^{n} (n-k+1)b_k
\]
\[
\leq S_n + ((n-1+1)b_1 + \sum_{k=2}^{n-1} (n-k+1)b_k) + b_n
\]
\[
\leq S_n + \sum_{k=1}^{n-1} (n-k+1)b_k + b_n
\]
\[
\leq 3 + \sum_{k=1}^{n-1} \frac{1}{n-k+1} \quad \text{(since } S_n \leq 2, b_n = 1) 
\]
\[
= 3 + \sum_{k=2}^{n} \frac{1}{k}
\]
\[
\leq 3 + \sum_{k=2}^{n} \int_{k-1}^{k} \frac{1}{x} \, dx 
\]
\[
= 3 + \int_{1}^{n} \frac{1}{x} \, dx 
\]
\[
= 3 + \ln n 
\]
\[
\leq 2 \ln n, \text{ if } n \geq 10.
\]

Hence we have
\[
S_n^* = S_n + O(\ln n) \varepsilon.
\]
Question 2
We first prove the following lemma.

Lemma If real numbers $a$, and $b$ satisfy $a < b$, then we have $fl(a) \leq fl(b)$.

Proof
According to the construction of IEEE floating point numbers, we have the following facts.

Fact 1 If $x$ is already a floating point number, then $x = fl(x)$, otherwise there exists floating point number $y$ and its next floating point number $z$ with $y < x < z$ such that $fl(x) = y$ whenever $y < x < \frac{y + z}{2}$, and $fl(x) = z$ whenever $\frac{y + z}{2} \leq x < z$.

Fact 2 There is no floating point number between $x$ and $fl(x)$ for any real number $x$.

We next discuss two cases.

Case 1: There is a floating point number $c$ such that $a \leq c \leq b$.

Since $a \leq c$, by Fact 2, we have $fl(a) \leq c$.

Also, since $c \leq b$, by Fact 2, we have $c \leq fl(b)$.

Combining these inequalities gives $fl(a) \leq fl(b)$.

Case 2: There is no floating point number $\in [a,b]$. Then none of $a$ and $b$ is floating point number.

Let $d$ be the largest floating point number with $d < a$, and $e$ be the smallest floating point number with $b < e$.

Then We have the following results according to Fact 1.

If $a \geq \frac{d + e}{2}$, then $fl(a) = fl(b) = e$.

If $b < \frac{d + e}{2}$, then $fl(a) = fl(b) = d$.

If $a < \frac{d + e}{2} \leq b$, then $fl(a) = d < e = fl(b)$.

Hence no matter which case happens, we have $fl(a) \leq fl(b)$ whenever $a < b$.

Since $a < b$, we have $a < \frac{a + b}{2}$, by Lemma, we get $fl(a) \leq fl(\frac{a + b}{2})$. Since $a$ is a floating point number, $a = fl(a)$.

These two relations tells us $a \leq fl(\frac{a + b}{2})$. Similar argument gives $fl(\frac{a + b}{2}) \leq b$.

Therefore we have

$$a \leq fl(\frac{a + b}{2}) \leq b.$$
Question 3

One can list the first few terms of the sequence of intervals to write down the pattern

\[ [a_{2n}, b_{2n}] = [-2^{-2n}, 2^{-2n+1}], \]

and

\[ [a_{2n+1}, b_{2n+1}] = [-2^{-2n}, 2^{-2n-1}]. \]

Next we prove the pattern above by induction on \( n \).

Step 1 (base case)
\([a_0, b_0] = [-1, 2] \) and \([a_1, b_1] = [-1, 2^{-1}] \) satisfy the pattern.

Step 2 (inductive step)
Assuming the pattern works for \( n = k \), that is,

\[ [a_{2k}, b_{2k}] = [-2^{-2k}, 2^{-2k+1}], \]

and

\[ [a_{2k+1}, b_{2k+1}] = [-2^{-2k}, 2^{-2k-1}], \]

we have

\[ [a_{2k+2}, b_{2k+2}] = [2^{-1}(-2^{-2k} + 2^{-2k-1}), 2^{-2k-1}] = [-2^{-2k-2}, 2^{-2k-1}], \]

and

\[ [a_{2k+3}, b_{2k+3}] = [-2^{-2k-2}, 2^{-1}(-2^{-2k-2} + 2^{-2k-1})] = [-2^{-2k-2}, 2^{-2k-3}], \]

which satisfy the pattern for \( n = k + 1 \).

By the pattern above, we have

\[ [a_{1074}, b_{1074}] = [-2^{-1074}, 2^{-1073}]. \]

When computing the midpoint of such interval, we have

\[ p_{1074} = \frac{-2^{-1074} + 2^{-1073}}{2} = \frac{2^{-1074}}{2} \in (0, 2^{-1074}) \]

will give 0 exactly, since the smallest subnormal number is \((-1)21-1023(0 + 2^{-52}) = 2^{-1074}\), and any positive number smaller than that will result in underflow to 0.

Hence 1075 steps are needed to get maximum accuracy.
Question 4

Part a

Taking limits on both sides of the given equation

\[ x_{n+1} = -\frac{x_n^2 - c}{2b} \]

gives

\[ x = -\frac{x^2 + c}{2b} , \]

that is,

\[ x^2 + 2bx + c = 0 . \]

Part b

Define

\[ g(x) = -\frac{x^2 + c}{2b} . \]

Then

\[ |g'(x)| = \frac{|x|}{|b|} \leq \frac{1}{2} \]

whenever \( |x| \leq \frac{|b|}{2} \).

Next, we are looking for conditions on \( b \) and \( c \) such that \( [-\frac{|b|}{2}, \frac{|b|}{2}] \) is an invariant interval under the function \( g \).

Case 1: \( b > 0 \)

Then function \( g(x) = -\frac{x^2 + c}{2b} \) takes maximum \( -\frac{c}{2b} \) at \( x = 0 \), and takes minimum \( -\frac{b^2 - c}{2b} \) at \( \pm \frac{b}{2} \).

In order to have the invariant property, we need

\[ -\frac{c}{2b} \leq \frac{b}{2} , \]

and

\[ -\frac{b^2 - c}{2b} \geq -\frac{b}{2} . \]

Straightforward computations give

\[ -b^2 \leq c \leq \frac{3}{4} b^2 . \]

Case 2: \( b < 0 \)

Then function \( g(x) = -\frac{x^2 + c}{2b} \) takes maximum \( -\frac{b^2 - c}{2b} \) at \( \pm \frac{b}{2} \), and takes minimum \( -\frac{c}{2b} \) at \( x = 0 \).
In order to have the invariant property, we need

\[-\frac{b^2}{4} - c \leq \frac{-b}{2},\]

and

\[-\frac{c}{2b} \geq \frac{b}{2}.

Straightforward computations give

\[-b^2 \leq c \leq \frac{3}{4}b^2.

Hence the conditions \(-b^2 \leq c \leq \frac{3}{4}b^2\), and \(b \neq 0\) guarantee the following

1. \([-\frac{|b|}{2}, \frac{|b|}{2}]\) is an invariant interval under the function \(g\).

2. \(|g'(x)| \leq \frac{1}{2}\) whenever \(x \in [-\frac{|b|}{2}, \frac{|b|}{2}]\).

By Theorem 2.3 (page 57 in Burden and Faires), we have exactly one fixed point \(x \in [-\frac{|b|}{2}, \frac{|b|}{2}]\). (Two roots to \(g(x) = x\) can not be \(\in [-\frac{|b|}{2}, \frac{|b|}{2}]\) at the same time as the sum of these two roots is \(-2b\).

Next, we apply Theorem 2.4 (page 62 in Burden and Faires) to conclude that our algorithm converges at a rate of \(O(2^{-n})\) or better if \(x_0 \in [-\frac{|b|}{2}, \frac{|b|}{2}]\).

The desired region \(-b^2 \leq c \leq \frac{3}{4}b^2\), and \(b \neq 0\) are sketched in the Cartesian system with axes \(b\) and \(c\) in the last page.

\(\square\)
Question 5

Part a

If $c = 0$, then

$$g = -b,$$

and the conclusion clearly holds.

If $c \neq 0$, then

$$g(x) = -b - \frac{c}{x},$$

and we have

$$|g'(x)| = \frac{|c|}{x^2} \leq \frac{|c|}{2|c|} = \frac{1}{2},$$

whenever $x^2 \geq 2|c|$.

\qed

Part b

Whenever $x^2 \geq 2|c|$ and $b^2 \geq \frac{9}{2}|c|$, we have that

$$|g(x)| = |b + \frac{c}{x}| \geq |b| - \frac{|c|}{|x|} \geq \frac{3}{\sqrt{2}} \sqrt{|c|} - \frac{|c|}{\sqrt{2}|c|} = \sqrt{2}|c|,$$

which implies that

$$g(x)^2 \geq 2|c|.$$

\qed

Part c

Case 1: $c = 0$.

Clearly this gives the desired property.

Case 2: $c \neq 0$, and $b^2 \geq \frac{9}{2}|c|$.

We provide two different approaches here.

Approach 1

The interval $A = \{ x \mid x^2 \geq 2|c| \} = (-\infty, -\sqrt{2}|c|] \cup [\sqrt{2}|c|, \infty)$ from part a and b is the union of two disjoint intervals $(-\infty, -\sqrt{2}|c|]$ and $(\sqrt{2}|c|, \infty)$. So we seek conditions on $b$ on $c$ to make either $(-\infty, -\sqrt{2}|c|)$ or $(\sqrt{2}|c|, \infty)$ invariant under the function $g$.

The condition $b^2 \geq \frac{9}{2}|c|$ from part b can be splitted into two regions $b \geq \frac{3}{\sqrt{2}} \sqrt{|c|}$ and $b \leq -\frac{3}{\sqrt{2}} \sqrt{|c|}$. We actually can prove the stronger statements.
1) $b \geq \frac{3}{\sqrt{2}} |c|$ makes that $g$ maps $A$ to $(-\infty, -\sqrt{2}|c|]$.

2) $b \leq -\frac{3}{\sqrt{2}} |c|$ makes that $g$ maps $A$ to $(\sqrt{2}|c|, \infty)$.

Proof for 1)

Notice that

$$g(x) = -b - \frac{c}{x} \leq -\frac{3}{\sqrt{2}} \sqrt{|c|} + \frac{|c|}{x} \leq -\frac{3}{\sqrt{2}} \sqrt{|c|} + \frac{|c|}{\sqrt{2}|c|} = -\sqrt{2}|c|.$$ 

Proof for 2)

Notice that

$$g(x) = -b - \frac{c}{x} \geq \frac{3}{\sqrt{2}} \sqrt{|c|} - \frac{|c|}{|x|} \geq \frac{3}{\sqrt{2}} \sqrt{|c|} - \frac{|c|}{\sqrt{2}|c|} = \sqrt{2}|c|.$$ 

Hence if $b \geq \frac{3}{\sqrt{2}} |c|$ holds, no matter whether $x_0$ is in $(-\infty, -\sqrt{2}|c|]$ or $(\sqrt{2}|c|, \infty)$, $x_1$ will be in $(-\infty, -\sqrt{2}|c|]$, and therefore the whole sequence $\{x_n\}_{n=1}^\infty$ will stay in $(-\infty, -\sqrt{2}|c|]$. Thus we can apply Theorem 2.3 from the textbook to get the existence and uniqueness of the fixed point in $(-\infty, -\sqrt{2}|c|]$, and further use Theorem 2.4 to deduce the convergence of $\{x_n\}_{n=1}^\infty$ to such fixed point at a rate at least $O(2^{-n})$. Same argument can give the convergence of $\{x_n\}_{n=1}^\infty$ with a rate at least $O(2^{-n})$ when $b \leq -\frac{3}{\sqrt{2}} |c|$ holds.

Altogether, if we have that the starting point $x_0$ satisfies $x_0^2 \geq 2|c|$ and the condition $b^2 \geq \frac{9}{2} |c|$, the sequence converges at the desired rate.

Approach 2

The condition $b^2 \geq \frac{9}{2} |c|$ implies the discriminant for the equation $g(x) = x$ is

$$b^2 - 4c \geq \frac{9}{2} |c| - 4|c| = \frac{1}{2} |c| > 0.$$ 

Hence we have two real roots

$$x_\alpha = \frac{-b + \sqrt{b^2 - 4c}}{2}$$

and

$$x_\beta = \frac{-b - \sqrt{b^2 - 4c}}{2}.$$ 

Next we prove that there is exactly one root in the set $A = (-\infty, -\sqrt{2}|c|] \cup [\sqrt{2}|c|, \infty)$.

If they were both in $A$, we would have $|x_\alpha||x_\beta| \geq (\sqrt{2}|c|)^2 = 2|c|$, which contradicts with the fact that $x_\alpha x_\beta = c$. On the other hand, if $b \geq \frac{3}{\sqrt{2}} |c|$, we have that

$$x_\beta = \frac{-b - \sqrt{b^2 - 4c}}{2} \leq \frac{-\frac{3}{\sqrt{2}} \sqrt{|c|} - \sqrt{(\frac{3}{\sqrt{2}} \sqrt{|c|})^2 - 4|c|}}{2} = -\sqrt{2}|c|.$$
and if \( b \leq -\frac{3}{\sqrt{2}} \sqrt{|c|} \), we have that
\[
x_\alpha = \frac{-b + \sqrt{b^2 - 4c}}{2} \geq \frac{3}{\sqrt{2}} \sqrt{|c|} + \sqrt{\left(-\frac{3}{\sqrt{2}} \sqrt{|c|}\right)^2 - 4|c|} = \sqrt{2}|c|.
\]

Let \( y \) be the number from \( \{x_\alpha, x_\beta\} \) which is in \( A \). Then \( y = -b - \frac{c}{y} \) implies
\[
|x_{n+1} - y| = \left|(-b - \frac{c}{x_n}) - (-b - \frac{c}{y})\right| = |c|\left|\frac{1}{x_n} - \frac{1}{y}\right| = \frac{|c|}{|y||x_n|}|x_n - y| \quad (*)
\]

Since \( y \in A \), we get \( y \geq \sqrt{2|c|} \), and by part b, if the starting point \( x_0 \) satisfies \( x_0^2 \geq 2|c| \), we have \( |x_n| \geq \sqrt{2|c|} \).

Plugging these estimates to (*) yields
\[
|x_{n+1} - y| \leq \frac{|c|}{|y||x_n|}|x_n - y| \leq \frac{1}{2}|x_n - y|.
\]

We can apply it repeatedly to obtain
\[
|x_n - y| \leq 2^{-n}|x_0 - y|,
\]
that is,
\[
x_n = y + O(2^{-n}).
\]

The sketch for \( \{(b, c) | b^2 \geq \frac{9}{2}|c|\} \) in a Cartesian system with axes \( b \) and \( c \) is in the last page.
Question 6

Part a

Taking limit on both sides of the given recursive relation

\[ x_{n+1} = x_n(2 - ax_n) \]

gives

\[ x = x(2 - ax), \]

which is

\[ x(1 - ax) = 0. \]

As \( a > 0 \), we get

\[ x = \frac{1}{a} \text{ or } x = 0. \]

Part b

The given recursive relation

\[ x_{n+1} = x_n(2 - ax_n) \]

can be written as

\[ ax_{n+1} - 1 = -(ax_n - 1)^2. \]

Applying this repeatedly to get

\[ |ax_n - 1| = |ax_0 - 1|^{2^n}. \]

In order to have the convergence of the sequence \( \{x_n\}_{n=0}^{\infty} \), we need that

\[ |ax_0 - 1| < 1, \]

which is

\[ |x_0 - \frac{1}{a}| < \frac{1}{a}. \]

Hence for \( x_0 \in (0, \frac{2}{a}) \), we have that

\[ \lim_{n \to \infty} x_n = \frac{1}{a}. \]

Note that \( x_0 = 0 \) or \( x_0 = \frac{2}{a} \) implies that \( x_n = 0 \) for any \( n \geq 1 \), and further that

\[ \lim_{n \to \infty} x_n = 0. \]

Hence whenever \( x_0 \in [\alpha, \beta] \) (with \( \alpha = 0, \beta = \frac{2}{a} \)), we have the convergence of \( \{x_n\}_{n=0}^{\infty} \).
Part c

Notice that

\[ |ax_n - 1| = |ax_0 - 1|^{2^n} \]

from part b tells

\[ |x_n - \frac{1}{a}| = \frac{1}{a}|ax_0 - 1|^{2^n}, \]

which indicates that

\[ x_n = \frac{1}{a} + O(|ax_0 - 1|^{2^n}). \]

Part d

Let \( I \) be the \( 2 \times 2 \) identity matrix.

We first generalize part a to \( 2 \times 2 \) matrices.

Taking limit on both sides of the given recursive relation

\[ X_{n+1} = X_n(2I - AX_n) \]

gives

\[ X = X(2I - AX), \]

which is,

\[ X(I - AX) = 0. \]

Hence the limit \( X \) for \( X_n \) satisfies the equation

\[ X(I - AX) = 0. \]

For the generalization to part b, we first establish the following lemma.

**Lemma** Let \( \rho(M) \) be the maximum of the absolute value of the eigenvalues of the matrix \( M \). If \( \rho(M) < 1 \), then

\[ \lim_{n \to \infty} M^n = 0, \]

which means that each entry of the matrix powers converges to 0.

**Proof**

Let \( J \) be the Jordan canonical (normal) form of \( M \).

Then \( J \) has the form
\[
\begin{pmatrix}
J_1 & 0 & 0 & \cdots & 0 \\
0 & J_2 & 0 & \cdots & 0 \\
\vdots & 0 & \ddots & \cdots & \vdots \\
0 & \cdots & 0 & J_{s-1} & 0 \\
0 & \cdots & \cdots & 0 & J_s
\end{pmatrix},
\]

where \( J_i \) is a \( m_i \times m_i \) matrix with eigenvalue \( \lambda_i \) on the diagonal and 1 on the superdiagonal, that is,

\[
J_i = \begin{pmatrix}
\lambda_i & 1 & 0 & \cdots & 0 \\
0 & \lambda_i & 1 & \cdots & 0 \\
\vdots & 0 & \ddots & \cdots & \vdots \\
0 & \cdots & 0 & \lambda_i & 1 \\
0 & \cdots & \cdots & 0 & \lambda_i
\end{pmatrix},
\]

Then there exists invertible matrix \( P \) such that

\[
M = PJP^{-1}.
\]

It is easy to see that

\[
M^n = PJ^nP^{-1},
\]

and since \( J \) is block-diagonal, we have

\[
J^n = \begin{pmatrix}
J_1^n & 0 & 0 & \cdots & 0 \\
0 & J_2^n & 0 & \cdots & 0 \\
\vdots & 0 & \ddots & \cdots & \vdots \\
0 & \cdots & 0 & J_{s-1}^n & 0 \\
0 & \cdots & \cdots & 0 & J_s^n
\end{pmatrix},
\]
and for \( n \geq m_i - 1 \),

\[
\begin{pmatrix}
\lambda_i^n & \binom{n}{1} \lambda_i^{n-1} & \binom{n}{2} \lambda_i^{n-2} & \cdots & \binom{n}{m_i-1} \lambda_i^{n-m_i+1} \\
0 & \lambda_i^n & \binom{n}{1} \lambda_i^{n-1} & \cdots & \binom{n}{m_i-2} \lambda_i^{n-m_i+2} \\
\vdots & 0 & \vdots & \cdots & \vdots \\
0 & \cdots & 0 & \lambda_i^n & \binom{n}{1} \lambda_i^{n-1} \\
0 & \cdots & \cdots & 0 & \lambda_i^n
\end{pmatrix}
\]

If \( \rho(M) < 1 \), then \( |\lambda_i| < 1 \) for all \( i \), and each entry of \( J_i^n \) therefore converges to 0, which indicates

\[
\lim_{n \to \infty} J_i^n = 0,
\]

and therefore

\[
\lim_{n \to \infty} M^n = \lim_{n \to \infty} PJ^n P^{-1} = 0.
\]

Assume \( A \) is invertible.

Notice that

\[
X_{n+1} = X_n(2I - AX_n)
\]

can be written as

\[
AX_{n+1} - I = -(AX_n - I)^2.
\]

Applying this repeatedly to get

\[
AX_n - I = -(AX_0 - I)^{2^n}. \quad (*)
\]

If \( \rho(AX_0 - I) < 1 \), by Lemma, we have

\[
\lim_{n \to \infty} (AX_0 - I)^n = 0.
\]

Notice that \( \{(AX_0 - I)^{2^n}\} \) is a subsequence of the convergent sequence \( \{(AX_0 - I)^n\} \), we get

\[
\lim_{n \to \infty} (AX_0 - I)^{2^n} = 0.
\]

Therefore we obtain that

\[
\lim_{n \to \infty} (AX_n - I) = 0.
\]

Since \( A \) is assumed to be invertible, multiplying through the equation above by \( A^{-1} \) gives

\[
\lim_{n \to \infty} (X_n - A^{-1}) = 0.
\]
That is to say, the assumptions that $\rho(AX_0 - I) < 1$ and that $A$ is invertible indicate

$$\lim_{n \to \infty} X_n = A^{-1}.$$ 

Finally, let’s generalize part c to the $2 \times 2$ matrix under the assumption that $\rho(AX_0 - I) < 1$.

Assume $AX_0$ is diagonalizable, then $AX_0 - I$ is also diagonalizable. Hence there’s invertible matrix $V$ and diagonal matrix $D$ such that

$$AX_0 - I = VDV^{-1}.$$ 

Here

$$D = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix},$$

where $\beta_1$ and $\beta_2$ are eigenvalues of $AX_0 - I$, and hence $|\beta_1| \leq \rho(AX_0 - I) < 1$, and $|\beta_2| \leq \rho(AX_0 - I) < 1$.

According to (*), we have

$$AX_n - I = -(AX_0 - I)^{2^n} = -VD^{2^n}V^{-1},$$

which indicates

$$X_n - A^{-1} = -A^{-1}VD^{2^n}V^{-1}.$$ 

Define $\phi(B) = \max\{|B_{11}|, |B_{12}|, |B_{21}|, |B_{22}|\}$ for any $2 \times 2$ matrix.

Then we have

$$\phi(A^{-1}VD^{2^n}) \leq \phi(A^{-1}V) \cdot \max\{\beta_1^{2^n}, \beta_2^{2^n}\} = \phi(A^{-1}V)(\rho(AX_0 - I))^{2n},$$

and further

$$\phi(X_n - A^{-1}) = \phi(A^{-1}VD^{2^n}V^{-1}) \leq 2\phi(A^{-1}VD^{2^n})\phi(V^{-1}) \leq 2\phi(A^{-1}V)(\rho(AX_0 - I))^{2n}\phi(V^{-1}),$$

which implies

$$(X_n - A^{-1})_{ij} \leq 2\phi(A^{-1}V)(\rho(AX_0 - I))^{2n}\phi(V^{-1})$$

for all $1 \leq i, j \leq 2$.

Since $\phi(A^{-1}V)$ and $\phi(V^{-1})$ depends only on $A$ and $X$, we have

$$(X_n - A^{-1})_{ij} = O((\rho(AX_0 - I))^{2n})$$

for all $1 \leq i, j \leq 2$, which is the convergence rate of the element of $X_n$ to $A^{-1}$. 

$\square$