227AT Control with Multiplicative Noise

Ayah Ahmad, Christian Ikeokwu, Ebonye Smith, Simeon Adebola

Introduction and Background

In this project, we will use different techniques to analyze *control systems*, which are discrete-time dynamical systems that have external inputs. For concreteness and simplicity, we work in the case where everything is a scalar, though the ideas generalize to vector systems.

Problems

1 Environment Implementation

We did it! :)

2 Control Noise: Derivations

In this problem we consider a simplified version of the multiplicative noise system where the noise is only on the input. More formally, we set $c = 1$ and $\alpha = \gamma = 0$, so $A_t = a$ and $C_t = 1$ deterministically. In other words, we analyze the system

$$X_{t+1} = aX_t + B_tU_t \qquad \forall t \geq 0 \tag{1}$$
$$Y_t = X_t, \tag{2}$$

(a) First, let us determine what the state second moment $\mathbb{E}[X_{t+1}^2|Y_t]$. Fix $t \geq 0$. Show that

$$\mathbb{E}[X_{t+1}^2|Y_t] = a^2Y_t^2 + 2abU_tY_t + (b^2 + \beta^2)U_t^2$$

Solution:

$$\begin{aligned}
\mathbb{E}[X_{t+1}^2|Y_t] &= \mathbb{E}[(aX_t + B_tU_t)^2|Y_t] \\
&= \mathbb{E}[(aY_t + B_tU_t)^2] \\
&= \mathbb{E}[a^2Y_t^2 + 2aY_tB_tU_t + B_t^2U_t^2] \\
&= a^2\mathbb{E}[Y_t^2|Y_t] + 2a\mathbb{E}[Y_tB_tU_t|Y_t] + \mathbb{E}[B_t^2U_t^2|Y_t] \\
&= a^2\mathbb{E}[Y_t^2|Y_t] + 2a\mathbb{E}[B_t]\mathbb{E}[U_t]\mathbb{E}[Y_t|Y_t] + \mathbb{E}[B_t^2]\mathbb{E}[U_t^2] \\
&= a^2Y_t^2 + 2abU_tY_t + (b^2 + \beta^2)U_t^2
\end{aligned}$$

□

(b) Now, we will determine what exactly the optimal greedy memory-1 policy $F^\star$ is. Fix $t \geq 0$. Define $F^\star\colon \mathbb{R}^{t+1} \to \mathbb{R}$ by

$$F_t^\star(Y_{(t)}) = \operatorname*{argmin}_{U_t \in \mathbb{R}} \mathbb{E}[X_{t+1}^2|Y_t]. \tag{3}$$

Show that

$$F_t^\star(Y_{(t)}) = -\frac{ab}{b^2 + \beta^2}Y_t^2, \quad \forall t \geq 0 \tag{4}$$

so that $F_t^\star$ is a *linear* function of *only* Y_t, and the strategy is the *same* regardless of the value of t. This optimal policy $F^\star$ is therefore called a *linear period-1* (or *linear time-invariant*) policy.

Solution:

$$F_t^\star(Y_{(t)}) = \operatorname*{argmin}_{U_t \in \mathbb{R}} \mathbb{E}[X_{t+1}^2|Y_t] = \operatorname*{argmin}_{U_t \in \mathbb{R}} a^2Y_t^2 + 2abU_tY_t + (b^2 + \beta^2)U_t^2 \tag{5}$$

$$\nabla_{U_t} F_t^\star(\cdot) = 0 + 2abY_t + 2(b^2 + \beta^2)U_t$$

$$\begin{aligned}
2abY_t + 2(b^2 + \beta^2)U_t^* &= 0 \\
\implies U_t^* &= -\frac{ab}{b^2 + \beta^2}Y_t
\end{aligned}$$

□

(c) With the optimal control $U_t = F_t^\star(Y_{(t)})$, where $F_t^\star$ was given in part (b), show that

$$\mathbb{E}[X_{t+1}^2|Y_t] = \frac{a^2\beta^2}{b^2+\beta^2}Y_t^2 \quad \forall t \geq 0. \tag{6}$$

Solution:

$$\begin{aligned}
\mathbb{E}[X_{t+1}^2|Y_t] &= a^2Y_t^2 + 2abU_t^\star Y_t + (b^2+\beta^2)U_t^{\star 2} \\
&= a^2Y_t^2 + 2abY_t \cdot -\frac{ab}{b^2+\beta^2}Y_t + (b^2+\beta^2)\left(-\frac{ab}{b^2+\beta^2}Y_t\right)^2 \\
&= a^2Y_t^2 - 2\frac{a^2b^2}{(b^2+\beta^2)}Y_t^2 + \frac{(b^2+\beta^2)a^2b^2}{(b^2+\beta^2)^2}Y_t^2 \\
&= a^2Y_t^2 \cdot \frac{(b^2+\beta^2)}{(b^2+\beta^2)} - 2\frac{a^2b^2}{(b^2+\beta^2)}Y_t^2 + \frac{a^2b^2}{(b^2+\beta^2)}Y_t^2 \\
&= \frac{a^2b^2+a^2\beta^2}{(b^2+\beta^2)}Y_t^2 - \frac{a^2b^2}{(b^2+\beta^2)}Y_t^2 \\
&= \frac{a^2b^2+a^2\beta^2-a^2b^2}{(b^2+\beta^2)}Y_t^2 \\
&= \frac{a^2\beta^2}{(b^2+\beta^2)}Y_t^2
\end{aligned}$$

□

(d) With the optimal control $U_t = F_t^\star(Y_{(t)})$, where where $F_t^\star$ was given in part (b), and using the result from part (c), show that

$$\mathbb{E}[X_t^2] = \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^t \quad \forall t \geq 0. \tag{7}$$

Solution:

Theorem 2.1. $\mathbb{E}[X_t^2] = \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^t$

Proof. By induction on t

Base case : $t = 0$.

$$\mathbb{E}[X_0^2] = \mathbb{E}[1^2] = 1 = \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^0$$

IH: Let $\mathbb{E}[X_k^2] = \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^k$

Consider $\mathbb{E}[X_{k+1}^2]$.

$$\begin{aligned}
\mathbb{E}[X_{k+1}^2] &= \mathop{\mathbb{E}}_{Y_k}[\mathbb{E}[X_{k+1}^2|Y_k]] \\
&= \mathop{\mathbb{E}}_{Y_k}\left[\frac{a^2\beta^2}{(b^2+\beta^2)}Y_k^2\right] \\
&= \frac{a^2\beta^2}{b^2+\beta^2}\mathbb{E}[Y_k^2] \\
&= \frac{a^2\beta^2}{b^2+\beta^2}\mathbb{E}[X_k^2] \\
&= \frac{a^2\beta^2}{b^2+\beta^2}\left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^k \\
&= \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^{k+1}
\end{aligned}$$

□

(e) With the optimal control $U_t = F_t^\star(Y_{(t)})$, where $F_t^\star$was given in part (b), and using the result from part (d), show that the system is stable in the second moment if and only if

$$|a| \leq \sqrt{1+\frac{b^2}{\beta^2}}. \tag{8}$$

Solution:

($\Rightarrow$) if $|a| \leq \sqrt{1+\frac{b^2}{\beta^2}}$ Then this implies

$$\begin{aligned}
\mathbb{E}[X_t^2] &= \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^t && \forall t \\
&\leq \left(\frac{\left(\sqrt{1+\frac{b^2}{\beta^2}}\right)^2\beta^2}{b^2+\beta^2}\right)^t \\
&= \left(\frac{b^2+\beta^2}{b^2+\beta^2}\right)^t \\
&= 1^2 = 1 \leq M && \text{for some fixed } M \geq 1 \in \mathbb{R}
\end{aligned}$$

($\Leftarrow$) if $\mathbb{E}[X_t^2] \leq M \quad \forall t$, for some fixed $M \in \mathbb{R}$ this implies

$$\mathbb{E}[X_t^2] = \left(\frac{a^2\beta^2}{b^2+\beta^2}\right)^t \leq M \qquad \forall t$$

$$\implies \left(\frac{a^2\beta^2}{b^2+\beta^2}\right) \leq M^{1/t} \qquad \forall t$$

$$\implies \left(\frac{a^2\beta^2}{b^2+\beta^2}\right) \leq \lim_{t\to\infty} M^{1/t}$$

$$\implies \left(\frac{a^2\beta^2}{b^2+\beta^2}\right) \leq 1$$

$$\implies a^2 \leq \frac{\beta^2+b^2}{\beta^2}$$

$$\implies |a| \leq \sqrt{\frac{\beta^2+b^2}{\beta^2}}$$

$$\implies |a| \leq \sqrt{1+\frac{b^2}{\beta^2}}$$

□

3 State and Control Noise: Derivations

In this problem we consider a more general version of the previous problem, where the noise is on the state as well as on the control. More formally, we set $c = 1$ and $\gamma = 0$, so $C_t = 1$ deterministically. In other words, we analyze the system

$$X_{t+1} = A_t X_t + B_t U_t \qquad \forall t \geq 0 \tag{9}$$
$$Y_t = X_t, \tag{10}$$

with $\mathbb{E}[X_0] = \mu$ and $\mathbb{E}[X_0^2] = \mu^2 + \sigma^2$, and attempt to minimize the loss $L(F) = \mathbb{E}[\sum_{t=0}^{\infty} X_{t+1}^2 + \lambda U_t^2]$.

(a) Using the same approach as Problem 2, show that an optimal control policy is

$$F_t^\star(Y_{(t)}) = -\frac{ab}{b^2 + \beta^2 + \lambda} Y_t \quad \forall t \geq 0 \tag{11}$$

and that the system is stabilizable in the second moment if and only if

$$\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2 + \lambda)^2)}{(b^2 + \beta^2 + \lambda)^2} \leq 1 \tag{12}$$

.

i) **Solution**:

$$\begin{aligned}
\mathbb{E}[X_{t+1}^2 | Y_t] &= \mathbb{E}[(A_t X_t + B_t U_t)^2] \\
&= \mathbb{E}[A_t^2 X_t^2 + 2A_t X_t B_t U_t + B_t^2 U_t^2 | Y_t] \\
&= (a^2 + \alpha^2)\mathbb{E}[X_t^2 | Y_t] + 2ab\mathbb{E}[U_t]\mathbb{E}[X_t | Y_t] + (b^2 + \beta^2)\mathbb{E}[U_t^2] \\
&= (a^2 + \alpha^2)\mathbb{E}[Y_t^2 | Y_t] + 2ab\mathbb{E}[U_t]\mathbb{E}[Y_t | Y_t] + (b^2 + \beta^2)\mathbb{E}[U_t^2 | Y_t] \\
&= (a^2 + \alpha^2)Y_t^2 + 2abU_t Y_t + (b^2 + \beta^2)U_t^2
\end{aligned}$$

□

ii) **Solution**:

$$\begin{aligned}
F_t^\star(Y_{(t)}) &= \operatorname*{argmin}_{U_t \in \mathbb{R}} \mathbb{E}[X_{t+1}^2 | Y_t] + \lambda U_t^2 \\
&= \operatorname*{argmin}_{U_t \in \mathbb{R}} (a^2 + \alpha^2)Y_t^2 + 2abU_t Y_t + (b^2 + \beta^2)U_t^2 + \lambda U_t^2 \\
&= \operatorname*{argmin}_{U_t \in \mathbb{R}} (a^2 + \alpha^2)Y_t^2 + 2abU_t Y_t + (b^2 + \beta^2 + \lambda)U_t^2
\end{aligned}$$

$$\nabla_{U_t} F_t^\star(\cdot) = 0 + 2abY_t + 2(b^2 + \beta^2 + \lambda)U_t$$

Using the first order conditions on convex functions

$$\begin{aligned}
2abY_t + 2(b^2 + \beta^2 + \lambda)U_t^* = 0 \\
\implies U_t^* = -\frac{ab}{b^2 + \beta^2 + \lambda} Y_t
\end{aligned}$$

□

iii) **Solution**:

$$\begin{aligned}
\mathbb{E}[X_{t+1}^2|Y_t] &= (a^2+\alpha^2)Y_t^2 + 2abU_t^*Y_t + (b^2+\beta^2)(U_t^*)^2 \\
&= (\alpha^2+a^2)Y_t^2 + 2ab\left(-\frac{ab}{b^2+\beta^2+\lambda}Y_t\right)Y_t + (b^2+\beta^2)\left(-\frac{ab}{b^2+\beta^2+\lambda}Y_t\right)^2 \\
&= \alpha^2 Y_t^2 + a^2Y_t^2 - \frac{2a^2b^2}{(b^2+\beta^2+\lambda)}Y_t^2 + \frac{(b^2+\beta^2)a^2b^2}{(b^2+\beta^2+\lambda)^2}Y_t^2 \\
&= \left(\alpha^2 + \frac{a^2(b^2+\beta^2+\lambda)^2}{(b^2+\beta^2+\lambda)^2} - \frac{2(b^2+\beta^2+\lambda)a^2b^2}{(b^2+\beta^2+\lambda)^2} + \frac{(b^2+\beta^2)a^2b^2}{(b^2+\beta^2+\lambda)^2}\right)Y_t^2 \\
&= \left(\alpha^2 + \frac{a^2(b^2+\beta^2+\lambda)^2}{(b^2+\beta^2+\lambda)^2} - \frac{(b^2+\beta^2+2\lambda)a^2b^2}{(b^2+\beta^2+\lambda)^2}\right)Y_t^2 \\
&= \left(\alpha^2 + \frac{a^2\beta^2b^2 + a^2\beta^4 + 2a^2\beta^2\lambda + a^2\lambda^2}{(b^2+\beta^2+\lambda)^2}\right)Y_t^2 \\
&= \left(\alpha^2 + \frac{a^2(\beta^2b^2 + \beta^4 + 2\beta^2\lambda + \lambda^2)}{(b^2+\beta^2+\lambda)^2}\right)Y_t^2 \\
&= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)Y_t^2
\end{aligned}$$

□

iv) **Solution**:

Theorem 3.1.

$$\mathbb{E}[X_t^2] = \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^t (\mu^2+\sigma^2) \tag{13}$$

Proof. By induction on t

Base case : $t = 0$.

$$\mathbb{E}[X_0^2] = (\mu^2+\sigma^2) = \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^0 (\mu^2+\sigma^2)$$

IH: Let $\mathbb{E}[X_k^2] = \left(\alpha^2 + \frac{a^2(b^2\beta^2+(\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^k (\mu^2+\sigma^2)$

Consider $\mathbb{E}[X_{k+1}^2]$.

$$\begin{aligned}
\mathbb{E}[X_{k+1}^2] &= \mathop{\mathbb{E}}_{Y_k}[\mathbb{E}[X_{k+1}^2|Y_k]] \\
&= \mathop{\mathbb{E}}_{Y_k}\left[\left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)Y_k^2\right] \\
&= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)\mathbb{E}[Y_k^2] \\
&= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)\mathbb{E}[X_k^2] \\
&= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)\left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^k (\mu^2+\sigma^2) \\
&= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^{k+1} (\mu^2+\sigma^2)
\end{aligned}$$

□

v) **Solution**:

($\Rightarrow$) if $\left(\alpha^2 + \frac{a^2(b^2\beta^2+(\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right) \leq 1$. Then

$$\begin{aligned}
\mathbb{E}[X_t^2] &= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^t (\mu^2+\sigma^2) \\
&\leq (1)^t(\mu^2+\sigma^2) \\
&\leq (\mu^2+\sigma^2) \leq M \qquad \text{for some fixed } M \geq \mu^2+\sigma^2 \in \mathbb{R}
\end{aligned}$$

($\Leftarrow$) if $\mathbb{E}[X_t^2] \leq M \quad \forall t$, for some fixed $M \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{E}[X_t^2] &= \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^t (\mu^2+\sigma^2) \quad \leq M \\
\implies \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right)^t &\leq \frac{M}{\mu^2+\sigma^2} \qquad \forall t \\
\implies \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right) &\leq \left(\frac{M}{\mu^2+\sigma^2}\right)^{1/t} \qquad \forall t \\
\implies \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right) &\leq \lim_{t\to\infty}\left(\frac{M}{\mu^2+\sigma^2}\right)^{1/t} \\
\implies \left(\alpha^2 + \frac{a^2(b^2\beta^2 + (\beta^2+\lambda)^2)}{(b^2+\beta^2+\lambda)^2}\right) &\leq 1
\end{aligned}$$

□

4 Observation Noise: Derivations

Now, we will consider the complementary problem to the control noise system. That is, we will consider a system where the noise is only on the observation. Formally, we set $b = 1$ and $\alpha = \beta = 0$, so $A_t = a$ and $B_t = 1$ deterministically. In other words, we analyze the system

$$X_{t+1} = aX_t + U_t \qquad \forall t \geq 0 \tag{14}$$
$$Y_t = C_t X_t. \tag{15}$$

Let us also assume (for this problem only) that $X_0 = 1$ deterministically (so $\mu = 1$ and $\sigma = 0$), and that our regularization parameter $\lambda = 0$, so that the loss we attempt to minimize is $L(F) = \mathbb{E}[\sum_{t=0}^{\infty} X_{t+1}^2]$.

(a) Show by induction on t that, if U is a linear memory-1 period-1 greedy control policy, i.e., if for all t we have $U_t = F_t(Y_t) = \theta Y_t$, then

$$\mathbb{E}[X_t^2] = (a^2 + 2ac\theta + (c^2 + \gamma^2)\theta^2)^t, \quad \forall t \geq 0 \tag{16}$$

Solution:

$$\begin{aligned}
\mathbb{E}[X_{t+1}^2] &= \mathbb{E}[(aX_t + U_t)^2] \\
&= \mathbb{E}[(aX_t + \theta C_t X_t)^2] \\
&= \mathbb{E}[(a + \theta C_t)^2 X_t^2] \\
&= \mathbb{E}[(a + \theta C_t)^2]\mathbb{E}[X_t^2] \\
&= (a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))\mathbb{E}[X_t^2]
\end{aligned}$$

$$\therefore \ \mathbb{E}[X_{t+1}^2] = (a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))\mathbb{E}[X_t^2] \tag{17}$$

Theorem 4.1.

$$\mathbb{E}[X_t^2] = (a^2 + 2ac\theta + (c^2 + \gamma^2)\theta^2)^t, \quad \forall t \geq 0 \tag{18}$$

Proof. By induction on t

Base case : $t = 0$.

$$\mathbb{E}[X_0^2] = 1^2 = 1 = (a^2 + 2ac\theta + (c^2 + \gamma^2)\theta^2)^0$$

IH: Let $\mathbb{E}[X_k^2] = (a^2 + 2ac\theta + (c^2 + \gamma^2)\theta^2)^k$ Consider $\mathbb{E}[X_{k+1}^2]$.

$$\begin{aligned}
\mathbb{E}[X_{k+1}^2] &= (a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))\mathbb{E}[X_t^2] && \text{by 17} \\
&= (a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))(a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))^k \\
&= (a^2 + 2ac\theta + \theta^2(c^2 + \gamma^2))^{k+1}
\end{aligned}$$

□

(b) Suppose that $U_t = F_t(Y_{(t)}) = \theta Y_t$ for all $t \geq 0$, and fix a particular t. Define

$$\theta^\star = \operatorname*{argmin}_{\substack{\theta \in \mathbb{R} \\ U_t = \theta Y_t}} . \tag{19}$$

Show that

$$\theta^\star = -\frac{ac}{c^2 + \gamma^2}. \tag{20}$$

This result shows that the optimal linear memory-1 period-1 greedy control policy is

$$F_t^\star(Y_{(t)}) = -\frac{ac}{c^2+\gamma^2}Y_t. \tag{21}$$

Contrast this to the optimal control policy $F_t^\star(Y_{(t)}) = -\frac{ab}{b^2+\beta^2}Y_t^2$ derived in Problem 2 **Solution**:

$$\begin{aligned}
\theta^\star &= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ \mathbb{E}[X_{t+1}^2] \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ (a^2+2ac\theta+\theta^2(c^2+\gamma^2))^{t+1} \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ \ln(a^2+2ac\theta+\theta^2(c^2+\gamma^2))^{t+1} \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ (t+1)\ln((a^2+2ac\theta+\theta^2(c^2+\gamma^2)) \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ \ln((a^2+2ac\theta+\theta^2(c^2+\gamma^2)) \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ \exp\{\ln((a^2+2ac\theta+\theta^2(c^2+\gamma^2))\} \\
&= \underset{\theta\in\mathbb{R}, U_t=\theta Y_t}{\operatorname{argmin}} \ (a^2+2ac\theta+\theta^2(c^2+\gamma^2)) = I(\cdot)
\end{aligned}$$

Using first order conditions on $I(\cdot)$

$$\nabla_\theta I(\cdot) = 0 + 2ac + 2\theta(c^2+\gamma^2) \tag{22}$$

$$\begin{aligned}
0 &= 2ac + 2\theta^*(c^2+\gamma^2) \\
\implies \theta^* &= -\frac{ac}{c^2+\gamma^2}
\end{aligned}$$

(c) With the control $U_t = F_t^\star(Y_{(t)})$, where F_t was given in part (b), show that

$$\mathbb{E}[X_t^2] = \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right)^t, \quad \forall t \geq 0. \tag{23}$$

Solution:

$$\begin{aligned}
E[X_{t+1}^2] &= (a^2+2ac\theta^*+(c^2+\gamma^2)\theta^{*2})^t \\
&= \left(a^2+2ac\left(-\frac{ac}{c^2+\gamma^2}\right)+(c^2+\gamma^2)\left(-\frac{ac}{c^2+\gamma^2}\right)^2\right)^t \\
&= \left(a^2+2ac\left(-\frac{ac}{c^2+\gamma^2}\right)+(c^2+\gamma^2)\frac{a^2c^2}{(c^2+\gamma^2)^2}\right)^t \\
&= \left(a^2+2ac\left(-\frac{ac}{c^2+\gamma^2}\right)+\frac{a^2c^2}{(c^2+\gamma^2)}\right)^t \\
&= \left(\frac{a^2c^2+a^2\gamma^2-a^2c^2}{c^2+\gamma^2}\right)^t \\
&= \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right)^t
\end{aligned}$$

(d) With the control $U_t = F_t^\star(Y_{(t)})$, where F_t was given in part (b), and using the result from part (c), show that the system is stable in the second moment if and only if

$$|a| \leq \sqrt{1 + \frac{c^2}{\gamma^2}}. \tag{24}$$

Solution:

($\Rightarrow$) if $|a| \leq \sqrt{1 + \frac{c^2}{\gamma^2}}$ Then this implies

$$\begin{aligned}
\mathbb{E}[X_t^2] &= \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right)^t && \forall t \\
&\leq \left(\frac{\left(\sqrt{1+\frac{c^2}{\gamma^2}}\right)^2 \gamma^2}{c^2+\gamma^2}\right)^t \\
&= \left(\frac{c^2+\gamma^2}{c^2+\gamma^2}\right)^t \\
&= 1^2 = 1 \leq M && \text{for some fixed } M \geq 1 \in \mathbb{R}
\end{aligned}$$

($\Leftarrow$) if $\mathbb{E}[X_t^2] \leq M \quad \forall t$, for some fixed $M \in \mathbb{R}$ this implies

$$\begin{aligned}
\mathbb{E}[X_t^2] = \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right)^t &\leq M && \forall t \\
\implies \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right) &\leq M^{1/t} && \forall t \\
\implies \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right) &\leq \lim_{t\to\infty} M^{1/t} \\
\implies \left(\frac{a^2\gamma^2}{c^2+\gamma^2}\right) &\leq 1 \\
\implies a^2 &\leq \frac{\gamma^2+c^2}{\gamma^2} \\
\implies |a| &\leq \sqrt{\frac{\gamma^2+c^2}{\gamma^2}} \\
\implies |a| &\leq \sqrt{1+\frac{c^2}{\gamma^2}}
\end{aligned}$$

□

5 Introduction to Policy Gradient

Using optimal control policies derived by hand, such as in Problem 2, Problem 3, and Problem 4, provably ensures that the control system is stable in the second moment under broad conditions. However, implementing the optimal control requires knowledge of the environment, namely the parameters $a, b, c, \alpha, \beta, \gamma$. At face value, this is an unrealistic assumption; most of the time we only have noisy estimates, at most, for these parameters. Thus, we introduces the *policy gradient* method to learn a control policy from data without having full knowledge of the environment.

The policy gradient method is conceptually very similar to gradient descent. It is an iterative procedure where at each iteration we estimate the gradient of the cost function $L(F) = \mathbb{E}[\sum_{t=0}^{\infty} X_{t+1}^2 + \lambda U_t^2]$ and then use the gradient to update the control policy.

One problem we have is that we cannot run gradient descent on control policies, since we are directly optimizing over functions. the solution to this this is a rather common idea; we parameterize our control policy $F = (F_0, F_1, ...)$ by some parameter θ, so it will be written as $F(\theta)$, where the policy at time t is $F_t(\cdot; \theta)$. Instead of taking a gradient step over F, we just take a gradient step over θ.

Algorithm 1 Our policy gradient algorithm.

function PolicyGradient
  Initialize at some parameter θ_0
  for $i \in 0, ..., M-1$ **do**
    for $j \in 1, ..., N$ **do**
      Begin new trajectory.
      for $t \in 0, ..., T$ **do**
        Collect observation Y_t^j from environment.
        Sample $W_t^j \sim \mathcal{N}(0, \omega^2)$
        Hand control Input $\tilde{U_t^j} = F_t(Y_{(t)}^j; \theta_i) + W_t^j$ to environment and collect loss ℓ_t^j
      end for
    end for
    Approximate $\nabla_\theta L(\theta_i) \approx \frac{1}{N} \sum_{j=1}^{N} \left(\sum_{t=0}^{T} \nabla_\theta \log \left(\pi_{\theta_i} \left(\tilde{U_t^j} | Y_{(t)}^j \right) \right) \right) \left(\sum_{t=0}^{T} \ell_t^j \right)$
    $\theta_i + 1 \leftarrow \theta_i - \eta \nabla_\theta L(\theta_i)$
  end for
  return $F(\theta_M)$
end function

(a) Suppose that $U_t = F_t(Y_{(t)}; \theta) = \theta Y_t$ for $\theta \in \mathbb{R}$. Show that

$$\nabla_\theta \log \left(\pi_\theta \left(\tilde{U_t} | Y_t \right) \right) = \frac{1}{\omega^2} \left(\tilde{U_t} - \theta Y_t \right) Y_t \quad \forall t \geq 0 \tag{25}$$

Solution:

$$\tilde{U_t} | Y_t = F_t(Y_{(t)}; \theta) + \mathcal{N}(0, \omega^2) \tag{26}$$
$$= \theta Y_t + \mathcal{N}(0, \omega^2) \tag{27}$$
$$\implies \tilde{U_t} | Y_t \sim \mathcal{N}(\theta Y_t, \omega^2) \tag{28}$$
$$\tag{29}$$

Thus $\pi_\theta\left(\tilde{U}_t|Y_t\right) = \frac{1}{\omega\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{U}_t-\theta Y_t}{\omega}\right)^2\right\}$ is the p.d.f for a $\mathcal{N}(\theta Y_t, \omega^2)$ random variable

$$\begin{aligned}
\log\left(\pi_\theta\left(\tilde{U}_t|Y_t\right)\right) &= \log\left(\frac{1}{\omega\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{U}_t-\theta Y_t}{\omega}\right)^2\right\}\right) \\
&= \log\left(\exp\left\{-\frac{1}{2}\left(\frac{\tilde{U}_t-\theta Y_t}{\omega}\right)^2\right\}\right) - \log\left(\omega\sqrt{2\pi}\right) \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\tilde{U}_t-\theta Y_t}{\omega}\right)^2\right\} - \log\left(\omega\sqrt{2\pi}\right)
\end{aligned}$$

Thus we can take the gradient

$$\begin{aligned}
\nabla_\theta \log\left(\pi_\theta\left(\tilde{U}_t|Y_t\right)\right) &= 2\cdot -\frac{1}{2}\left(\frac{\tilde{U}_t-\theta Y_t}{\omega}\right)\cdot -\frac{Y_t}{\omega} + 0 \\
&= (\tilde{U}_t-\theta Y_t)\cdot\frac{1}{\omega}\cdot\frac{Y_t}{\omega} \\
&= (\tilde{U}_t-\theta Y_t)\cdot\frac{Y_t}{\omega^2} \qquad\qquad \forall t \geq 0
\end{aligned}$$

□

6 Input Noise and Observation Noise: Policy Gradient

In this problem, we will show via policy gradient that the optimal greedy linear memory-1 period-1 control policy is:

- The optimal control policy overall, in the input-noise setting of Problem 2.
- *Not* the optimal control policy overall, in the observation-noise setting of Problem 4.

More specifically, we evaluate the following three classes of policies on our two settings:

- The class of linear memory-1 period-1 control policies:

$$F_t(Y_{(t)};\theta) = \theta_0 Y_t. \tag{30}$$

- The class of affine memory-2 period-1 control policies:

$$F_t(Y_{(t)};\theta) = \begin{cases} \theta_0 + \theta_1 Y_t, & t = 0 \\ \theta_0 + \theta_1 Y_t + \theta_2 Y_{t-1}, & t \geq 1 \end{cases} \tag{31}$$

- The class of affine memory-1 period-2 control policies:

$$F_t(Y_{(t)};\theta) = \begin{cases} \theta_0 + \theta_1 Y_t, & t \text{ is even} \\ \theta_2 + \theta_3 Y_t, & t \text{ is odd} \end{cases} \tag{32}$$

(b) Change **driver.py** to train each policy on a control noise system with $b = 1$ and $\beta = 1$ and visualize the results. Which policies tend to do well? What level of a can each policy stabilize? Include the provided visualizations.

Solution: The policy that tends to do the best when $\beta = 1, b = 1, \gamma = 0, c = 0$ is the Linear Memory-1, Period-1 policy. The graph below was generated for $a = 0, 0.9, 1.0, 1.4, 2, 3.0$. As we can see, that means that it can stabilize almost any value of a.

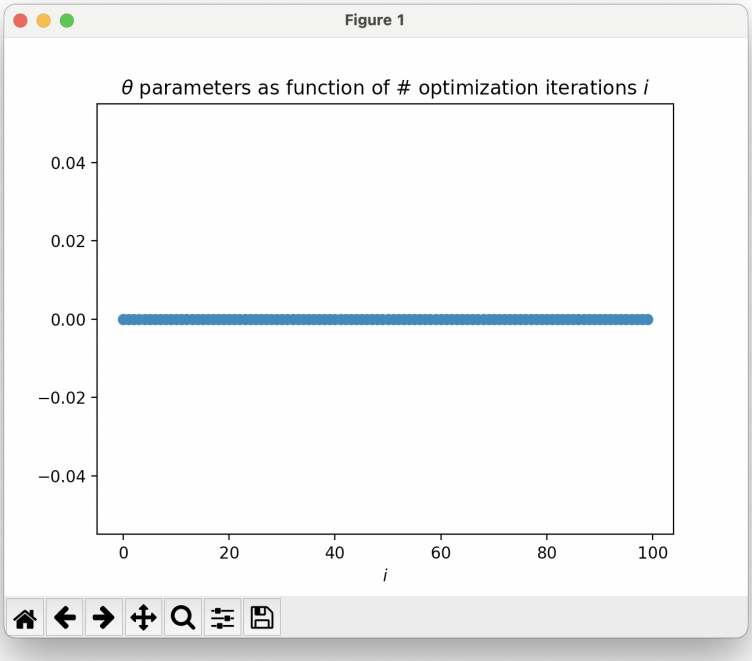

Figure 1: Learnable Weight Memory-1, Period-1 Linear Module: $b = 1, \beta = 1, c = 0, \gamma = 0$

The graphs below were generated for the Learnable Weight Memory-1, Period-2 Affine Module, for $a = 0, 0.9, 1.0, 1.4, 2, 3.0$. They can stabilize values of $a \geq \sqrt{2}$.

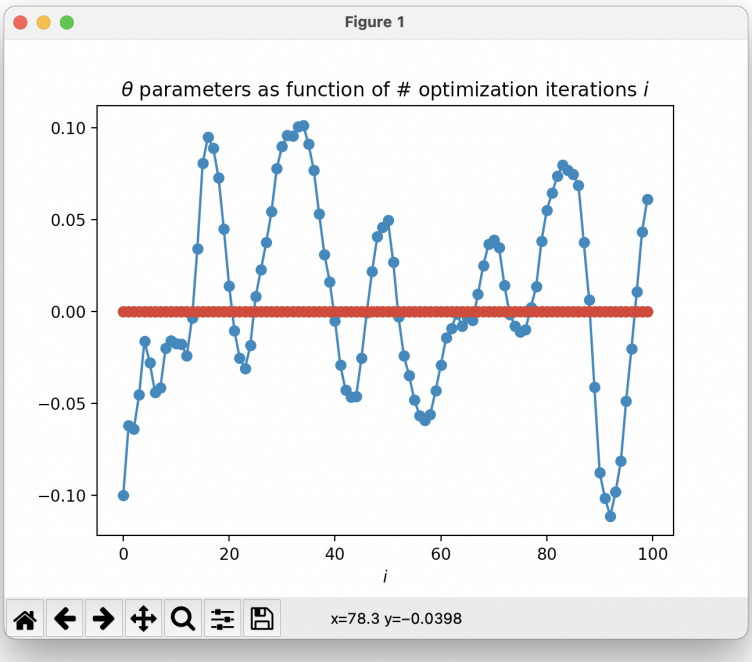

Figure 2: Learnable Weight Memory-1, Period-2 Affine Module: $a = 0, b = 1, \beta = 1, c = 0, \gamma = 0$

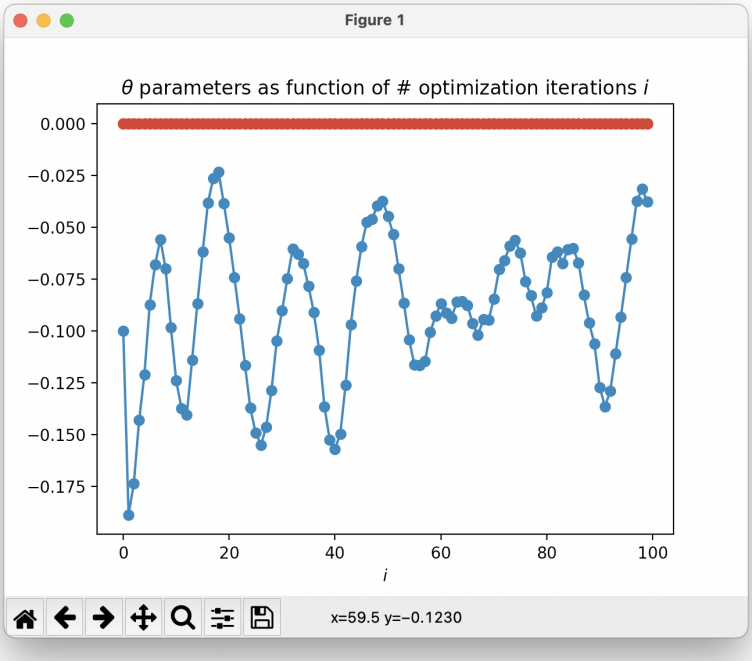

Figure 3: Learnable Weight Memory-1, Period-2 Affine Module: $a = 0.9, b = 1, \beta = 1, c = 0, \gamma = 0$

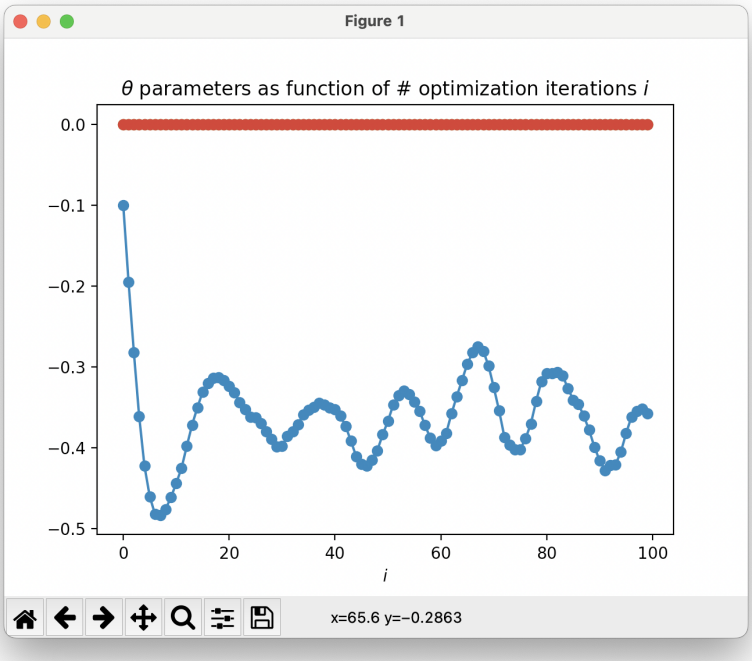

Figure 4: Learnable Weight Memory-1, Period-2 Affine Module: $a = 1.4, b = 1, \beta = 1, c = 0, \gamma = 0$

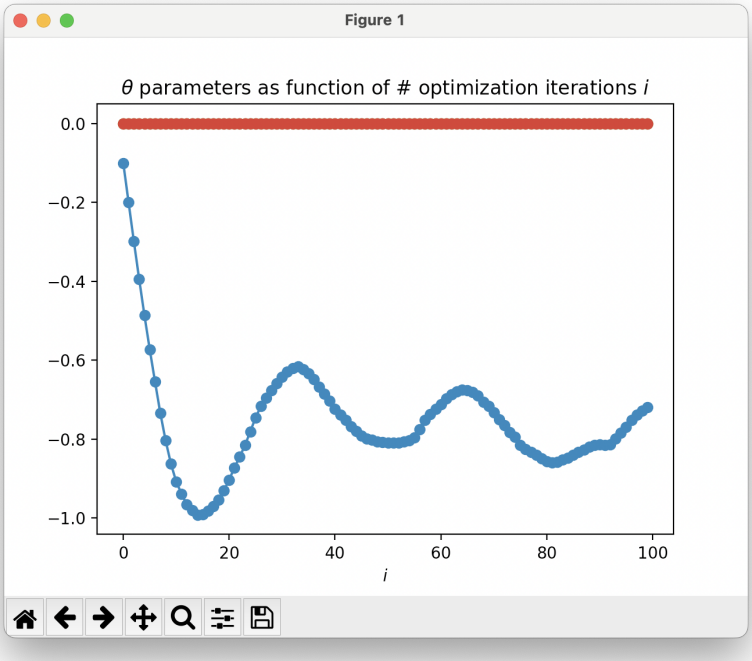

Figure 5: Learnable Weight Memory-1, Period-2 Affine Module: $a = 2, b = 1, \beta = 1, c = 0, \gamma = 0$

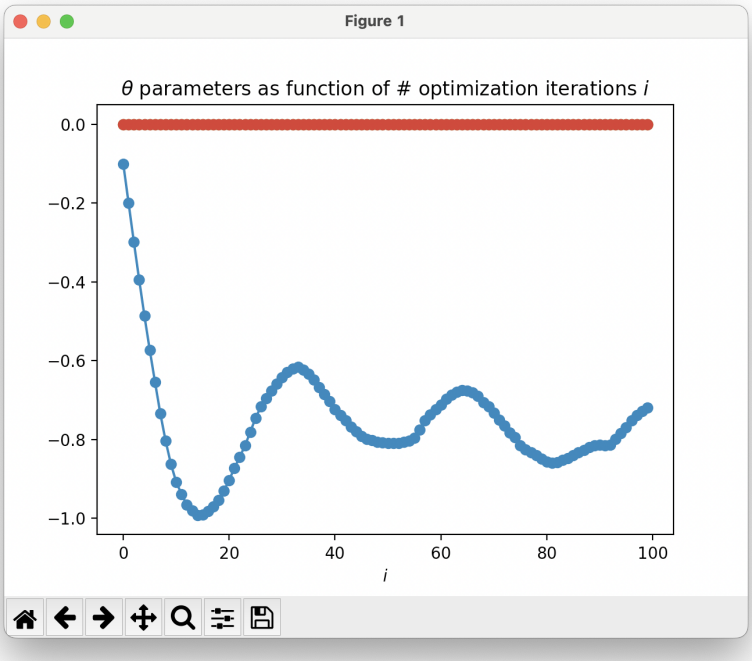

Figure 6: Learnable Weight Memory-1, Period-2 Affine Module: $a = 3, b = 1, \beta = 1, c = 0, \gamma = 0$

The graphs below were generated for the Learnable Weight Memory-2, Period-1 Affine Module, for $a = 0, 1.0, 1.4, 2, 3$. This seems to converge for $a \leq \sqrt{2}$, which we can verify visually.

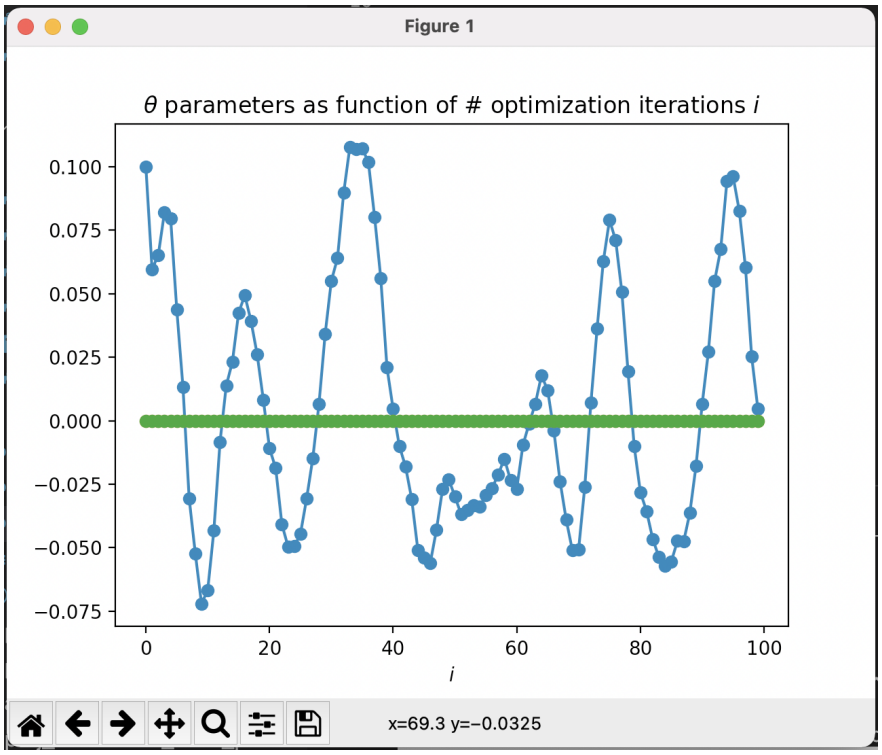

Figure 7: Learnable Weight Memory-2, Period-1 Affine Module: $a = 0, b = 1, \beta = 1, c = 0, \gamma = 0$

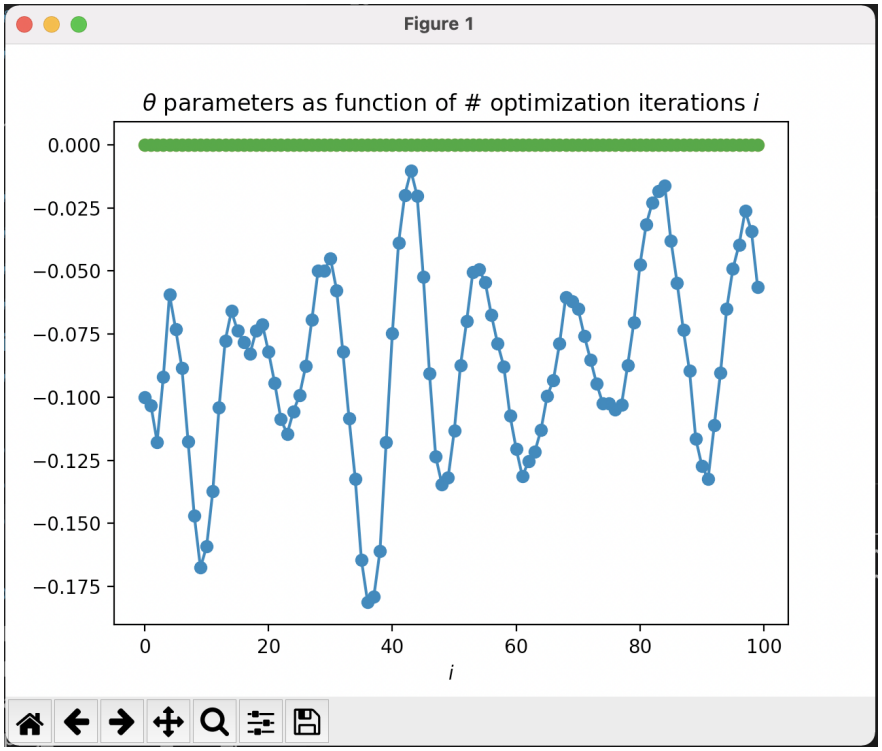

Figure 8: Learnable Weight Memory-2, Period-1 Affine Module: $a = 0.9, b = 1, \beta = 1, c = 0, \gamma = 0$

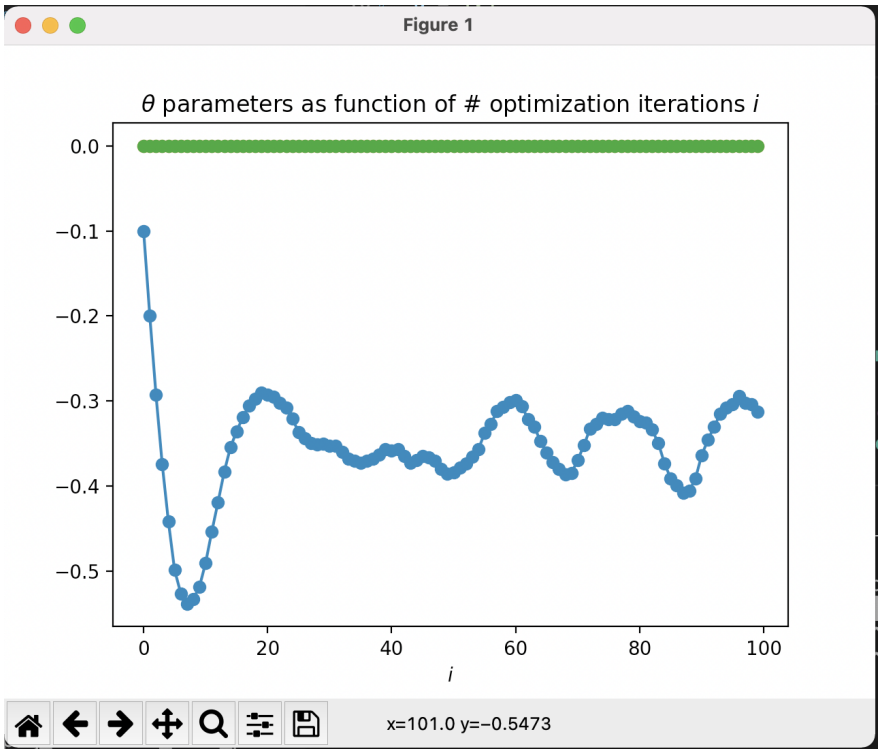

Figure 9: Learnable Weight Memory-2, Period-1 Affine Module: $a = 1.4, b = 1, \beta = 1, c = 0, \gamma = 0$

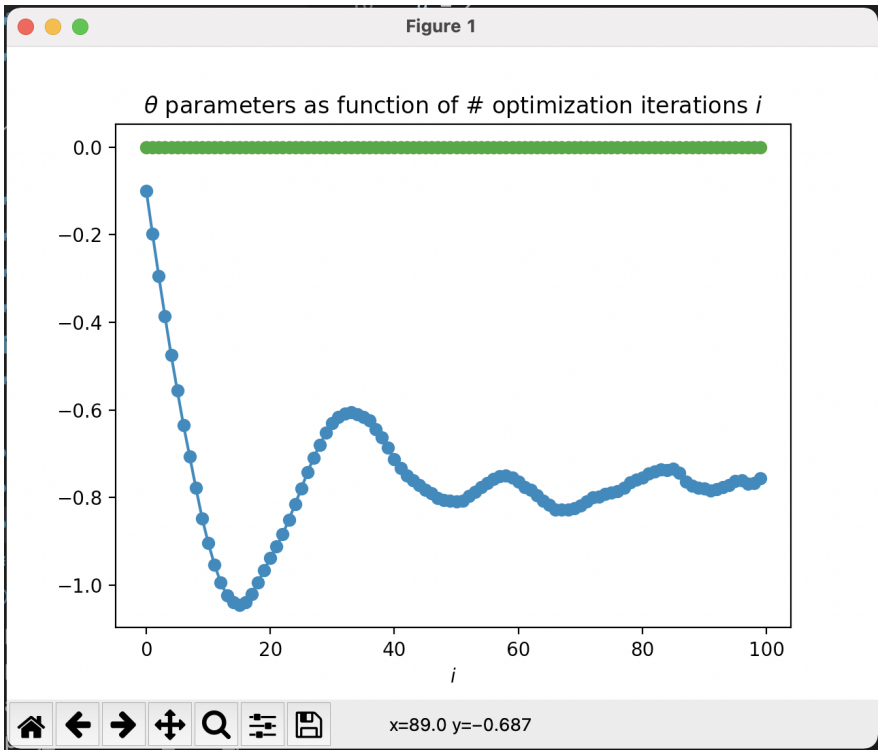

Figure 10: Learnable Weight Memory-2, Period-1 Affine Module: $a = 2, b = 1, \beta = 1, c = 0, \gamma = 0$

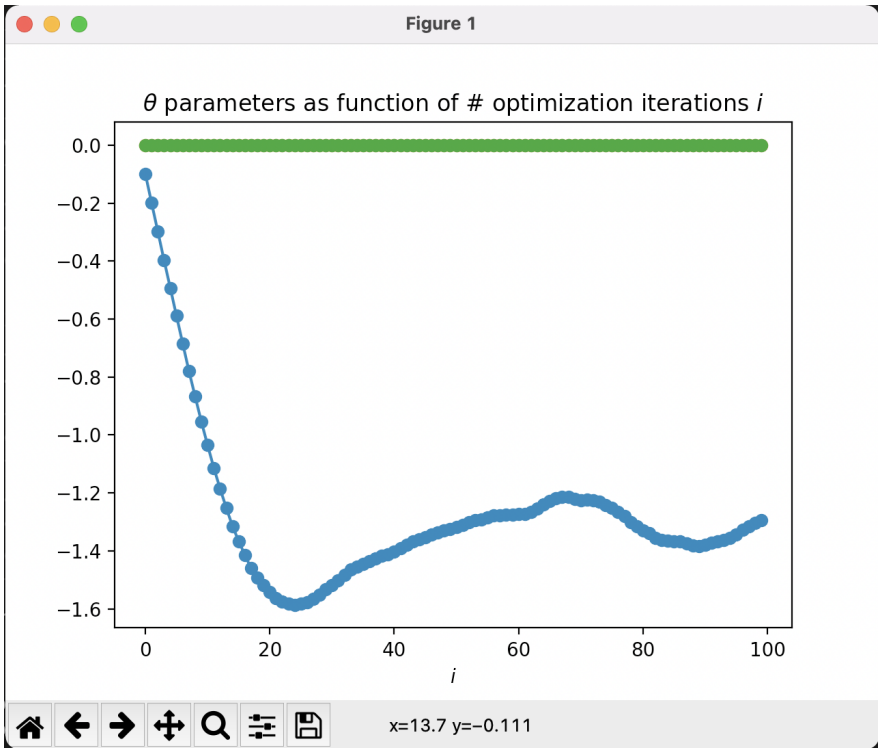

Figure 11: Learnable Weight Memory-2, Period-1 Affine Module: $a = 3, b = 1, \beta = 1, c = 0, \gamma = 0$

(c) Change **driver.py** to train each policy on an observation noise system with $c = 1$ and $\gamma = 1$. Which policies tend to do well? Why? What level of a can each policy stabilize? Include the provided visualizations.

Solution: The graphs below were generated for the Learnable Weight Memory-1, Period-1 Linear Module, for $a = 0, 0.9, 1.4, 2, 3.0$.

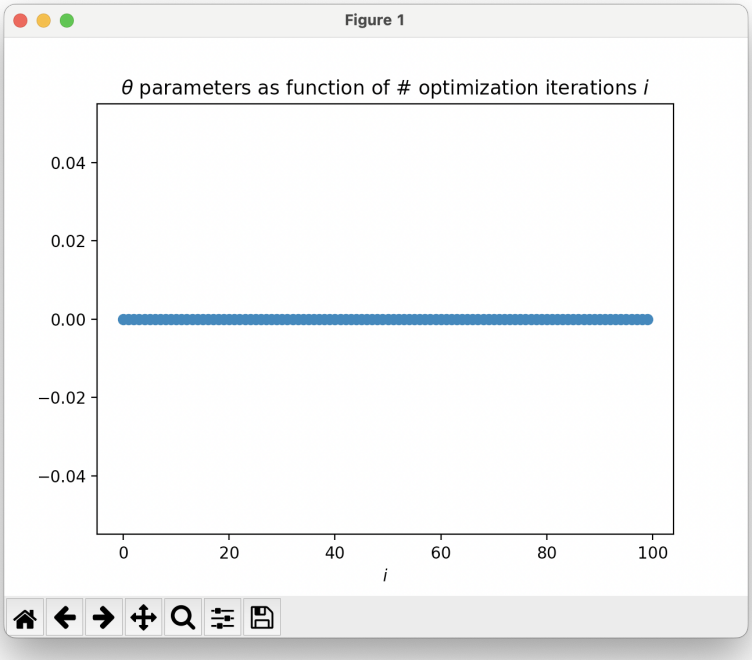

Figure 12: Learnable Weight Memory-1, Period-1 Linear Module: $a = 0, b = 0, \beta = 0, c = 1, \gamma = 1$

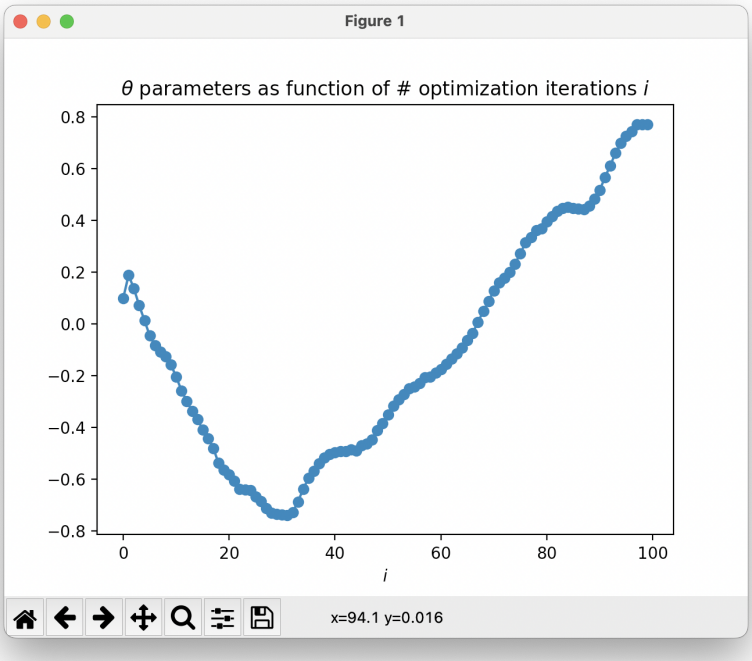

Figure 13: Learnable Weight Memory-1, Period-1 Linear Module: $a = 0.9, b = 0, \beta = 0, c = 1, \gamma = 1$

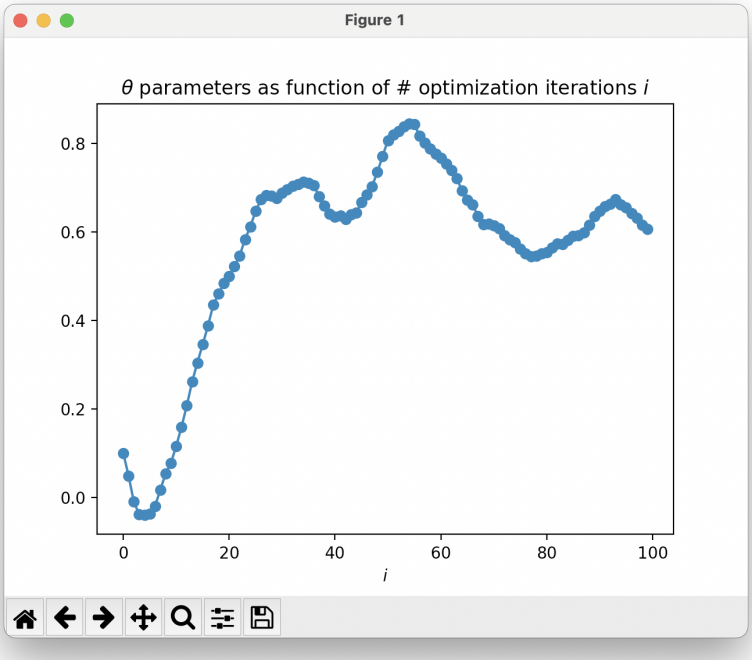

Figure 14: Learnable Weight Memory-1, Period-1 Linear Module: $a = 1.4, b = 0, \beta = 0, c = 1, \gamma = 1$

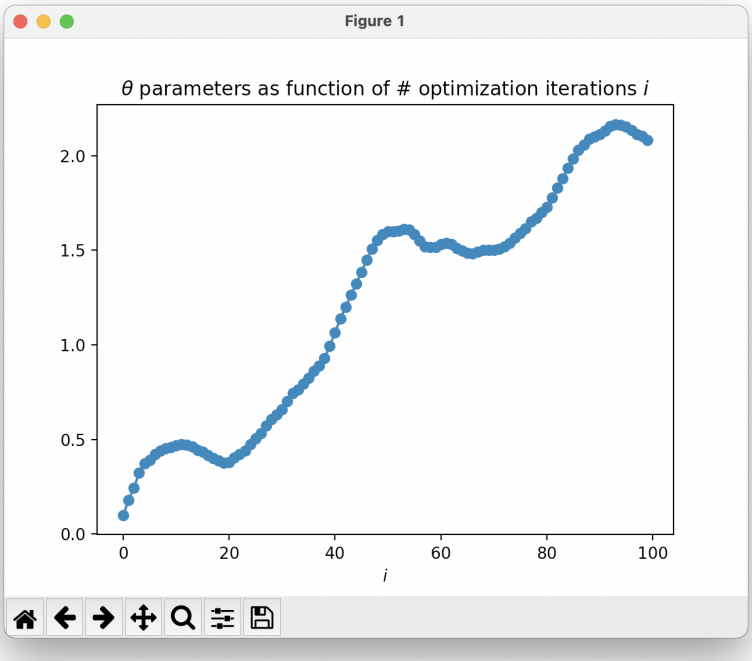

Figure 15: Learnable Weight Memory-1, Period-1 Linear Module: $a = 2, b = 0, \beta = 0, c = 1, \gamma = 1$

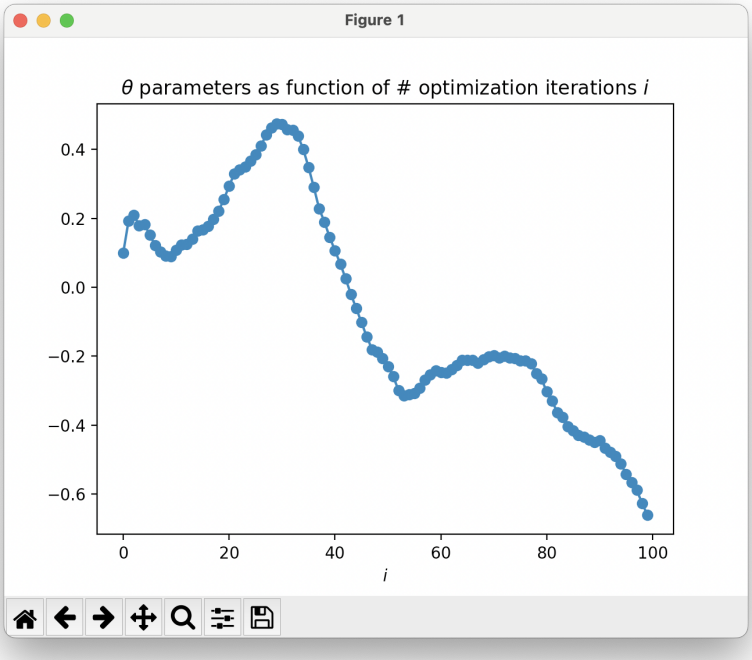

Figure 16: Learnable Weight Memory-1, Period-1 Linear Module: $a = 3, b = 0, \beta = 0, c = 1, \gamma = 1$

The graphs below were generated for the Learnable Weight Memory-1, Period-2 Affine Module, for $a = 0, 1.0, 1.4, 2, 3.0$. It is stabilizable for $a \leq \sqrt{2}$. We can visually see this from the graphs; the $a = 2$ and $a = 3$ values diverge.

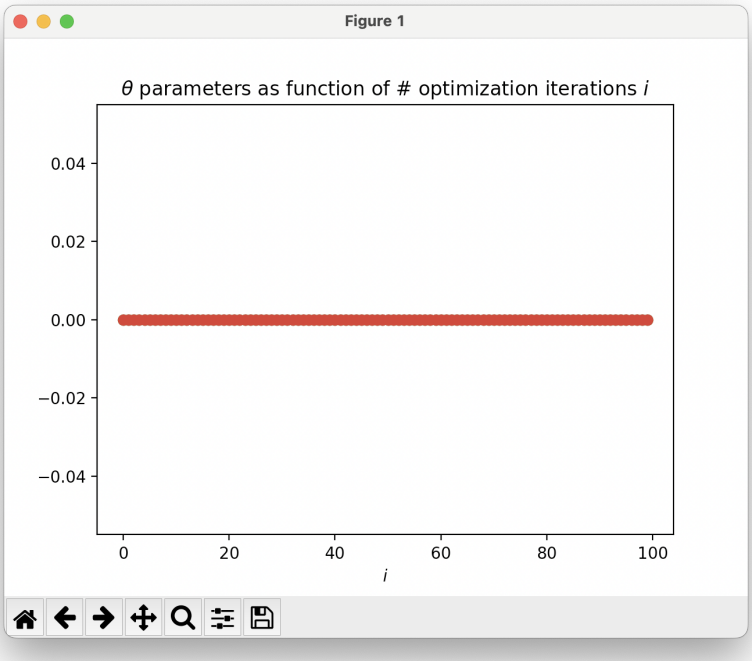

Figure 17: Learnable Weight Memory-1, Period-2 Affine Module: $a = 0, b = 0, \beta = 0, c = 1, \gamma = 1$

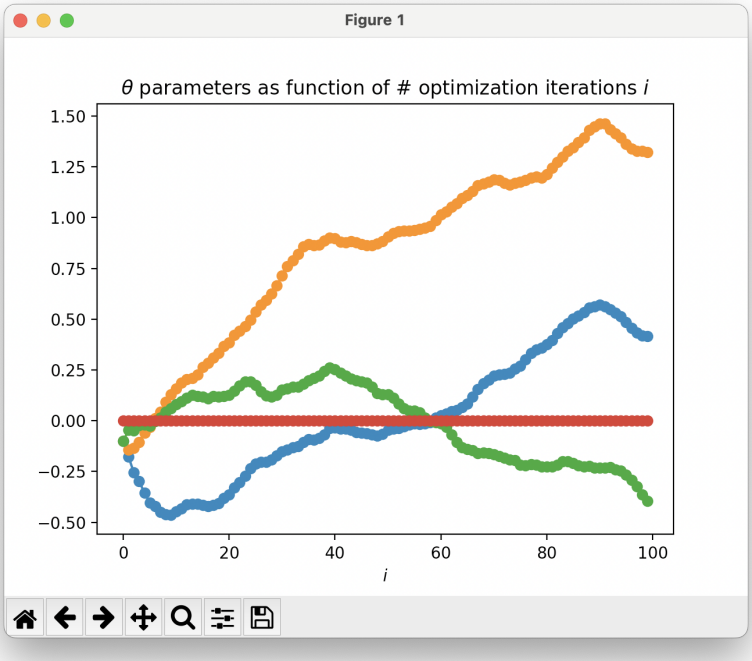

Figure 18: Learnable Weight Memory-1, Period-2 Affine Module: $a = 0.9, b = 0, \beta = 0, c = 1, \gamma = 1$

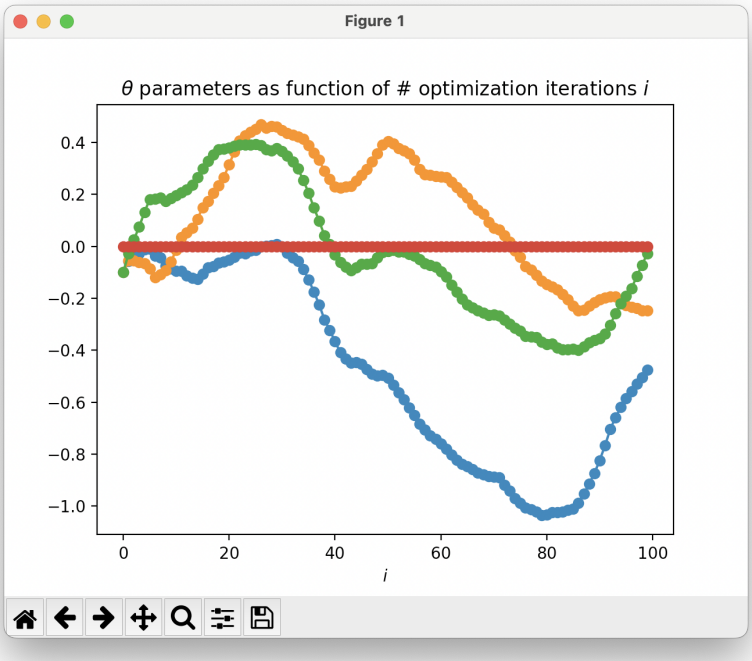

Figure 19: Learnable Weight Memory-1, Period-2 Affine Module: $a = 1.4, b = 0, \beta = 0, c = 1, \gamma = 1$

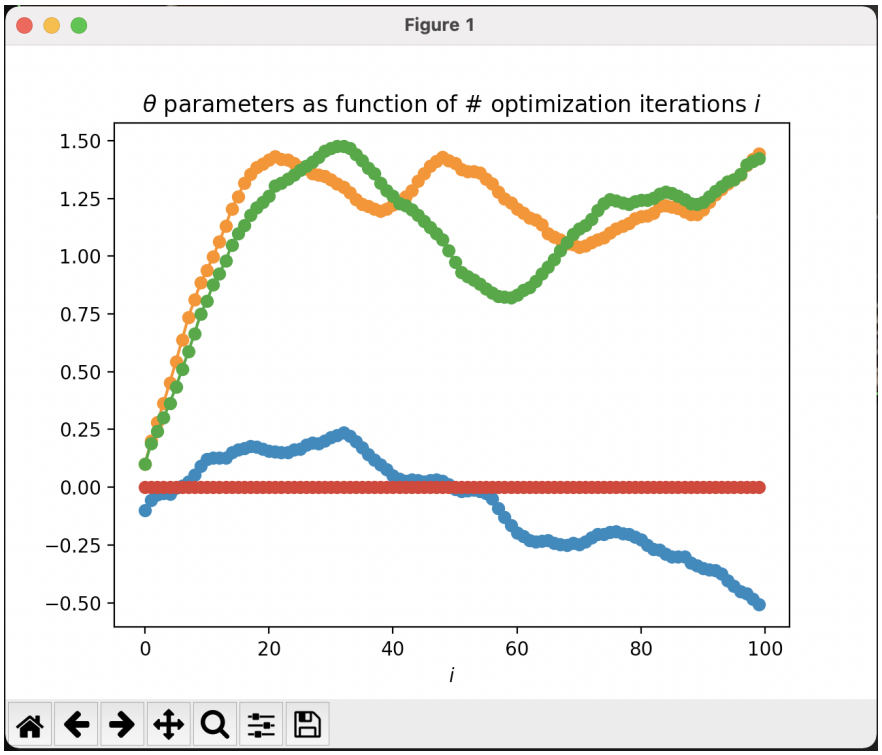

Figure 20: Learnable Weight Memory-1, Period-2 Affine Module: $a = 2, b = 0, \beta = 0, c = 1, \gamma = 1$

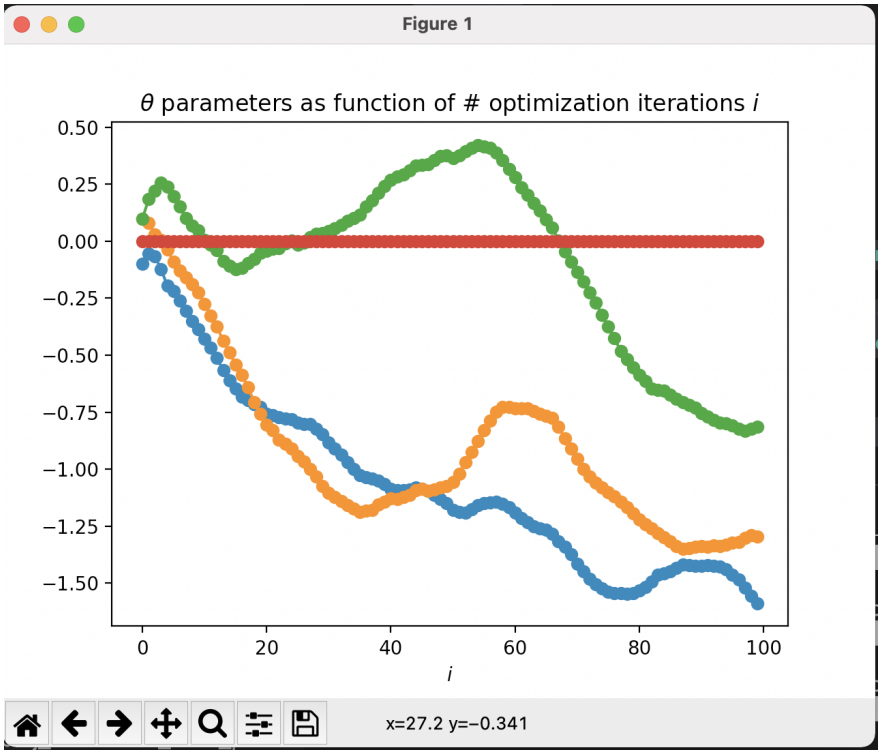

Figure 21: Learnable Weight Memory-1, Period-2 Affine Module: $a = 3, b = 0, \beta = 0, c = 1, \gamma = 1$

The graphs below were generated for the Learnable Weight Memory-2, Period-1 Affine Module, for $a = 0, 0.9, 1.4, 2, 3.0$. It generally doesn't converge for $a \neq 0$.

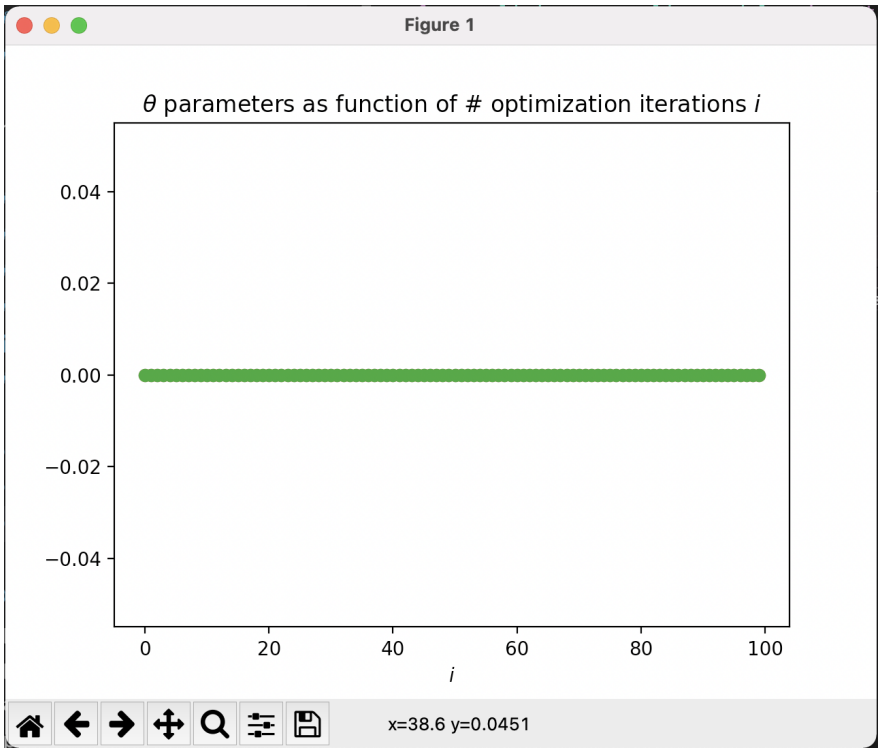

Figure 22: Learnable Weight Memory-2, Period-1 Affine Module: $a = 0, b = 0, \beta = 0, c = 1, \gamma = 1$

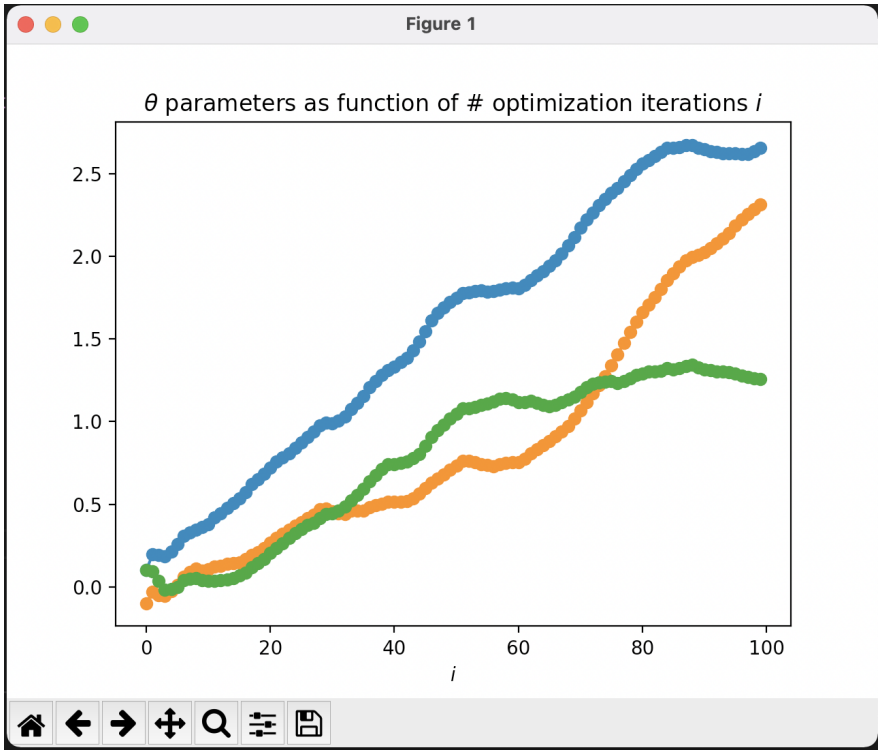

Figure 23: Learnable Weight Memory-2, Period-1 Affine Module: $a = 0.9, b = 0, \beta = 0, c = 1, \gamma = 1$

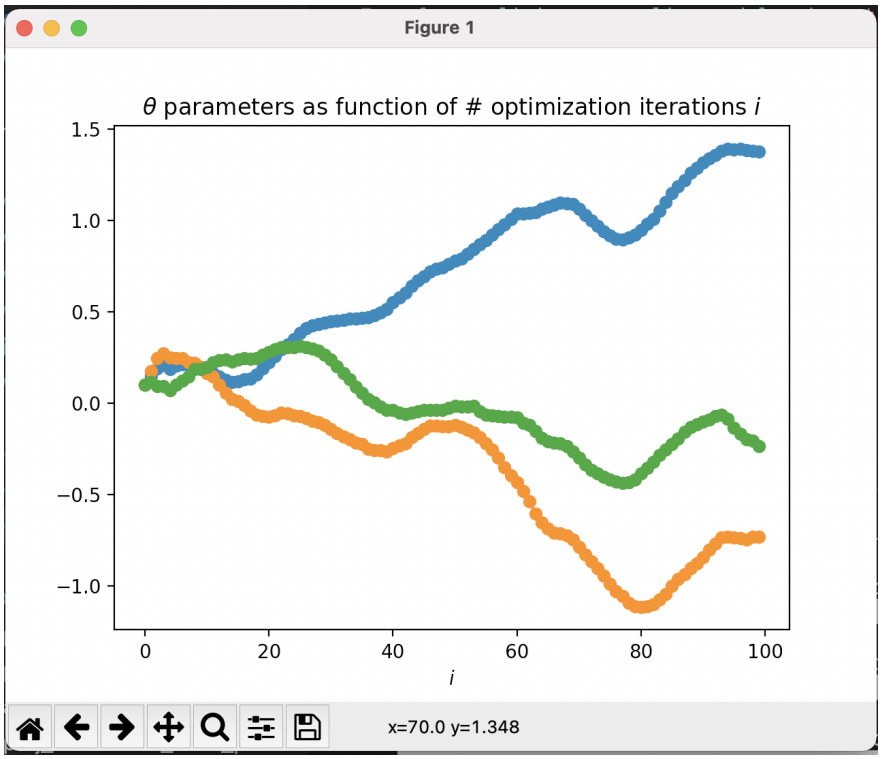

Figure 24: Learnable Weight Memory-2, Period-1 Affine Module: $a = 1.4, b = 0, \beta = 0, c = 1, \gamma = 1$

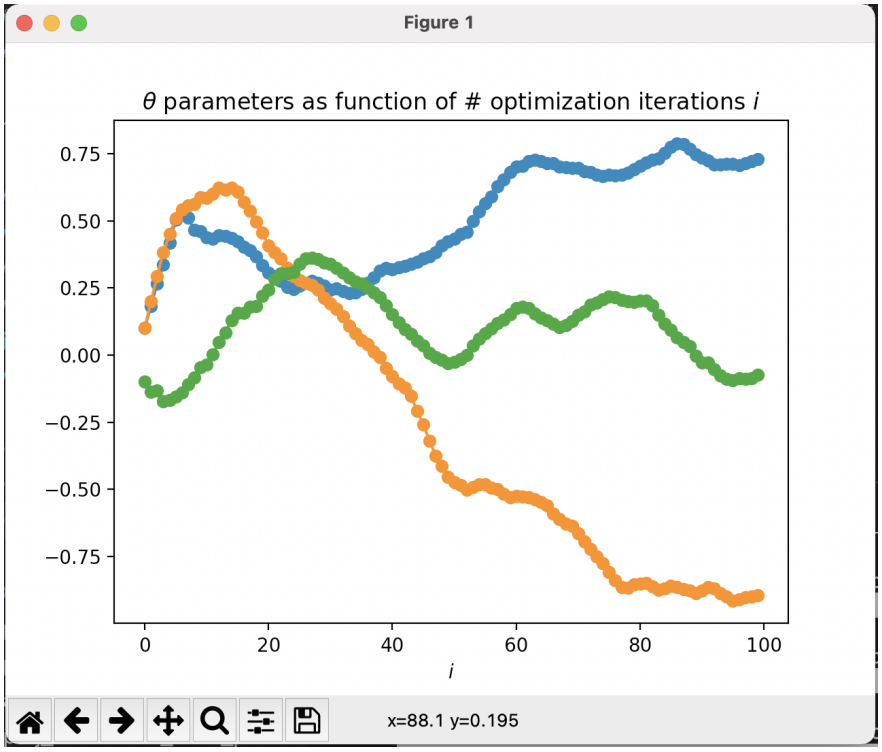

Figure 25: Learnable Weight Memory-2, Period-1 Affine Module: $a = 2.0, b = 0, \beta = 0, c = 1, \gamma = 1$

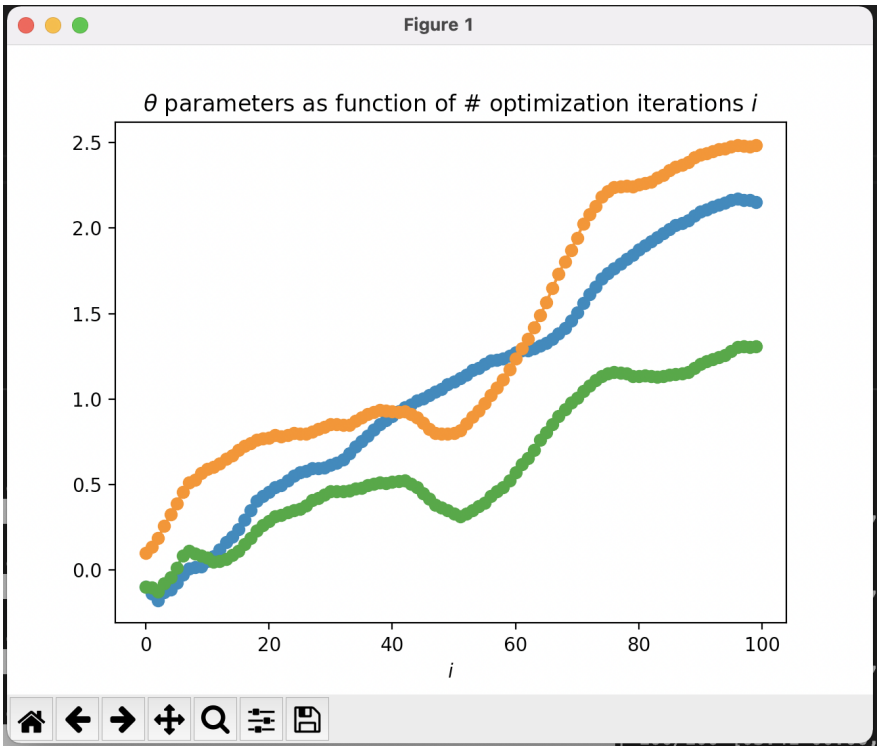

Figure 26: Learnable Weight Memory-2, Period-1 Affine Module: $a = 3.0, b = 0, \beta = 0, c = 1, \gamma = 1$

7 Literature Review

In this review, we examine Subramanian et al's paper "Learning a Neural-Network Controller for a Multiplicative Observation Noise System" and their approach to controlling multiplicative noise. In the paper, the authors propose the use of a periodic controller structure and a greedy training procedure to stabilize a linear control system that has multiplicative observation noise. Consider a linear control system:

$$X_{n+1} = aX_n - U_n \tag{33}$$

$$Y_n = Z_n X_n \tag{34}$$

where X_n is the state, U_n is the control, Y_n is the observation, and Z_n represents multiplicative noise ($Z_n \sim \mathcal{N}(0,1)$) that has corrupted the observation. Linear strategies have proven to be ineffective in terms of stabilizing this control system, and non-linear strategies consistently outperform linear strategies. However, with these hand-crafted non-linear strategies comes increased complexity making them sub-optimal. Therefore, a large gap remains between achieveability and performance.

It is important to note that previous work has shown that it is better to use memory, M, at time-step (n) to generate the control U_n. This means that we consider the previous observations $Y_{n-1}, Y_{n-2}, \ldots, Y_{n-M+1}$ to generate the control U_n. Thus, this paper uses the idea of memory in addition to neural networks to design controllers for the control problem presented above. More specifically, Subramanian et al seek to reduce the gap between the known bounds of a , as presented by Ding et al "When multiplicative noise stymies control", for which it is possible to stabilize this linear system in a second-moment sense. This is related to our work in Problems 2, 3, 4 of this project. Next, the system was allowed to cycle through different neural networks as controllers periodically, which is denoted by the parameter P. Lastly, the length of each of the training stages is denoted by parameter G. Specifically, the second moment of the true state is greedily minimized at the end of each training stage. Overall, the goal is second moment stability given that $a \in \mathbb{R}^+$ is fixed and known.

Then, the paper discusses the architecture and training procedure of the neural networks and how appropriately choosing these components allows for the opportunity of learning a neural network based control strategy that outperforms the complex, hand-crafted control strategies we mentioned before, and stabilizes the multiplicative observation noise control system.

Subramanian et al also demonstrate that performance improves when allowing control strategies to use more memory, however, after a while diminishing returns sets in. The paper showed that neural network based approaches give higher values of a^* for which the second moment stability is possible. The best results from prior work denoted by PBS achieved an a^* value of 1.032 while strategies such as M1-P2-G4 achieved 1.026, M2-P2-G2 achieved 1.097 and M4-P4-G4 achieved 1.156.

Generally, we see that as M, P and G increase, performance generally increases. The paper also considers the interplay of these parameters, states that diminishing returns is seen from increasing their values but states that future work will focus on the open problem of exact quantification of how memory and period affect performance and their optimality.