I. Numerical Examples

Example 1: We consider a one-dimensional state x_n and observation y_n satisfying:

$$\begin{cases} x_{n+1} = x_n + \eta \cos(3x_n\alpha_n) + \sqrt{\eta}\sigma w_n, \\ y_{n+1} = y_n + 2\eta x_n + \sqrt{\eta}\sigma_1 v_n, \end{cases} \tag{1}$$

where $\sigma = 0.7$, $\sigma_1 = 0.2$, and $x_0 = y_0 = 0$. We plot the training loss function with both constant LR (denoted by Const lr) and adaptive LR (denoted by EpochAda lr) with constant initials in Figure 1.

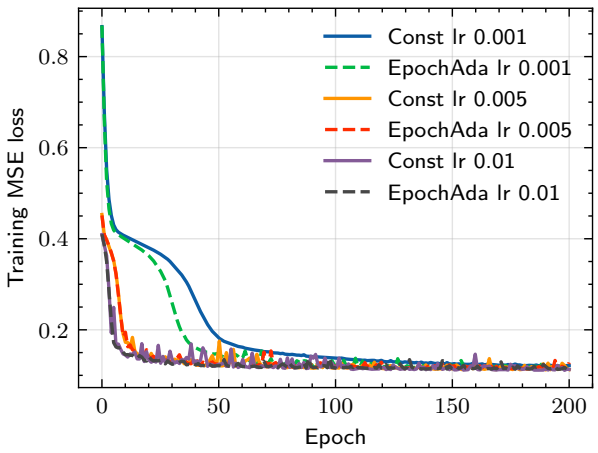

Fig. 1. Example 1: MSE loss functions

The loss graph for constant learning rate is plotted by the solid line, and that of adaptive learning rate is given by dashed one. From Figure 1, it demonstrated that the loss function with adaptive learning rate with initial $\rho_0 = 0.001$ decreases much faster after around 20 epochs compared to that with constant learning rate. For initials $\rho_0 = 0.005$ and 0.01, such difference are less pronounced. The path of adaptive learning rates is shown in Figure 2 with initials $\rho_0 = 0.001, 0.005$, and 0.01, respectively.

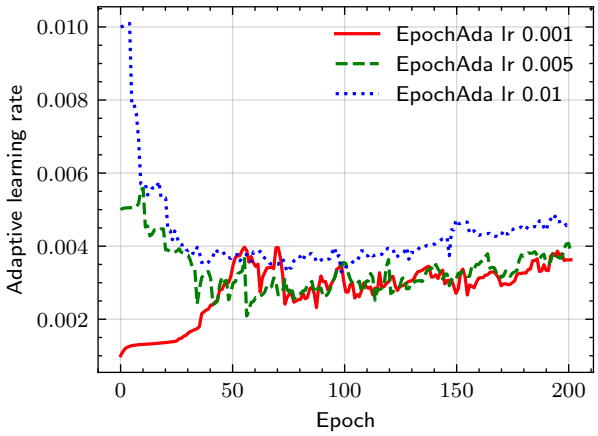

Fig. 2. Example 1: Adaptive learning rates with initials $0.001, 0.005$, and 0.01.

It shows that the adaptive learning rate appears to converge to around 0.004. With initial learning rate 0.004, Table I exhibits the corresponding relative errors of the out-of-sample filtering results.

TABLE I
Example 1: Relative errors of the state x_n and the deep filtering results.

ρ_0	0.001	0.005	0.01
EpochAda LR	0.1612	0.1721	0.1629

From the table, it can be seen that the choice of initial learning rate has little impact on the relative error, which indicates its robustness. Finally, Figure 3 plots a sample path of the state x_n and the corresponding paths of deep filters with constant LR and adaptive LR.

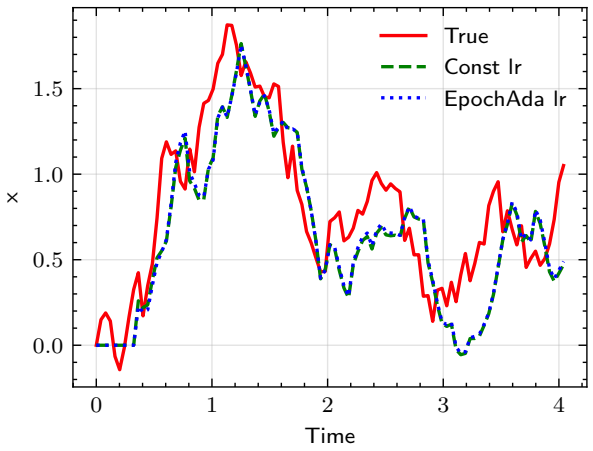

Fig. 3. Example 1: Sample paths of the state and deep filtering with constant LR $\rho = 0.004$ and the adaptive LR with initial $\rho_0 = 0.004$.

Example 2: Consider a two-dimensional linear in (x, y) model:

$$\begin{cases} x_{n+1} = x_n + \eta F(\alpha_n)x_n + \sqrt{\eta}\sigma(\alpha_n)w_n, \\ y_{n+1} = y_n + \eta G(\alpha_n)x_n + \sqrt{\eta}\sigma_1 v_n, \end{cases} \tag{2}$$

where the Markov chain $\alpha_n \in \{1, 2\}$, the initial condition $x_0 = (1, -1)'$, and

$$F(1) = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \sigma(1) = \begin{bmatrix} 1 & -0.1 \\ 0 & 1 \end{bmatrix}, G(1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$F(2) = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}, \sigma(2) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, G(2) = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix},$$

and the observation noise matrix

$$\sigma_1 = \begin{bmatrix} 0.2 & 0.05 \\ 0 & 0.2 \end{bmatrix}.$$

The plot of the training loss function with both constant LR and adaptive LR with constant initials was exhibited in Figure 4. It indicates that if we choose initial learning rate $\rho_0 = 0.001$, the loss function with adaptive learning rate strategy decreases much faster after around 20 epochs.

For initials $\rho_0 = 0.005$ and 0.01, such differences are less pronounced.

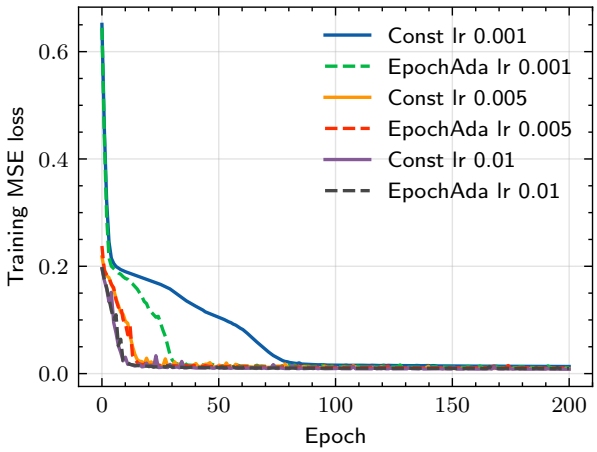

Fig. 4. Example 2: MSE loss functions

The evolution of adaptive learning rate can be found in Figure 5 with initials $\rho_0 = 0.001, 0.005$, and 0.01, respectively. It can be seen that the adaptive learning rate appears to converge to around 0.003.

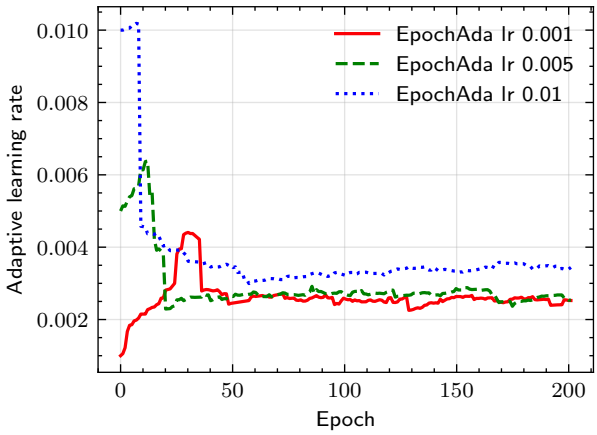

Fig. 5. Example 2: Adaptive learning rates with initials $0.001, 0.05$, and 0.01

Table II gives the corresponding relative errors of the out-of-sample filtering outcomes with initial learning rate 0.003.

TABLE II
EXAMPLE 2: RELATIVE ERRORS OF THE STATE x_n AND THE DEEP FILTERING RESULTS.

ρ_0	0.001	0.005	0.01
EpochAda LR	0.1095	0.1053	0.1038

It is seen that the choice of initial learning rate has little impact on the relative error, which is a sign of robustness. Finally, Figure 6 shows a sample path of the state x_n and the corresponding paths of deep filters with constant LR and adaptive LR.

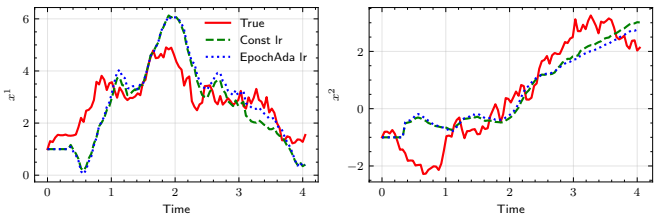

Fig. 6. Example 2: Sample paths of the state and deep filtering with constant LR $\rho = 0.003$ and the adaptive LR with initial $\rho_0 = 0.003$.

Example 3: Consider a two-dimensional nonlinear model:

$$\begin{cases} x_{n+1} = x_n + \eta \begin{bmatrix} \cos((0.1x_n^0 + 0.3x_n^1)\alpha_n) \\ \sin(0.3x_n^1\alpha_n) \end{bmatrix} + \sqrt{\eta}\sigma w_n, \\ y_{n+1} = y_n + \eta G x_n + \sqrt{\eta}\sigma_1 v_n, \end{cases} \tag{3}$$

where $x_n = (x_n^0, x_n^1)'$ and the Markov chain $\alpha_n \in \{1, 2\}$ and

$$\sigma = \begin{bmatrix} 1 & -0.3 \\ 0 & 1 \end{bmatrix}, G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0.2 & 0.05 \\ 0 & 0.2 \end{bmatrix}$$

The initial state is chosen as $x_0 = (1, -1)'$. Like in the first two examples, Figure 7 exhibits the training loss functions for constant LRs $(0.001, 0.005, 0.01)$ and adaptive LRs with initials $\rho_0 = 0.001, 0.005$, and 0.01. It illustrates similar behaviors as in the previous models.

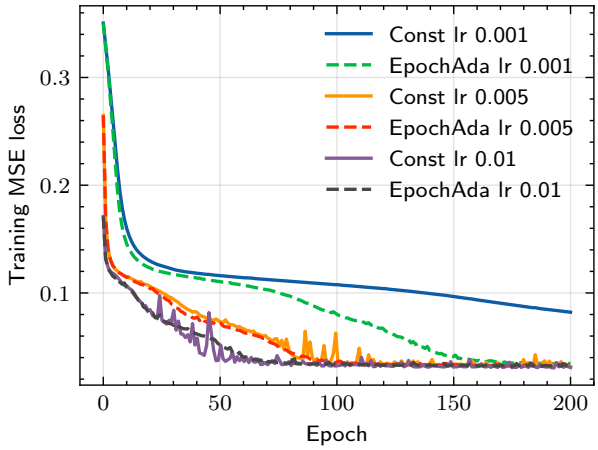

Fig. 7. Example 3: MSE loss functions

The path of adaptive LRs is also shown similar convergence in Figure 8 and appears to converge to around 0.002.

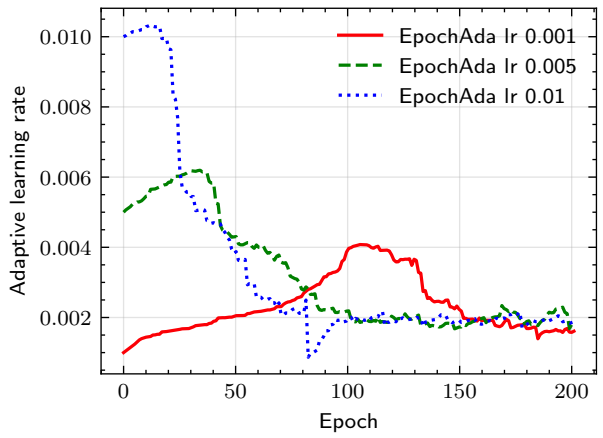

Fig. 8. Example 3: Adaptive learning rates with initials $0.001, 0.05$, and 0.01

The relative errors of the state x_n and the deep filtering results are given in Table III.

TABLE III
EXAMPLE 3: RELATIVE ERRORS OF THE STATE x_n AND THE DEEP FILTERING RESULTS.

ρ_0	0.001	0.005	0.01
EpochAda LR	0.0925	0.0.908	0.0910

Finally, a sample path of the state x_n and that of out-of-sample deep filtering with constant LR 0.002 and adaptive LR with initial $\rho_0 = 0.002$ are provided in Figure 9.

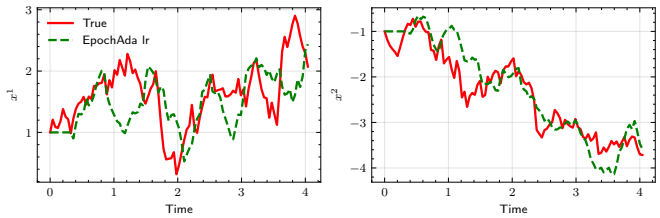

Fig. 9. Example 3: Sample paths of the state and deep filtering with constant LR $\rho = 0.002$ and the adaptive LR with initial $\rho_0 = 0.002$.

Example 4: Consider a six-dimensional nonlinear tracking model: Finally, we consider the following six-dimensional nonlinear model where the state x_n and the observation y_n satisfying

$$\begin{cases} x_{n+1} = x_n + \eta F x_n + \sqrt{\eta} w_n, \\ y_{n+1} = y_n + \eta h(x_n) + \sqrt{\eta}\sigma_1 v_n, \end{cases} \tag{4}$$

where $w_n \backsim N(0, I_6)$, $v_n \backsim N(0, I_6)$ with I_6 being a 6×6 dimensional identity matrix, the matrix F is

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -\omega_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -\omega_2^2 & 0 \end{bmatrix},$$

$\sigma_1 = 0.2$, and the function $h(x)$ is defined as

$$h(x) = [\sqrt{x_0^2 + x_3^2}, \tan^{-1}(x_3/x_0), x_1, x_2, x_4, x_5]'$$

where the state $x = (x_0, x_1, x_2, x_3, x_4, x_5)'$. For experiments, we take $\omega_1 = 0.3, \omega_2 = 0.9$ and $x_0 = (0.8, 0.2, 1, -1, 0.5, 1)'$. The training loss graph and the path of adaptive learning rate are provided in Figure 10 and Figure 11, respectively.

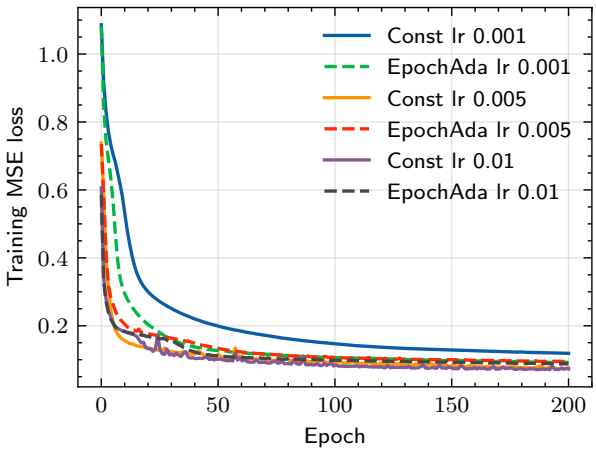

Fig. 10. Example 4: MSE loss functions

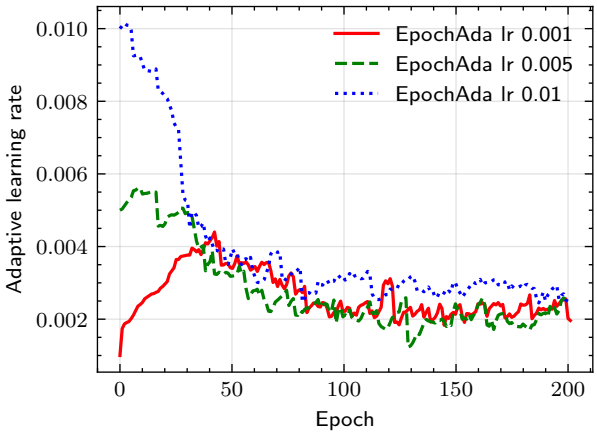

Fig. 11. Example 4: Adaptive learning rates with initials $0.001, 0.05$, and 0.01

The relative errors of deep filtering with adaptive learning rate outcomes are provided in Table IV. These relative errors indicates the robustness of deep filtering with respect to the initial selection of the adaptive LRs.

TABLE IV
EXAMPLE 4: RELATIVE ERRORS OF THE STATE x_n AND THE DEEP FILTERING RESULTS.

ρ_0	0.001	0.005	0.01
EpochAda LR	0.1822	0.1848	0.1808

As shown in Figure 11, the adaptive learning rate converges to around 0.0025. We present a sample path of the state

x_n and the corresponding out-of-sample deep filtering results with constant LR 0.0025 and adaptive LR $\rho_0 = 0.0025$ in Figure 12.

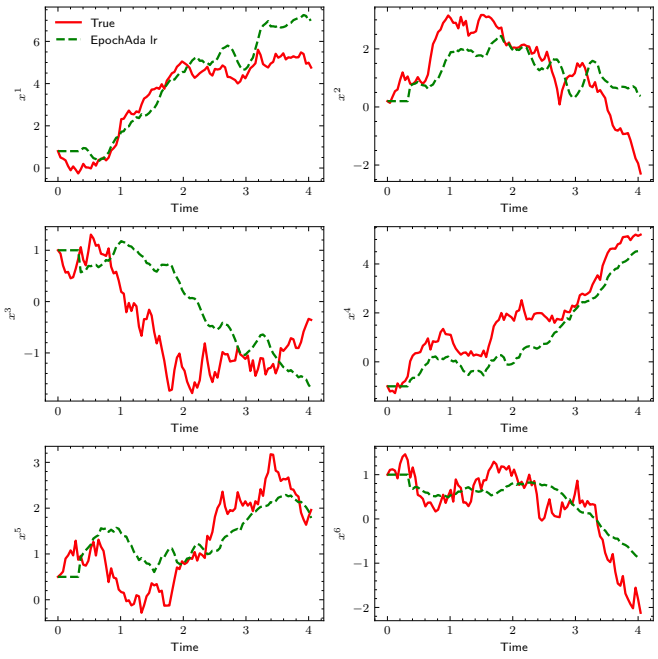

Fig. 12. Example 4: Sample paths of the state and deep filtering with constant LR $\rho = 0.0025$ and the adaptive LR with initial $\rho_0 = 0.0025$.