Lecture 14. Online Adaptation

Parameter Convergence
Uncertainty Estimation

Online Adaptation

- Generalization of RLS
- Parameter convergence
- Uncertainty estimation

Recursive Least Square: Summary

$$y_{k+1} = f_{\theta_k}(x_k)$$

Parameter update law $\theta_k = \theta_{k-1} + [H_k]^{-1} G_{k-1}^T e_k$

A priori prediction error $e_k = y_k - \hat{y}_k \qquad \hat{y}_k = f_{\theta_{k-1}}(x_{k-1})$

Gradient $G_{k-1} = \nabla_{\theta|\theta_{k-1}} f_\theta(x_{k-1})$

Hessian $H_k = \nabla^2_{\theta|\theta_k} L_k \qquad H_k = G_{k-1}^T G_{k-1} + \lambda H_{k-1}$

*** Note the similarity to Newton's update rule in optimization**

General Learning Gain

$$H_k = \lambda H_{k-1} + G_{k-1}^T G_{k-1}$$

$$H_k = \lambda_{1,k-1} H_{k-1} + \lambda_{2,k-1} G_{k-1}^T G_{k-1}$$

$\lambda_{1,k}$	$\lambda_{2,k}$	Learning rate
1	0	constant
1	1	Least square
< 1	1	Least square with forgetting factor

Stochastic Gradient Descent

- The constant learning rate case corresponds to stochastic gradient descent (SGD).

$$\theta_k = \theta_{k-1} - F G_{k-1}^T (f_{\theta_{k-1}}(x_{k-1}) - y_k)$$

This is the gradient of $\frac{1}{2}\|y_k - f_{\theta_{k-1}}(x_{k-1}))\|^2$

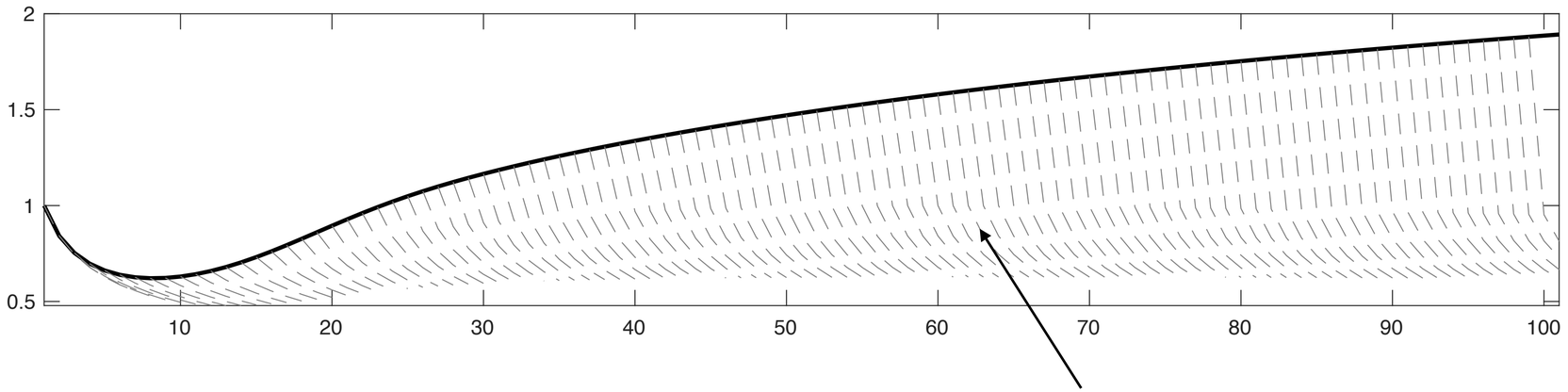
- Optimization problem: $J(\theta) = \sum_k J_k(\theta)$
- Batch optimization: $\theta^i = \theta^{i-1} - F\nabla_\theta J(\theta) = \theta^{i-1} - F\sum_k \nabla_\theta J_k(\theta)$
- Stochastic gradient descent: $\theta_k = \theta_{k-1} - F\nabla_\theta J_{k-1}(\theta)$

SGD vs RLS

- Both are optimization algorithms that can be applied to either offline learning or online adaptation.
- RLS is more suitable for online adaptation due to the adaptive learning rate, which results in faster convergence as well as more robust learning.

Example Revisited

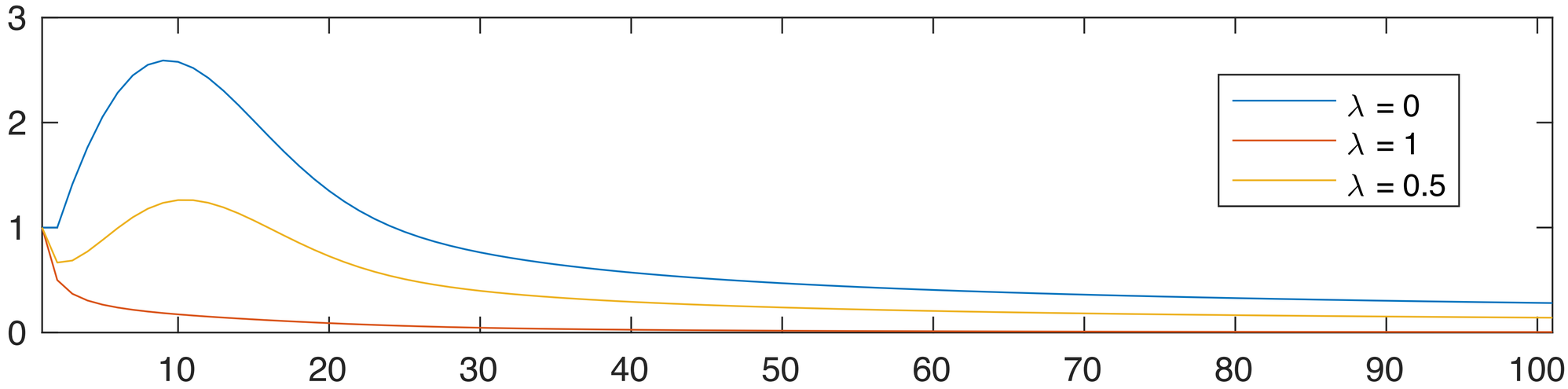
- Dynamics $x_{k+1} = a(k)\sin(x_k)$
- Time varying component $a(k) = 0.99 + 0.01k$



https://github.com/changliuliu/AdaptablePrediction

The Learning Rate in RLS

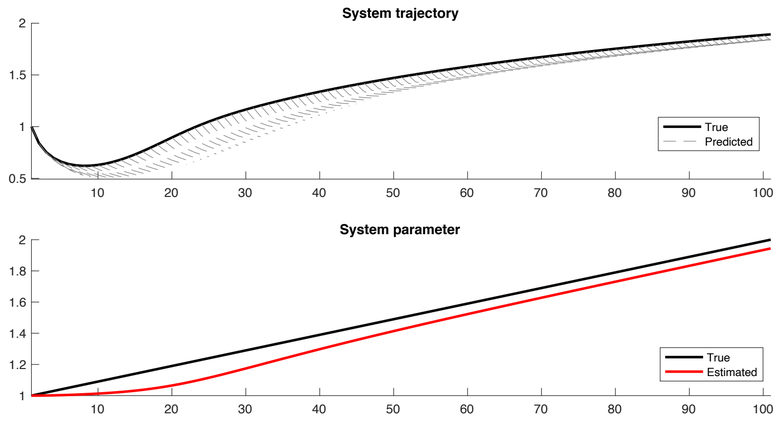
- Different forgetting factor corresponds to different learning rate
- In this example, the learning rates converge. It does not always converge.



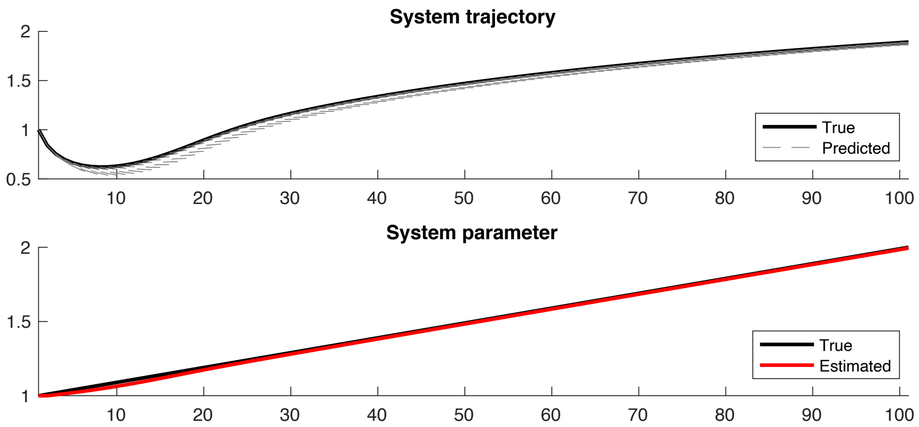
The Learning Rate in SGD

- Increasing the learning rate can improve the prediction accuracy.
- It can be unstable!

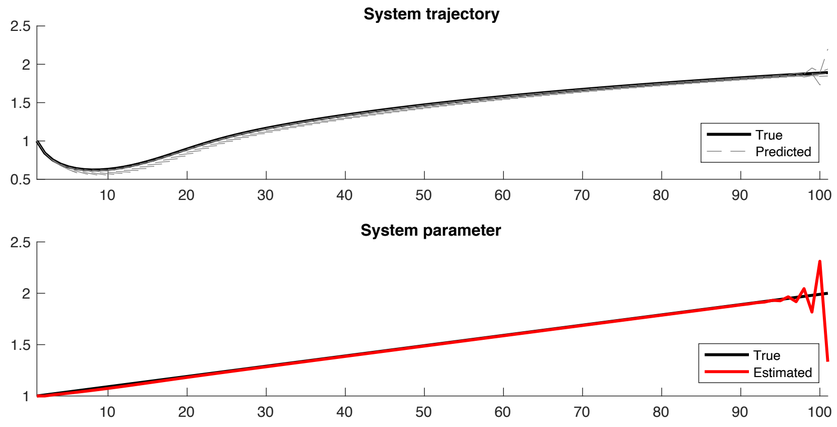
Learning rate 0.1



Learning rate 1

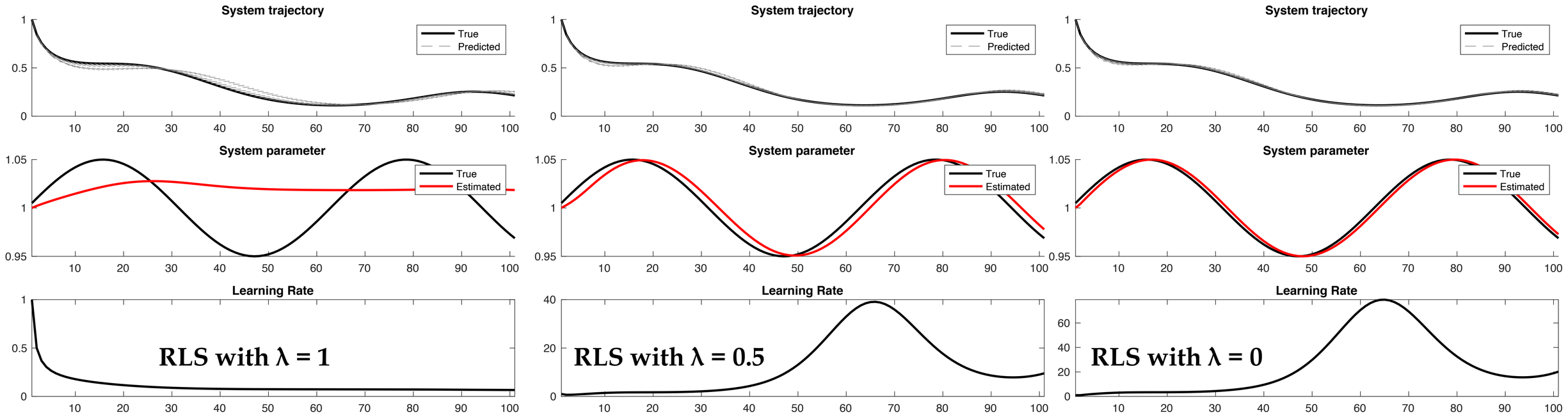


Learning rate 1.7



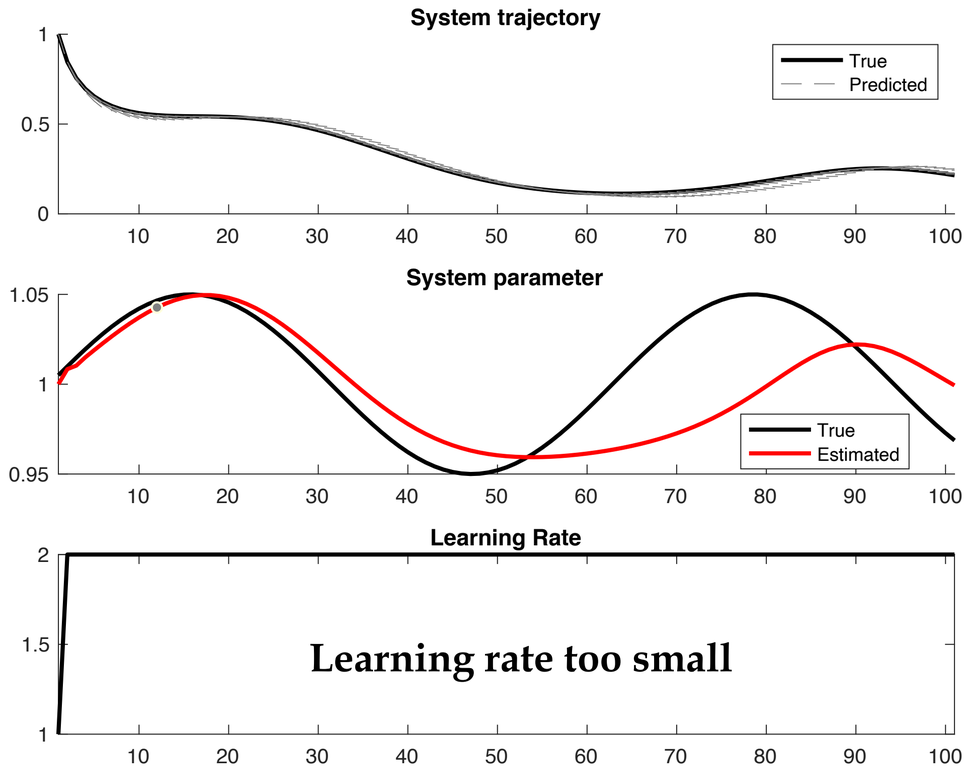
More examples with RLS

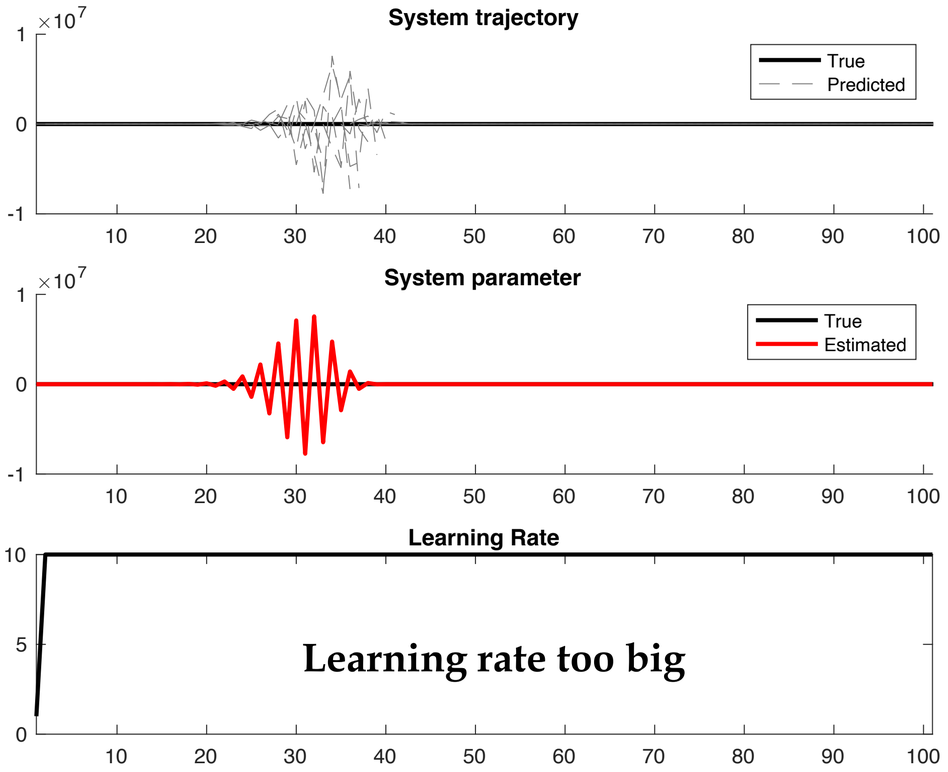
- Dynamics $x_{k+1} = a(k)\sin(x_k)$
- Time varying component $a(k) = 1 + 0.05\sin(t/10)$



More examples with SGD

- Dynamics $x_{k+1} = a(k)\sin(x_k)$
- Time varying component $a(k) = 1 + 0.05\sin(t/10)$





More on Learning Rate

- There will always be phase lag in parameter estimation no matter what the learning rate is.
- The desired learning rate depends both on the changing rate of the parameter and the current input.

Online Adaptation

- Generalization of RLS
- Parameter convergence
- Uncertainty estimation

Parameter Convergence

- Whether the adaptation is stable?
- Whether it will converge to the true value?
- Whether the prediction error goes to zero?

Online Adaptation - Linear Time-Invariant Case

$$y_{k+1} = \theta^T x_k \qquad y_{k+1} \in \mathbb{R}; \quad \theta, x_k \in \mathbb{R}^n$$

Parameter update law $\theta_k = \theta_{k-1} + F_k x_{k-1} e_k$

A priori prediction error $e_k = y_k - \hat{y}_k \qquad \hat{y}_k = \theta_{k-1}^T x_{k-1}$

RLS Learning gain $F_k = \frac{1}{\lambda}\left\{F_{k-1} - \frac{F_{k-1}x_{k-1}x_{k-1}^T F_{k-1}}{\lambda + x_{k-1}^T F_{k-1} x_{k-1}}\right\} \qquad F_k = \left[\sum_{i=0}^{k-1} \lambda^{k-1-i} x_i x_i^T\right]^{-1}$

SGD Learning gain $F_k = \text{const}$

Parameter Error Dynamics

$$\theta_k = \theta_{k-1} + F_k x_{k-1} e_k$$

$$\tilde{\theta}_k = \theta - \theta_k$$

$$\begin{aligned}
\tilde{\theta}_k &= \tilde{\theta}_{k-1} - F_k x_{k-1} e_k \\
&= \tilde{\theta}_{k-1} - F_k x_{k-1} [y_k - \theta_{k-1}^T x_{k-1}] \\
&= \tilde{\theta}_{k-1} - F_k x_{k-1} \tilde{\theta}_{k-1}^T x_{k-1} \\
&= \tilde{\theta}_{k-1} - F_k x_{k-1} x_{k-1}^T \tilde{\theta}_{k-1}
\end{aligned}$$

$$\tilde{\theta}_k = (I - F_k x_{k-1} x_{k-1}^T) \tilde{\theta}_{k-1}$$

Parameter Error Dynamics - Scalar Case

$$\tilde{\theta}_k = (I - F_k x_{k-1} x_{k-1}^T)\tilde{\theta}_{k-1} \qquad \theta, x_k \in \mathbb{R}$$

$$F_k = \left[\sum_{i=0}^{k-1} \lambda^{k-1-i} x_i x_i^T\right]^{-1}$$

$$1 - F_k x_{k-1}^2 \in [0, 1]$$

$$\|\tilde{\theta}_k\| \leq \|\tilde{\theta}_{k-1}\|$$

The error dynamics are stable

$$F_k = \text{const}$$

$1 - F_k x_{k-1}^2$ depends on the data

$\tilde{\theta}_k$ can be diverging

Parameter Error Dynamics - Scalar Case

$$\tilde{\theta}_k = (I - F_k x_{k-1} x_{k-1}^T)\tilde{\theta}_{k-1} \qquad \theta, x_k \in \mathbb{R}$$

$$F_k = \left[\sum_{i=0}^{k-1} \lambda^{k-1-i} x_i x_i^T\right]^{-1}$$

$$1 - F_k x_{k-1}^2 \in [0, 1]$$

$$\|\tilde{\theta}_k\| \leq \|\tilde{\theta}_{k-1}\|$$

The error dynamics are stable

Necessary condition for convergence to 0:
persistent excitation: $x_k^2 \not\equiv 0$ as $k \to \infty$
(which implies $1 - F_k x_{k-1}^2 \not\equiv 1$ as $k \to \infty$)

λ smaller

$\rightarrow F_k$ bigger

$\rightarrow 1 - F_k x_{k-1}^2$ smaller

$\rightarrow$Faster convergence

Parameter Error Dynamics - Vector Case

$$\tilde{\theta}_k = (I - F_k x_{k-1} x_{k-1}^T)\tilde{\theta}_{k-1}$$

$$F_k = \left[\sum_{i=0}^{k-1} \lambda^{k-1-i} x_i x_i^T\right]^{-1}$$

All Eigen-values of $I - F_k x_{k-1} x_{k-1}^T$ belong to [0,1]

$$\|\tilde{\theta}_k\| \leq \|\tilde{\theta}_{k-1}\|$$

The error dynamics are stable

$$F_k = \text{const}$$

$I - F_k x_{k-1} x_{k-1}^T$ depends on the data

$\tilde{\theta}_k$ can be diverging

Parameter Error Dynamics - Vector Case

Claim: $F_k = \left[\sum_{i=0}^{k-1} \lambda^{k-1-i} x_i x_i^T\right]^{-1}$ All Eigen-values of $I - F_k x_{k-1} x_{k-1}^T$ belong to [0,1]

Proof: Let η be the Eigen value and v the associated Eigen vector.

$$v - F_k x_{k-1} x_{k-1}^T v = \eta v$$

$$\Leftrightarrow H_k v - x_{k-1} x_{k-1}^T v = \eta H_k v$$

$$\Rightarrow v^T H_k v - v^T x_{k-1} x_{k-1}^T v = \eta v^T H_k v$$

Since $0 \leq v^T H_k v - v^T x_{k-1} x_{k-1}^T v \leq v^T H_k v$

$$\eta \in [0, 1]$$

Parameter Error Dynamics - LTI case

- Theorem [Stability]: Under RLS parameter adaptation for $y_{k+1} = \theta^T x_k$ where $y_{k+1} \in \mathbb{R}; \quad \theta, x_k \in \mathbb{R}^n$,
  - The error dynamics $\tilde{\theta}_k = \theta - \theta_k$ is stable and $\|\tilde{\theta}_k\| \leq \|\tilde{\theta}_{k-1}\|$ for all k;
  - One necessary condition for $\lim_{k\to\infty} \tilde{\theta}_k = 0$ (from any initial estimate) is **persistent excitation**, i.e., $\|x_k\| \not\equiv 0$ as $k \to \infty$.

Q: Is it a sufficient condition?

Prediction Error Dynamics

Parameter Error $\tilde{\theta}_k = \tilde{\theta}_{k-1} - F_k x_{k-1} e_k$

Prediction Error $e_k = \tilde{\theta}_{k-1}^T x_{k-1} \leq \|\tilde{\theta}_{k-1}\| \|x_{k-1}\|$

The prediction error highly depends on the input data x_k

Fact 1: If the parameter error converges to zero, then the prediction error converges to zero.

Fact 2: If the prediction error converges to zero, the parameter error may not converge.

Fact 3: $\frac{|e_k|}{\|x_{k-1}\|}$ is bounded.

Prediction Error Dynamics

- For online adaptation, the input data is not random, but follows its own dynamics.
- The analysis of the prediction error should take that dynamics into consideration.
- In general, the prediction error converges to zero under RLS.

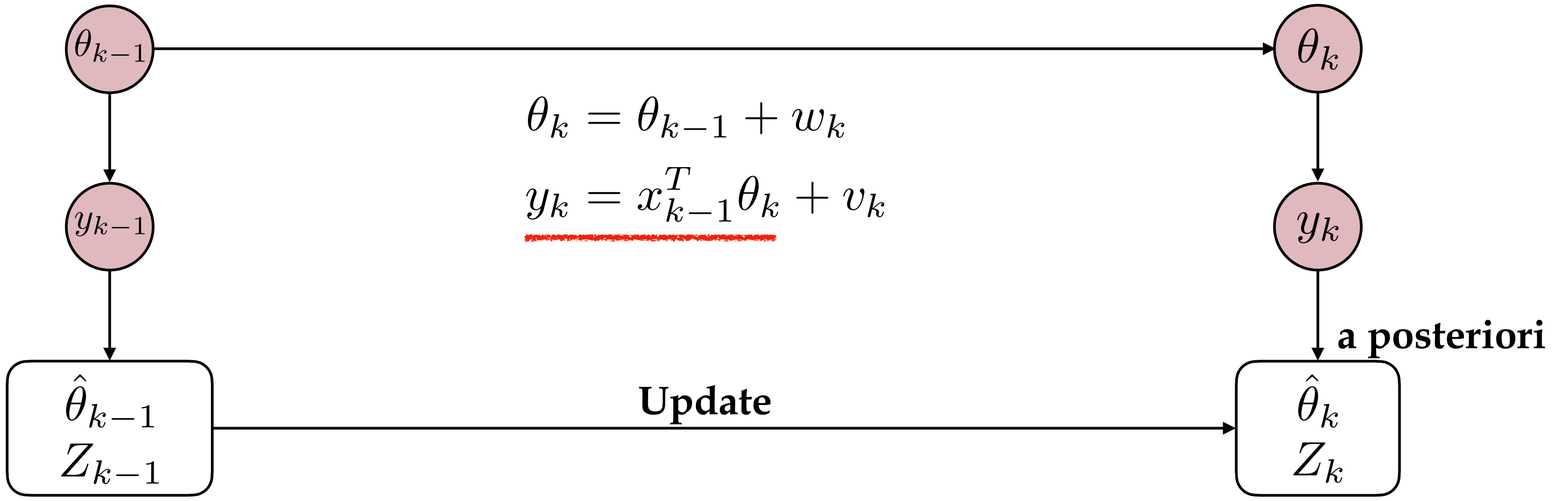
Advanced Tools For Convergence Analysis

- Lyapunov methods
- Hyperstability and passivity
- To be discussed later

Online Adaptation

- Generalization of RLS
- Parameter convergence
- Uncertainty estimation

Parameter Estimation with KF



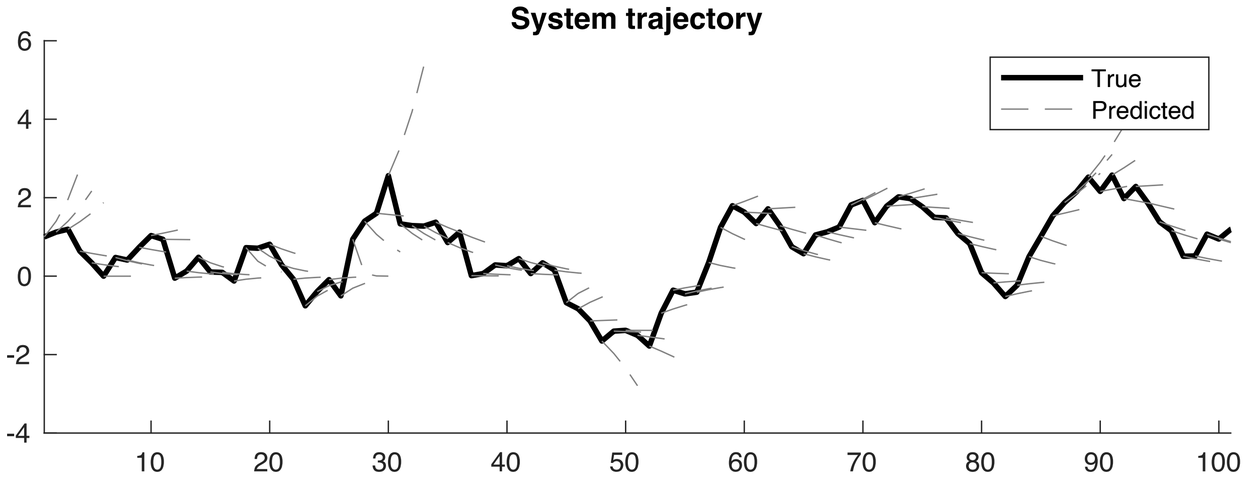
$$\hat{\theta}_k = \hat{\theta}_{k-1} + (Z_{k-1} + W_k)x_{k-1}[x_{k-1}^T(Z_{k-1} + W_k)x_{k-1} + V_k]^{-1}[y_k - x_{k-1}^T\hat{\theta}_{k-1}]$$

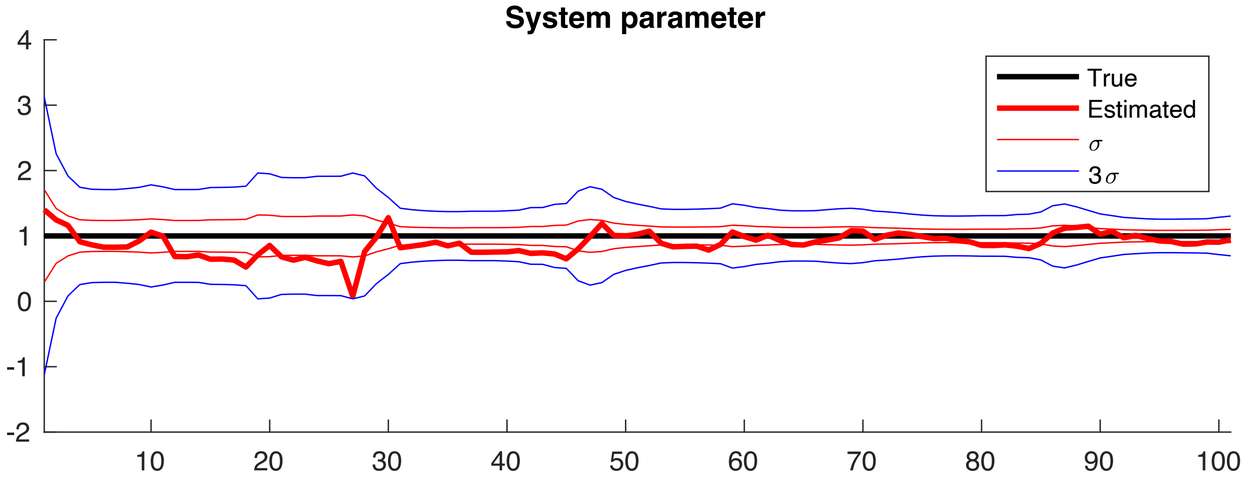
$$Z_k = (Z_{k-1} + W_k) - (Z_{k-1} + W_k)x_{k-1}(x_{k-1}^T(Z_{k-1} + W_k)x_{k-1} + V_k)^{-1}x_{k-1}^T(Z_{k-1} + W_k)$$

Covariance of the parameter error

Example

- $x_{k+1} = \theta x_k + w_k$
- $\theta = 1$, $x_0 = 1$, $\hat{\theta}_0 = 1.4$, $W = 0.25$
- $\lambda = 0.8$
- $\sigma_k = \sqrt{Z_k}$





Online Adaptation

- Generalization of RLS
- Parameter convergence
- Uncertainty estimation