VIOLATION-GUIDED LEARNING FOR CONSTRAINED FORMULATION IN NEURAL-NETWORK TIME SERIES PREDICTION

Benjamin W. Wah and Minglun Qian

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
E-mail:{wah,m-qian}@manip.crhc.uiuc.edu

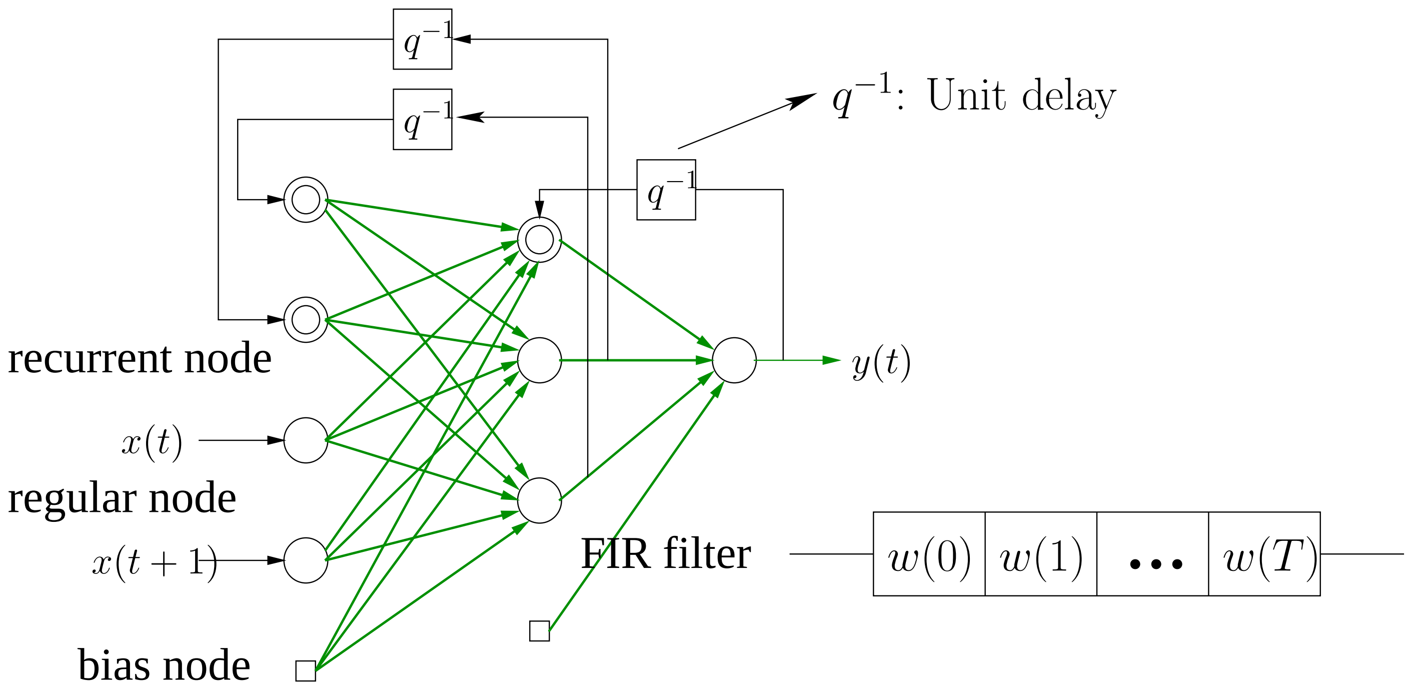
Outline

- **Motivations**
- **ANN models for time-series prediction**
- **Constrained formulation for ANN training**
- **Violation-guided Back-Propagation(VGBP) algorithm**
  - Gradient descents in the w subspace
  - Probabilistic acceptances in the w subspace
  - Relax-and-tighten strategy
- **Experimental results**
- **Conclusions**

ANN Models for Time Series Prediction

- Time-series prediction
  - Given a sequence of values observed in the past, predict future values
- Existing architectures
  - Recurrent neural networks (RNN)
  - Memory-based neural networks (TDNN and FIR-NN)
  - Dynamic recurrent neural network (DRNN): FIR + feedback without delay
- Proposed architecture: recurrent FIR neural network (RFIR)
  - No consensus on which architecture is better [Horne][Hallas]
  - Training algorithm is more important than architecture [Koskela]
  - *RFIR*: FIR + recurrent feedback with time delay

Key Point 1: Recurrent FIR Architecture



- Unit delay $\Rightarrow$ easier to derive gradient compared with DRNN

Performance Metrics

- Normalized mean square error (nMSE):

$$\varepsilon = \frac{1}{\sigma^2 N} \sum_{t=t_0}^{t_1} (o(t) - d(t))^2, \qquad (1)$$

  - σ^2 is the variance of the true time series in $[t_0, t_1]$
  - $o(t)$ is actual output at t, $d(t)$ is desired output
  - N is number of patterns in the measurement
- Open-loop single-step measurement: external input is true observed data
- Close-loop iterative measurement: external input is predicted output

Traditional Formulations for ANN Training

- Unconstrained formulation

$$\min_{w} \quad E(w) = \frac{1}{n}\sum_{t=1}^{n}(o_t(w) - d_t)^2 \tag{2}$$

- Training algorithms
  - BP/BP variants and gradient-based methods
  - Genetic algorithms
  - Simulated annealing
- Issues
  - No guidance when search reaches a non-zero local minimum of $E(w)$
  - Nonuniform errors across patterns – not good for prediction

Key Point 2: Proposed Constrained Formulations

- Each pattern treated as an additional constraint:

$$h_t(w) = (o_t(w) - d_t)^2 \leq \tau, \tag{3}$$

  - τ decreases towards 0 as looser constraints are satisfied
  - Non-zero constraints provide guidance when search reaches a sub-optimum of the objective function

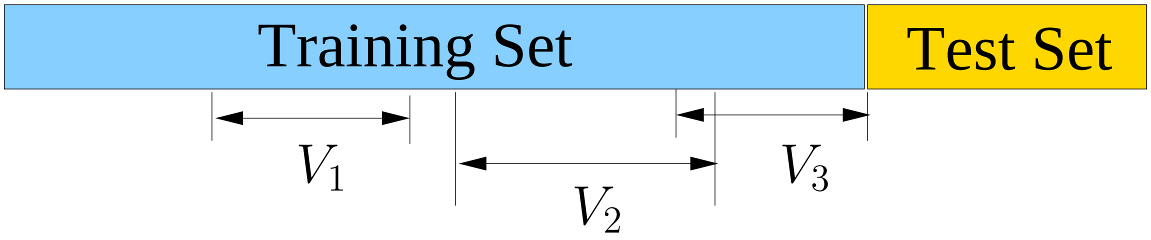
Traditional Cross-Validation

- Divide historical data into two *disjoint* sets
  - Training set
  - Cross-validation set

- Issues
  - Hard to choose appropriate validation set: how long?
  - Data used for cross-validation cannot be used for training
  - Only one validation set is used at any time: not good when time series is multi-stationary

Key Point 3: Proposed Cross-Validation Method

- Multiple validation set(s) within training set



- Iterative and single-step validation errors added as new constraints
- Advantages
  - Training patterns fully used
  - Multiple validation sets cover multiple regimes in a multi-stationary time-series
  - Flexibility in choosing validation sets

Constrained Formulation with Cross-Validation

- Constrained formulation

$$
\begin{array}{ll}
\min_w & E(w) = \frac{1}{n}\sum_{t=1}^{n} max\{(o_t(w) - d_t)^2 - \tau, 0\} \\
\text{s.t.} & h_t(w) = (o_t(w) - d_t)^2 \le \tau, \\
& h_i^I(w) = \varepsilon_i^I \le \tau_i^I, \qquad \text{(iterative validation)} \\
& h_i^S(w) = \varepsilon_i^S \le \tau_i^S, \qquad \text{(single-step validation)}
\end{array} \tag{4}
$$

- Constrained formulation solved by violation-guided back-propagation (VGBP) based on *discrete Lagrange-multiplier theory* [Wah & Wu]
- Transform Eq (4) into augmented Lagrangian function:

$$
\begin{aligned}
L(w,\lambda) = \; & E(w) + \sum_{t=1}^{n} \left(\lambda_t max\{0, h_t - \tau\} + \tfrac{1}{2} max^2\{0, h_t - \tau\}\right) + \\
& \sum_{k=1}^{v} \sum_{i=I,S} \left(\lambda_k^i max\{0, \varepsilon_k^i - \tau_k^i\} + \tfrac{1}{2} max^2\{0, \varepsilon_k^i - \tau_k^i\}\right)
\end{aligned} \tag{5}
$$

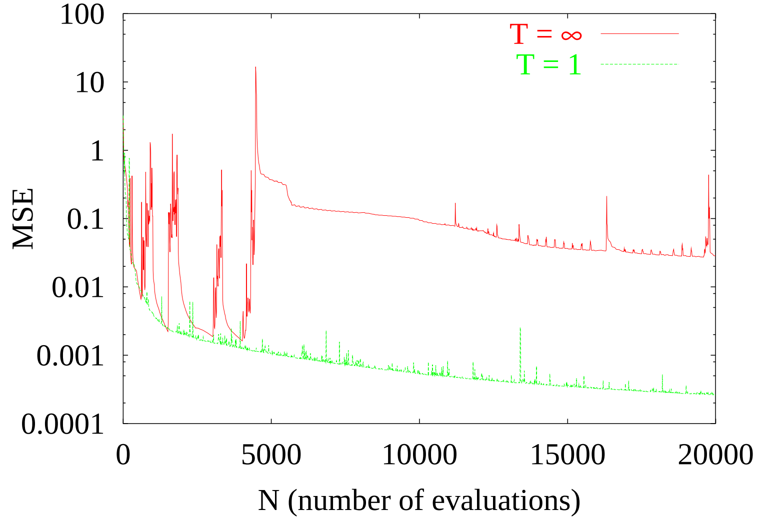
Key Point 4: Search for Saddle Points

- Saddle point
  - Local minimum of $L(w, \lambda)$ in w subspace
  - Local maximum of $L(w, \lambda)$ in λ subspace
- Gradient descents and stochastic acceptances in w subspace by VGBP
  - Using BP to generate approximate gradient direction in $L(w, \lambda)$
  - Accepting trial points with Metropolis probability using fixed T

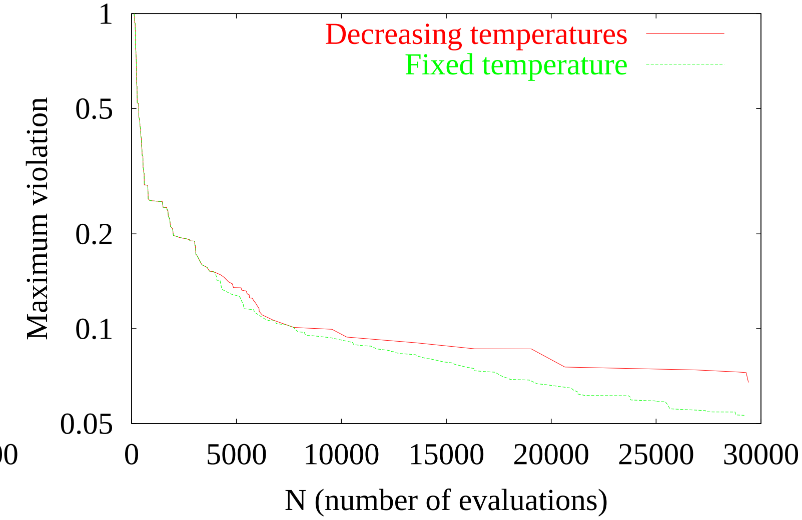
$$A_T(\mathbf{w}', \mathbf{w})|_\lambda = exp\left\{\frac{(L(\mathbf{w}) - L(\mathbf{w}'))^+}{T}\right\} \tag{6}$$

    where $x^+ = min\{0, x\}$ and T is temperature
- Gradient assents in λ subspace by deterministic increases of λ
  - Big violation $\Rightarrow$ increased λ $\Rightarrow$ more contribution

Justification for using Fixed T



(a) Annealing *vs* deterministic acceptance



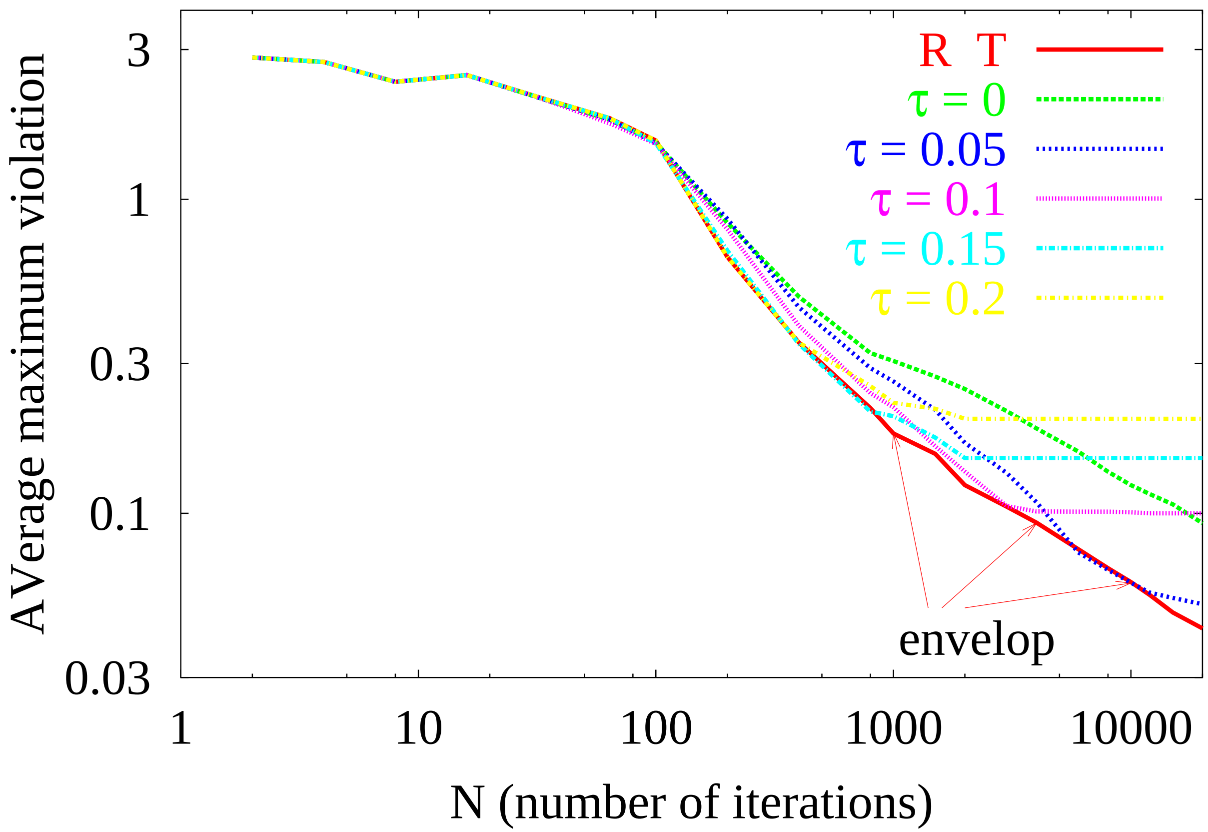
(b) Fixed T *vs* decreasing T

- Annealing avoids search going to very bad regions frequently
- Very low temperatures freeze search
  - Both BP and low-temperature search perform local search

Key Point 5: Relax-and-Tighten Strategy

- Observations
  - Looser constraints
    $\Rightarrow$ Faster convergence and larger maximum violation at convergence
  - Tighter constraints
    $\Rightarrow$ Slower convergence and smaller maximum violation at convergence
- Relax-and-Tighten strategy
  - Loose constraints in the beginning and tighten gradually
    $\Rightarrow$ Faster convergence, and smaller maximum violation at convergence

Relax-and-Tighten Strategy



Chaotic Time Series

- Benchmarks

Time Series	Description	Training Set	Single-Step Pred	Iterative Pred
Sunspots	yearly sunspots number	1700-1920	1921-1994	–
Laser	laser intensity	1-1000	1001-1100	1001-1100
Mackey-Glass(17)	differential equation	1-500	501-2000	501-600
Mackey-Glass(30)	differential equation	1-500	501-2000	501-600
Henon Map	bi-variate equation	1-5000	1-5000	–
Lorenz Attractor	differential equations	1-4000	4001-4150	–
Ikeda Attractor	plane wave	1-10000	10001-12000	–

  - Sunspots and Laser time series from real data
  - The rests are artificial chaotic time series

- Goals
  - less weights
  - better performance

Sunspots and Laser Time Series

- **Sunspots**

Method	No. of Free Variables	Training	Single-Step Testing			
		1700-1920	1921-55	1956-79	1980-94	1921-94
AR(12)	13	0.128	0.126	0.36	0.306	0.238
TAR	18	0.097	0.097	0.28	0.306	0.197
WNet	113	0.082	0.086	0.35	0.313	0.219
SSNet	N/A	-	0.077	N/A	N/A	N/A
DRNN	30	0.105	0.091	0.273	N/A	N/A
COMM	N/A	0.079	0.065	0.24	0.188	0.148
ScaleNet	N/A	0.086	0.057	0.13	N/A	N/A
VGBP	**11**	0.0559	**0.0337**	**0.0524**	**0.0332**	**0.0397**

- **Laser**

Method	Number of weights	Training	Single Step Prediction		Iterative Prediction	
		100-1000	1001-1050	1001-1100	1001-1050	1001-1100
FIRNN	1105	0.00044	0.00061	0.023	0.0032	0.0434
ScaleNet	N/A	0.00074	0.00437	0.0035	N/A	N/A
VGBP (Run 1)	**461**	0.00036	**0.00043**	**0.0034**	0.0054	**0.0194**
VGBP (Run 2)	**461**	0.00107	**0.00030**	**0.00276**	**0.0030**	**0.0294**

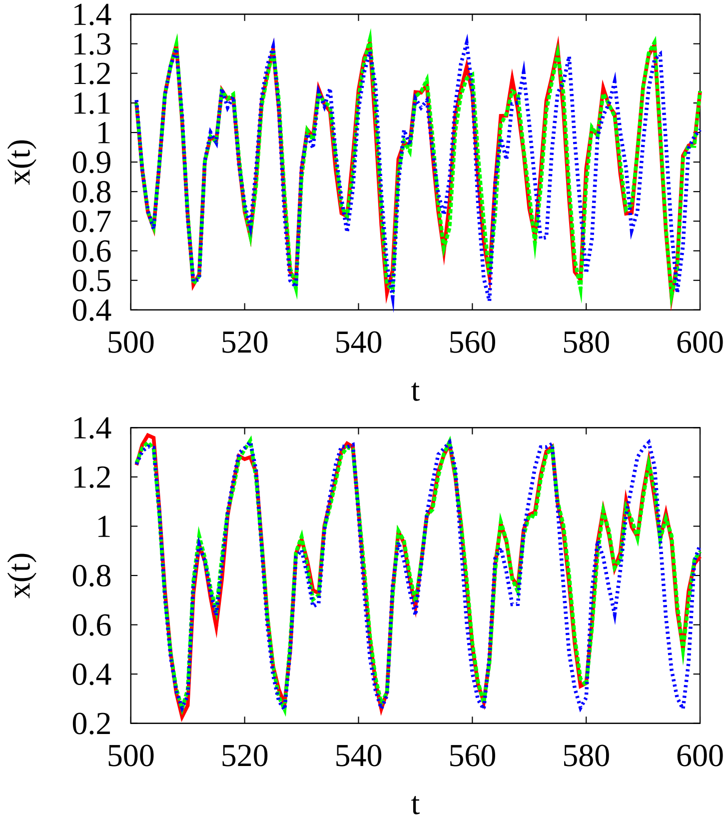
Artificial Chaotic Time Series

- **Single-step prediction**

Bench-Mark	Training Set	Testing Set	Performance Metrics		Design Methods				
					C.C.	Linear	FIR-NN	DRNN	VGBP
MG17	1-500	501-2000	$nMSE$		0.6686	0.320	0.00985	0.00947	0.000057
			# of weights		0	N/A	196	197	121
MG30	1-500	501-2000	$nMSE$		0.3702	0.375	0.0279	0.0144	0.000374
			# of weights		0	N/A	196	197	121
Henon	1-5000	5001-10000	$nMSE$		1.633	0.874	0.0017	0.0012	0.000034
			# of weights		0	N/A	385	261	209
Lorenz	1-4000	4001-5500	$nMSE$	x	0.0768	0.036	0.0070	0.0055	0.000034
				z	0.2086	0.090	0.0095	0.0078	0.000039
			# of weights		0	N/A	1070	542	527
Ikeda	1-10000	10001-11500	$nMSE$	$Re(x)$	2.175	0.640	0.0080	0.0063	0.00023
				$Im(x)$	1.747	0.715	0.0150	0.0134	0.00022
			# of weights		0	N/A	2227	587	574

Iterative Prediction for Mackey-Glass

- VGBP: green lines, nMSEs being 0.018(0.0064) for MG17(MG30)
- Wan's: red lines, nMSEs being 0.3832(0.1487) for MG17(MG30)



Conclusions

- Five key points
  - Combined FIR and recurrent structure in RFIR NN
  - Guidance based on violated patterns in a constrained formulation
  - New cross-validation for handling multi-stationary time series
  - Efficient and stable violation-guide back-propagation algorithm
  - Relax-and-tighten strategy for improved speed and convergence
- Most important sources for performance improvement
  - Constrained formulation
  - Relax-and-tighten strategy

Violation Guided Back-Propagation

```
procedure VGBP
    set initial w = (w, λ), η_0, T, N_S
    run one pass of the feedforward process
    while stopping condition is not satisfied do
        for k ← 1 to N_S do
            for t ← t_0 to t_1
                for i ← 1 to N_o
                    e_i(t) = λ_t e_i(t)
                end_for
            end_for
            run BP to obtain δw
            accept w' = w + δw using Eq.(6)
            set τ ← 0.95τ if max{h} ≤ 0.1τ
        end_for
        adjust λs according to constraint violations
        adjust η according to acceptance ratio
    end_while
end_procedure
```

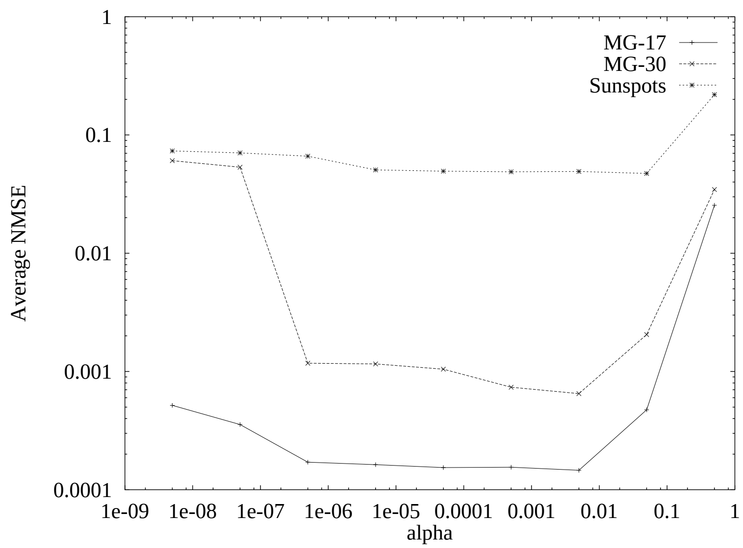
Choice of Temperature

- Set T according to

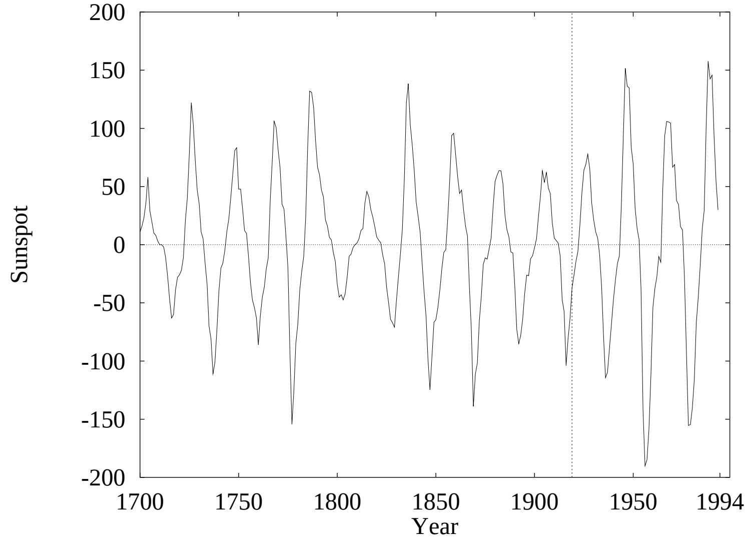
$$T = \alpha N_p R, \tag{7}$$

  - N_p: number of training patterns
  - R: magnitude of desired output data

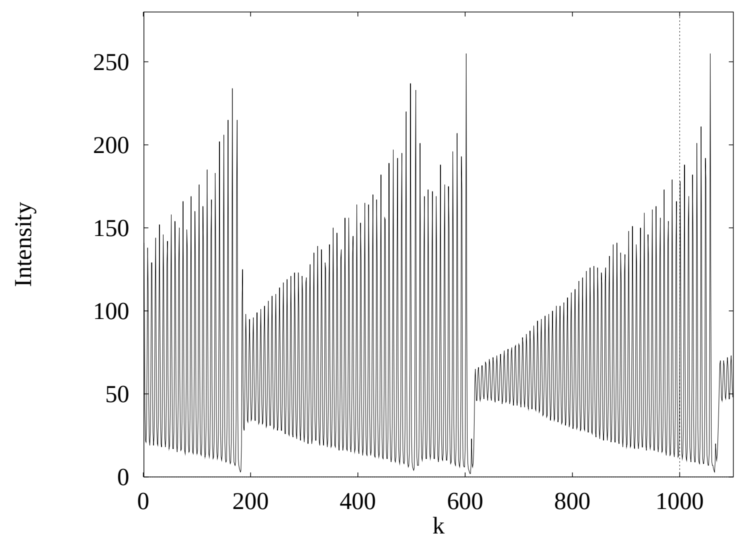
- When $\alpha \in [10^{-6}, 10^{-2}]$, performance insensitive to T



Sunspots and Laser Time Series

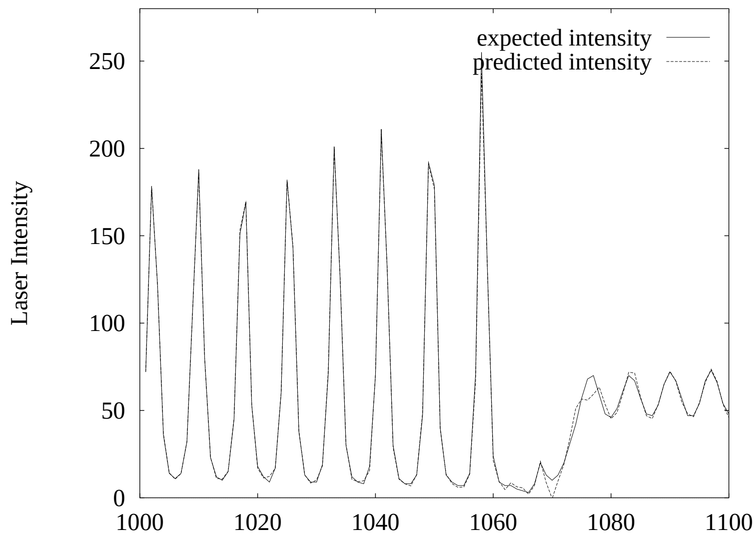


(a) Sunspots time series

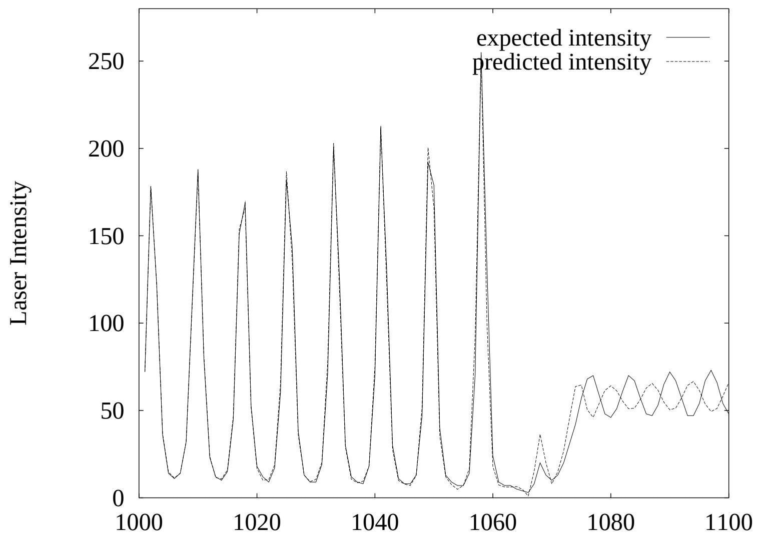


(b) Laser time series

Laser Time Series Prediction

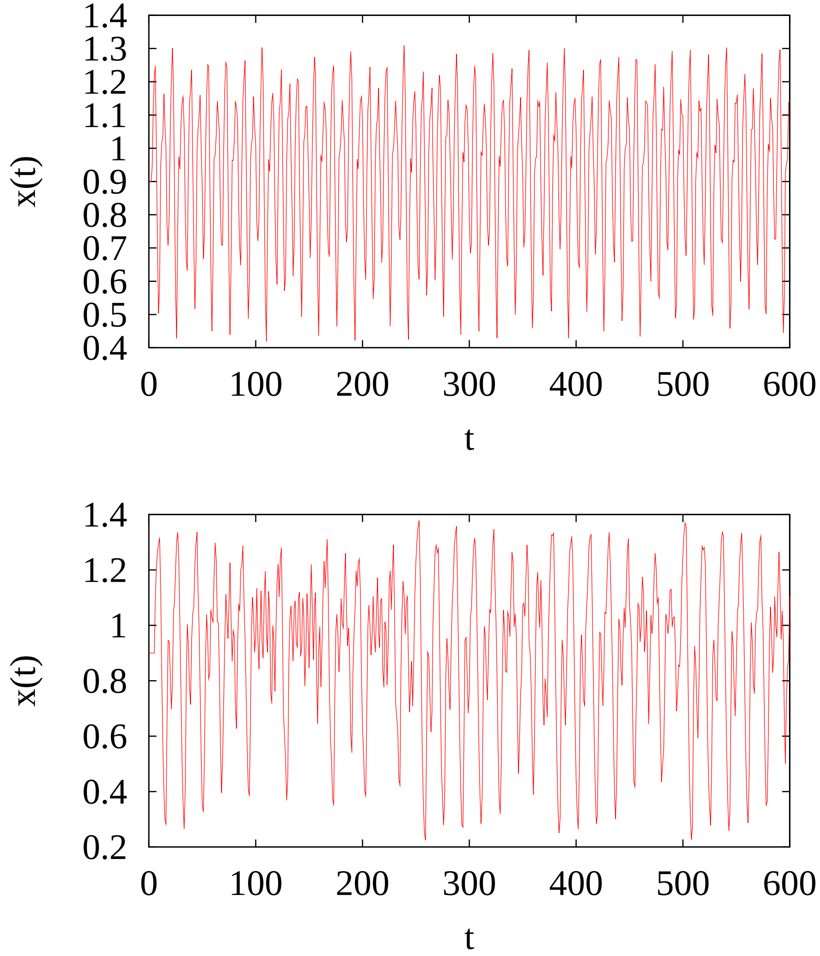


(a) Single-step prediction

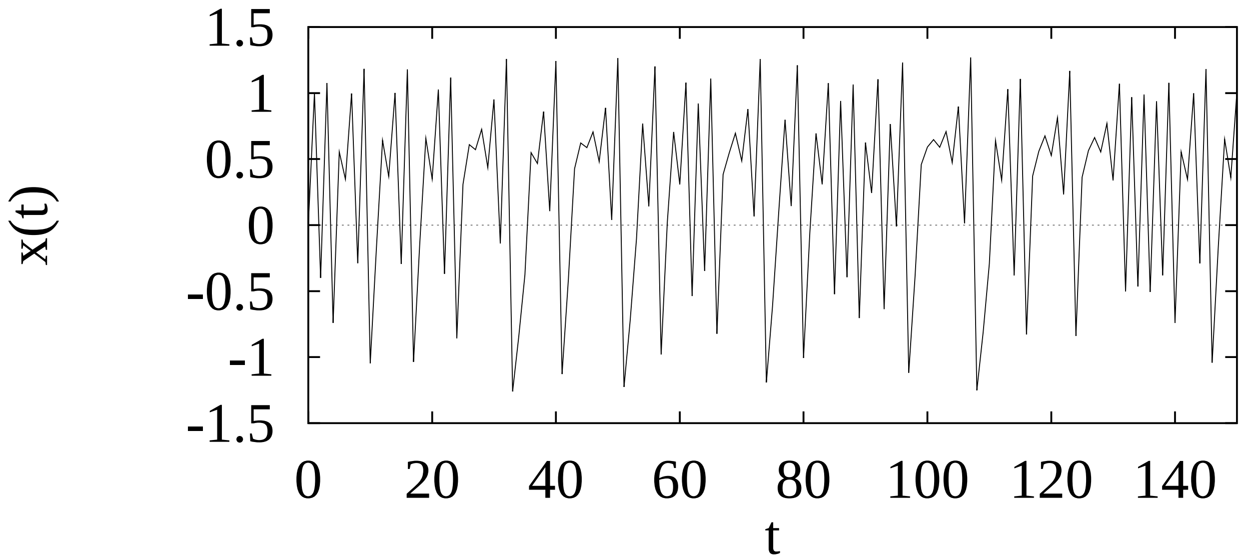


(b) Iterative prediction

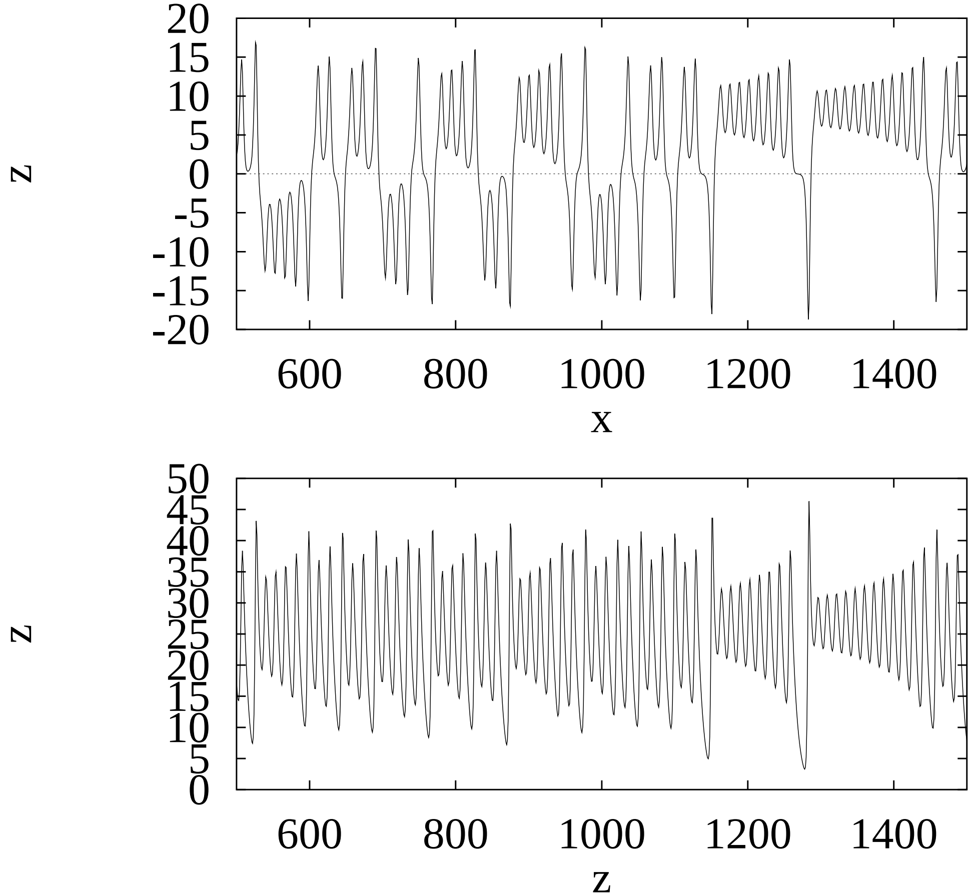
Mackey-Glass Time Series



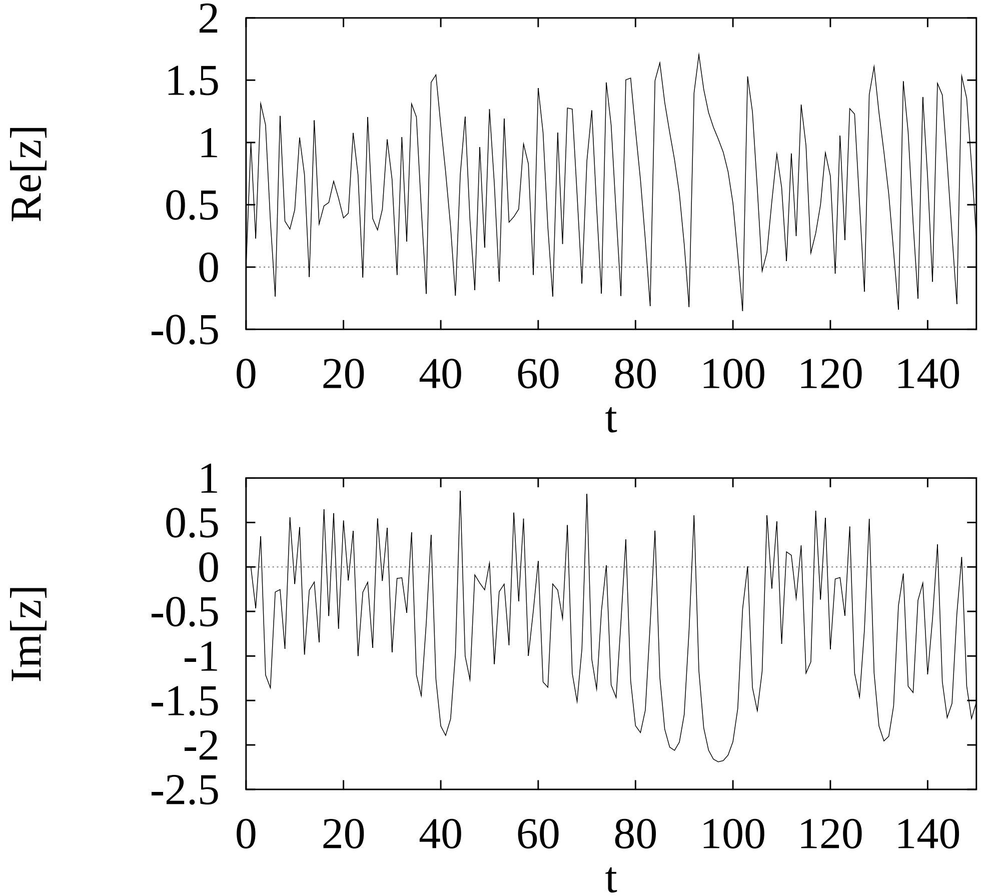
Henon Time Series



Lorenz Time Series



Ikeda Time Series



General Framework