Physica Scripta

OPEN ACCESS

RECEIVED
2 June 2025

REVISED
19 August 2025

ACCEPTED FOR PUBLICATION
17 September 2025

PUBLISHED
13 November 2025

PAPER

Regularisation and the least squares problem in the analysis of echo state networks

Yuwei Lu and Joab R Winkler*

School of Computer Science, The University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom
* Author to whom any correspondence should be addressed.

E-mail: ylu136@sheffield.ac.uk **and** j.r.winkler@sheffield.ac.uk

Keywords: Echo state networks, regularisation, condition estimation, time series

Abstract

An echo state network (ESN) is a recurrent neural network that has several advantages with respect to a deep neural network, including its fast training phase and the absence of vanishing and exploding gradients. The training phase reduces to solving the least squares (LS) problem $\mathbf{w}_{\text{ls}} = \arg\min_{\mathbf{v}} \|\mathbf{vX} - \mathbf{y}\|_2^2$, where $\mathbf{X}$ is the reservoir matrix, and $\mathbf{X}$ and $\mathbf{y}$ are functions of the training data. It is common to add regularisation to this problem because, it is claimed, it minimises the adverse effects of overfitting. Recent work in deep neural networks, physics informed neural networks and regression has shown, however, that regularisation does not solve the problem of overfitting, and thus this paper considers the application of regularisation to ESNs by analysing their predictive abilities on several time series, including a non-linear communication channel, the Hénon map and multiple superimposed oscillations. It is shown that the solution $\mathbf{w}_{\text{ls}}$ of the LS problem is, for many problems, stable with respect to a perturbation in $\mathbf{y}$ for a wide range of parameter values of an ESN, and thus regularisation must not be applied to these problems. Each problem must, however, be considered because the need, or otherwise, to apply regularisation is dependent on many parameters of an ESN. Furthermore, regularisation is not benign because its use when a condition on the rate of decay of the singular values of $\mathbf{X}$ is not satisfied leads to a large error between the theoretically exact and regularised solutions of the LS problem.

1. Introduction

A recurrent neural network (RNN) is a neural network that captures temporal dependencies in sequential data, and it is used in, for example, text processing [1], speech recognition [2] and the prediction of chaotic time series [3]. It is trained by backpropagation through time, which is an iterative algorithm for the determination of the weights of the network, and it may suffer from vanishing and exploding gradients. An echo state network (ESN), which is shown in figure 1, is a simplified form of an RNN that does not suffer from these problems. It has three components, specifically, an input unit, a reservoir that contains the neurons, and an output unit, and only the output weights of the network are trained [4, 5]. The weights of the neurons in the reservoir are assigned random values that do not change during training, which marks a difference between an RNN and an ESN. There are therefore many fewer parameters to train than in an RNN, which reduces the computational cost of training the network. An ESN can be extended to include feedback of a component of its output back into the reservoir, such that the state of the reservoir at time $t = n$ is a function of the output of the ESN at time $t = n - 1$ [6, 7].

Training an ESN reduces to the calculation of the solution $\mathbf{w}_{\text{ls}}$ of the the least squares (LS) problem,

$$\mathbf{w}_{\text{ls}} = \arg\min_{\mathbf{v}} = \|\mathbf{vX} - \mathbf{y}\|^2 = \mathbf{yX}^{\dagger}, \qquad \|\cdot\| = \|\cdot\|_2, \tag{1}$$

where the reservoir matrix $\mathbf{X} = \mathbf{X}(\mathbf{u})$ defines the state of the reservoir for the input data $\mathbf{u}$, $\mathbf{y}$ is the desired output for the data $\mathbf{u}$, and the components of $\mathbf{w}_{\text{ls}}$ are the weights v_i, where $\mathbf{v} = \{v_i\}$, that minimise the error between

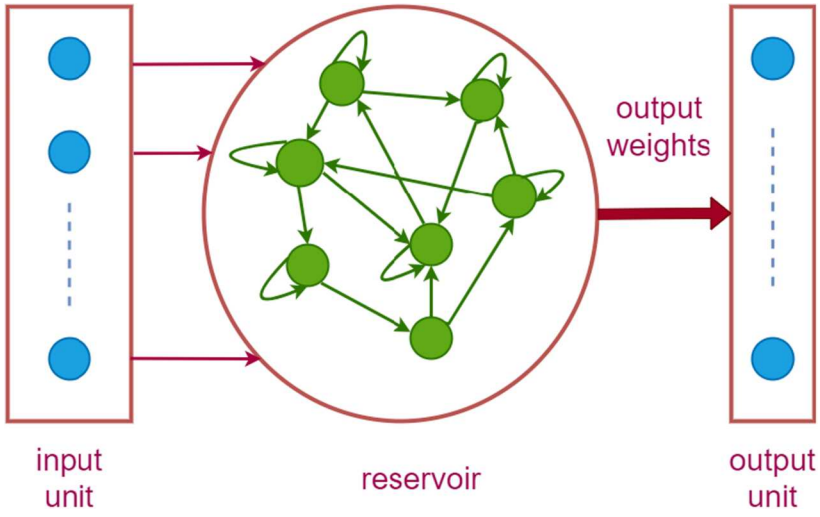

Figure 1. The input unit, reservoir that contains the neurons and output unit of an ESN.

the computed output $\mathbf{vX}$ and the desired output $\mathbf{y}$. The pair $(\mathbf{u}, \mathbf{y})$ defines the training data from which the weight vector $\mathbf{w}_{\text{ls}}$ is computed. This vector is then used to make predictions on new data.

Practical applications of an ESN require that it be numerically stable, that is, a perturbation of order ϵ in the input to an ESN yield a perturbation of the same order in $\mathbf{w}_{\text{ls}}$ and predictions on new data. This paper addresses one aspect of the stability of an ESN, and it is shown that three problems must be considered for a complete study of the stability of an ESN.[1]

Let

$$\mathbf{u} = [u(1) \;\; u(2) \;\; \cdots \;\; u(S) \;\; u(S+1) \;\; \cdots],$$

be a time series, the first S entries of which, $u(i)$, $i = 1, \ldots, S$, are used to train an ESN, and let the reservoir have N neurons.[2] The state $\mathbf{x}(n)$ of the reservoir at time $n = 1, \ldots, S$, is

$$\mathbf{x}(n) = (1 - \alpha)\mathbf{x}(n-1) + \alpha(\tanh(\mathbf{W}^{\text{in}}[1; u(n)] + \mathbf{W}\mathbf{x}(n-1))), \tag{2}$$

where the function $\tanh(\cdot)$ is called the activation function,

$$\mathbf{x}(n) = [x_1(n) \;\; x_2(n) \;\; \cdots \;\; x_N(n)]^T \in \mathbb{R}^N,$$

α is the leakage, $0 < \alpha \leqslant 1$, and $\mathbf{W}^{\text{in}} \in \mathbb{R}^{N\times 2}$ and $\mathbf{W} \in \mathbb{R}^{N\times N}$ are random matrices, each of whose entries is drawn from the uniform distribution in the open interval $(-1, 1)$. The vector $\mathbf{u}$ is normalised by $\|\mathbf{u}\|_\infty$ in order that the vector that multiplies $\mathbf{W}^{\text{in}}$ in (2) is balanced. It is assumed the reservoir is at rest on initialisation and thus $\mathbf{x}(0) = \mathbf{0}$.

Equation (2) is executed for $n = 1, \ldots, S$, which allows the matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ u(1) & u(2) & \cdots & u(S) \\ \mathbf{x}(1) & \mathbf{x}(2) & \cdots & \mathbf{x}(S) \end{bmatrix} \in \mathbb{R}^{(N+2)\times S}, \tag{3}$$

to be constructed, and the output weight vector $\mathbf{w}_{\text{ls}} \in \mathbb{R}^{N+2}$ is the solution of the LS problem (1), where

$$\mathbf{y} = [u(2) \;\; u(3) \;\; \cdots \;\; u(S+1)] \in \mathbb{R}^S, \tag{4}$$

which is one time step ahead of the input $\{u(i)\}_{i=1}^S$, and

$$\mathbf{X}^\dagger = \begin{cases} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T & \text{if} \quad N + 2 \geqslant S, \\ \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} & \text{if} \quad N + 2 < S. \end{cases}$$

[1] The work in this paper is therefore the first part of a detailed study of the stability of ESNs.

[2] A washout period is allowed for by only considering the time series $\mathbf{u}$ after the transient response has decayed.

The predicted output $\tilde{y}(S+k+1)$ of the ESN on a new data sample $u(S+k), k=1,2,\ldots,$ is

$$\tilde{y}(S+k+1) = \mathbf{w}_{\text{ls}} \begin{bmatrix} 1 \\ u(S+k) \\ \mathbf{x}(S+k) \end{bmatrix}, \quad k = 1, 2, \ldots \tag{5}$$

The term on the right hand side is a function of the data samples $u(n)$ and state $\mathbf{x}(n)$ of the reservoir at times $n=1,\ldots,S+k$, and the term on the left hand side is the output at time $n=S+k+1$, and thus (5) is a one step ahead predictor.

It is noted above that the output $\mathbf{y}$ of an ESN must be stable with respect to a change $\delta\mathbf{u}$ in $\mathbf{u}$, but this issue has not been considered. Numerical issues associated with this stability, using methods of computational linear algebra, are discussed in section 2 and it is shown that three problems must be considered to address this issue fully. The most challenging problem requires the development of a condition number of the LS problem (1) because the perturbation $\delta\mathbf{X}$ in $\mathbf{X}$ is a structured matrix. This paper considers, therefore, a simpler problem in which $\delta\mathbf{X}=\mathbf{0}$, which will allow extension to the situation that occurs when the structure of $\delta\mathbf{X}$ is included in the analysis of the condition of the LS problem. The numerical condition and regularisation of this simpler LS problem are considered in sections 3 and 4, respectively. Examples that demonstrate the theoretical analysis are in section 5 and future work is discussed in section 6. The paper is summarised in section 7.

This discussion leads to the aims of the paper:

- Tikhonov regularisation is applied to (1) in order to minimise overfitting [4, section 3], [5, section 4.2], [8, section 8.1.1].[3] It leads to the minimisation

$$\mathbf{w}(\lambda) = \arg\min_{\mathbf{v}} \{\|\mathbf{v}\mathbf{X} - \mathbf{y}\|^2 + \lambda \|\mathbf{v}\|^2\}, \qquad \mathbf{w}(0) = \mathbf{w}_{\text{ls}}, \tag{6}$$

  where the regularisation parameter $\lambda \geqslant 0$ controls the extent that the constraint on $\|\mathbf{v}\|$ is imposed on $\mathbf{w}_{\text{ls}}$. It has, however, been shown in physics informed neural networks [9, section 2.3], regression [10] and deep learning [11, section 1.2] that regularisation does not solve the problem of overfitting, and thus the objective of regularisation with respect to ESNs must be determined.

  This objective is realised by noting that regularisation imposes stability on the solution of an ill conditioned set of linear algebraic equations, assuming the coefficient matrix is exact. The application of regularisation requires that the discrete Picard condition, which is a condition on the rate of decay of the singular values of $\mathbf{X}$, be satisfied [12]. It is necessary to confirm that this condition is satisfied before regularisation is applied because it is not benign. In particular, the application of regularisation leads to a large error in $\mathbf{w}(\lambda^{\text{opt}})$, where λ^{opt} is the optimal value of λ, with respect to $\mathbf{w}(0)$ if the discrete Picard condition is not satisfied.[4] This property of regularisation highlights the importance of an investigation into the application of regularisation to (1).

- The stability and error of $\mathbf{w}_{\text{ls}}$ must be determined for the conditions $N+2 \geqslant S$ and $N+2 < S$ because they influence the predictive ability of an ESN. It is shown that the criterion for this selection is associated with overfitting and the bias-variance trade-off.

The motivation for this paper follows from the observation that the need, or otherwise, to apply regularisation and the determination of the optimal value of λ are not addressed in the literature on ESNs. For example, the value $\lambda=1$ is used in [13, section 4] to train an ESN to generate the figure 8, but it is not verified that the discrete Picard condition is satisfied, and thus that regularisation is required, and the reason for the selection of this value of λ is not stated. The Lorenz time series [14], Mackey-Glass time series [15] and NARMA10 time series [16, section 4] are considered in [17, section 4] using values of λ of (10^{-5} and 10^{-4}), (10^{-5} and 10^{-4}), and (10^{-4} and 10^{-3}), respectively, but the justification for the application of regularisation and these values of λ is not stated. Also, the values $\lambda=10^{-5}$ and $\lambda=10^{-3}$ are used in [18, section 3.2] for the analysis of the Lorenz time series, but the reason for using these two values of λ is not stated. Similarly, the value $\lambda=10^{-6}$ is used in [4, section 3] for the analysis of the Mackey-Glass time series, but it is not justified. More generally, the discrete Picard condition and condition estimation of the LS problem are not considered in the literature on ESNs, even though the objective of regularisation is the imposition of stability on its solution.

The results of the paper are now summarised:

[3] Tikhonov regularisation is also known as ridge regression and it will, for brevity, be termed regularisation.

[4] Methods for calculating the optimal value of λ are considered in section 4.1.

- Experiments on many time series show that for a wide range of parameter values of an ESN, for example, the value of α, the length of the training data and the number of neurons in the reservoir, the LS problem (1) is well conditioned with respect to a perturbation in $\mathbf{y}$. It follows that, for many time series, regularisation must not be imposed, but each problem must be considered to determine if it is required because, as noted above, its incorrect application leads to a large error between the theoretically exact and regularised solutions of the LS problem.
- Two methods, generalised cross validation (GCV) [19, 20] and the L-curve [21, section 4.6], are compared for the determination of λ^{opt}. The L-curve yields consistently better results, and in particular, it returns $\lambda^{\text{opt}} = 0$ if regularisation must not be applied because the discrete Picard condition is not satisfied, but the GCV often returns $\lambda^{\text{opt}} \gg 0$, which is unsatisfactory.
- Analysis of the error and stability of $\mathbf{w}_{\text{ls}}$ show that $N + 2 \geqslant S$ must be satisfied because this condition yields the solution $\mathbf{w}_{\text{ls}}$ that is more stable with respect to a perturbation in $\mathbf{y}$ than the solution obtained when $N + 2 < S$. Also, the error in the solution obtained with the condition $N + 2 < S$ is large because $\mathbf{y}$ does not, in general, lie in the row space of $\mathbf{X}$, but this error is zero if $N + 2 \geqslant S$.

2. The stability of an ESN

The conditions for which the output of an ESN is stable, and the conditions for which it is unstable, must be determined because they are critical for practical applications of an ESN. Analysis of this stability and instability requires that three problems be considered, and this paper addresses one of these problems, specifically, the numerical condition of the LS problem (1) whose solution $\mathbf{w}_{\text{ls}}$ is the weight vector that is used for predictions on new data (5). These three problems are now described because they provide the motivation for the work described in this paper.

A condition number of the problem (1) is usually computed by linear error analysis, which assumes that a perturbation $\delta\mathbf{u}$ in $\mathbf{u}$, and therefore, from (4), a perturbation $\delta\mathbf{y}$ in $\mathbf{y}$, does not cause a change $\delta\mathbf{X}$ in the state $\mathbf{X}$ of the neurons in the reservoir. This linear analysis is inadequate because it is shown in [22] that to first order in $\mathbf{x}(n-1)$ and $u(n)$, $n = 1, \ldots, S$,

$$\begin{bmatrix} \delta\mathbf{x}(1) \\ \delta\mathbf{x}(2) \\ \vdots \\ \delta\mathbf{x}(S) \end{bmatrix} = \begin{bmatrix} \mathbf{l}_{11} & & & & \\ \mathbf{l}_{21} & \mathbf{l}_{22} & & & \\ & & \vdots & & \\ \mathbf{l}_{S1} & \mathbf{l}_{S2} & \mathbf{l}_{S3} & \cdots & \mathbf{l}_{SS} \end{bmatrix} \begin{bmatrix} \delta u(1) \\ \delta u(2) \\ \vdots \\ \delta u(S) \end{bmatrix}, \tag{7}$$

where $\mathbf{l}_{ij} \in \mathbb{R}^N$, $i, j = 1, \ldots, S$, and the coefficient matrix $\mathbf{L} \in \mathbb{R}^{NS \times S}$,

$$\mathbf{L} = \{\mathbf{l}_{ij}\}_{i,j=1}^{S} = \mathbf{L}(\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(S-1), u(1), u(2), \ldots, u(S)),$$

is lower triangular because of causality. Furthermore, it follows from (3) that

$$\delta\mathbf{X} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \delta u(1) & \delta u(2) & \cdots & \delta u(S) \\ \delta\mathbf{x}(1) & \delta\mathbf{x}(2) & \cdots & \delta\mathbf{x}(S) \end{bmatrix} \in \mathbb{R}^{(N+2) \times S},$$

where, from (7),

$$\delta\mathbf{x}(n) = \sum_{j=1}^{n} \mathbf{l}_{nj} \delta u(j), \qquad n = 1, \ldots, S,$$

and thus $\delta\mathbf{X}(3{:}N+2, :)$ is a structured matrix whose NS entries are functions of S random variables $\delta u(n)$, $n = 1, \ldots, S$. This structure must be considered when the effect of a perturbation $\delta\mathbf{u}$ on the state of the neurons in the reservoir is considered.

Stability analysis of an ESN requires that three problems be considered:

- Problem 1: The determination of the perturbations $\delta\mathbf{x}(n)$, $n = 1, \ldots, S$, and thus the perturbation $\delta\mathbf{X}$ in the state of the neurons in the reservoir due to a perturbation $\delta\mathbf{u}$ in the data $\mathbf{u}$, as shown in (7).
- Problem 2: The determination of the perturbation $\delta\mathbf{w}_{\text{ls}}$ in the output $\mathbf{w}_{\text{ls}}$ of an ESN, due to the perturbations $\delta\mathbf{u}$ (and therefore $\delta\mathbf{y}$) and $\delta\mathbf{X}$.

- Problem 3: The determination of the perturbation $\delta\tilde{y}(S+k+1)$ in the predicted output $\tilde{y}(S+k+1)$,

$$\delta\tilde{y}(S+k+1) = \delta\mathbf{w}_{\rm ls}\begin{bmatrix}1\\ u(S+k)\\ \mathbf{x}(S+k)\end{bmatrix} + \mathbf{w}_{\rm ls}\begin{bmatrix}0\\ \delta u(S+k)\\ \delta\mathbf{x}(S+k)\end{bmatrix},$$

for $k = 1, 2, \ldots$, to first order, from (5).

Problem 1 is addressed by considering the Jacobian matrix of the reservoir, whose entries must be calculated from the structured form of $\delta\mathbf{X}$ in order that a meaningful estimate of the numerical condition of the reservoir be computed.

Full consideration of Problem 2 requires that the effect on $\mathbf{w}_{\rm ls}$ of the structured matrix $\delta\mathbf{X}$ and a random vector $\delta\mathbf{u}$ be included. Standard methods of error analysis of this non-linear problem assume that $\delta\mathbf{X}$ is random, which may therefore lead to a large overestimate of the numerical condition of the LS problem (1). A better estimate requires that the structured form of $\delta\mathbf{X}$ be considered, but this problem deserves a separate study because the structured form of $\delta\mathbf{X}$ must be included in the non-linear error analysis of the LS problem. It is therefore assumed in this paper that $\delta\mathbf{X} = \mathbf{0}$, and the condition of (1) with respect to a random perturbation $\delta\mathbf{y}$ is considered. In particular, a refined condition number that allows a distinction to be made between well conditioned and ill conditioned LS problems is developed and it is shown this distinction is essential for the application of regularisation to (1). This refined condition number will allow full consideration of Problem 2 in which the structured form of $\delta\mathbf{X}$ is included in the condition number of (1).

Problem 3 requires that the explicit form of $\delta\mathbf{x}(S+k)$ be considered, and the development of a refined condition number for $\tilde{y}(S+k+1)$ therefore requires that Problems 1 and 2 be addressed.

3. The condition of the LS problem

This section considers the numerical condition of the LS problem (1) with respect to a perturbation in $\mathbf{y}$, assuming $\delta\mathbf{X} = \mathbf{0}$. A refined condition number, called the effective condition number $\eta(\mathbf{X}, \mathbf{y})$, for this perturbation is developed and it allows the criterion for the application of regularisation to be established.

The condition number $\kappa(\mathbf{X})$ of $\mathbf{X}$ is independent of $\mathbf{y}$ but $\mathbf{w}_{\rm ls}$ is a function of $\mathbf{X}$ and $\mathbf{y}$, and it may therefore be an inaccurate measure of the true numerical condition of the LS problem. This issue is overcome by the effective condition number $\eta(\mathbf{X}, \mathbf{y})$, which is a function of $\mathbf{X}$ and $\mathbf{y}$, and it therefore provides more refined information than $\kappa(\mathbf{X})$ on the condition of the LS problem. An expression for it is developed in Theorem 1 and its properties are then considered.

Theorem 1. *Let $\Delta\mathbf{y}$ be the upper bound of the relative error in $\mathbf{y}$, and let $\Delta\mathbf{w}_{\rm ls}$ be the maximum relative error in $\mathbf{w}_{\rm ls}$ due to the perturbation $\delta\mathbf{y}$,*

$$\frac{\|\delta\mathbf{y}\|}{\|\mathbf{y}\|} \leqslant \Delta\mathbf{y} \qquad \text{and} \qquad \Delta\mathbf{w}_{\rm ls} = \max_{\delta\mathbf{y}\in\mathbb{R}^S}\frac{\|\delta\mathbf{w}_{\rm ls}\|}{\|\mathbf{w}_{\rm ls}\|}.$$

Then

$$\Delta\mathbf{w}_{\rm ls} = \left(\frac{\|\mathbf{y}\|\,\|\mathbf{X}^{\dagger}\|}{\|\mathbf{w}_{\rm ls}\|}\right)\Delta\mathbf{y} = \left(\frac{\|\mathbf{y}\|\,\|\mathbf{X}^{\dagger}\|}{\|\mathbf{y}\mathbf{X}^{\dagger}\|}\right)\Delta\mathbf{y} = \eta(\mathbf{X},\,\mathbf{y})\,\Delta\mathbf{y}, \tag{8}$$

where $\eta(\mathbf{X}, \mathbf{y})$ is the effective condition number.

Proof. It follows from (1) that

$$\|\delta\mathbf{w}_{\rm ls}\| \leqslant \|\delta\mathbf{y}\|\,\|\mathbf{X}^{\dagger}\|,$$

and (8) follows from the division of this inequality by $\|\mathbf{w}_{\rm ls}\| = \|\mathbf{y}\mathbf{X}^{\dagger}\|$. □

The relationship between $\kappa(\mathbf{X})$ and $\eta(\mathbf{X}, \mathbf{y})$ is considered in the next section. This relationship also addresses one of the aims of this paper stated in section 1, specifically, the relative magnitudes of $N+2$ and S.

3.1. The magnitudes of $\kappa(\mathbf{X})$ and $\eta(\mathbf{X}, \mathbf{y})$

Let $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{X}$, and thus it follows from (8) that

$$\eta(\mathbf{X},\ \mathbf{y}) = \frac{\|\mathbf{y}\|}{\sigma_r\|\mathbf{y}\mathbf{V}\boldsymbol{\Sigma}^{\dagger}\mathbf{U}^T\|} = \frac{\|\mathbf{d}\|}{\sigma_r\|\mathbf{d}\boldsymbol{\Sigma}^{\dagger}\|} = \frac{1}{\sigma_r}\left(\frac{\sum_{i=1}^{S} d_i^2}{\sum_{i=1}^{r}\left(\frac{d_i}{\sigma_i}\right)^2}\right)^{\frac{1}{2}},$$

where $\sigma_i,\ i = 1,\ldots,r = \min\,(N + 2,\ S)$, are the singular values of $\mathbf{X}$, $\mathbf{d} = \{d_i\}_{i=1}^{S} = \mathbf{y}\mathbf{V} \in \mathbb{R}^S$, and thus

$$\eta(\mathbf{X},\ \mathbf{y}) = \frac{\sigma_1}{\sigma_r}\left(\frac{\sum_{i=1}^{S} d_i^2}{\sum_{i=1}^{r} d_i^2\left(\frac{\sigma_1}{\sigma_i}\right)^2}\right)^{\frac{1}{2}} \leqslant \kappa(\mathbf{X})\left(\frac{\sum_{i=1}^{S} d_i^2}{\sum_{i=1}^{r} d_i^2}\right)^{\frac{1}{2}}. \tag{9}$$

It is convenient to partition $\mathbf{d}$ and $\mathbf{V}$ as

$$\begin{aligned} \mathbf{d} &= [\mathbf{d}_1\ \ \mathbf{d}_2],\quad \mathbf{d}_1 \in \mathbb{R}^r,\quad\ \ \mathbf{d}_2 \in \mathbb{R}^{S-r}, \\ \mathbf{V} &= [\mathbf{V}_1\ \ \mathbf{V}_2],\ \ \mathbf{V}_1 \in \mathbb{R}^{S\times r},\ \ \mathbf{V}_2 \in \mathbb{R}^{S\times(S-r)}, \end{aligned}$$

where

$$\mathbf{V}_1^T\mathbf{V}_1 = \mathbf{I}_r,\quad \mathbf{V}_2^T\mathbf{V}_2 = \mathbf{I}_{S-r},\quad \mathbf{V}_1^T\mathbf{V}_2 = \mathbf{0}_{r,S-r},\quad \mathbf{V}_2^T\mathbf{V}_1 = \mathbf{0}_{S-r,r},$$

and thus

$$\frac{\sum_{i=1}^{S} d_i^2}{\sum_{i=1}^{r} d_i^2} = \frac{\mathbf{d}\mathbf{d}^T}{\mathbf{d}_1\mathbf{d}_1^T} = \frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}\mathbf{V}_1\mathbf{V}_1^T\mathbf{y}^T}.$$

It follows from (9) that

$$\eta(\mathbf{X},\ \mathbf{y}) \leqslant \kappa(\mathbf{X})\left(\frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}\mathbf{V}_1\mathbf{V}_1^T\mathbf{y}^T}\right)^{\frac{1}{2}} = \kappa(\mathbf{X})\left(\frac{\|\mathbf{y}\|}{\|\mathbf{y}\mathbf{V}_1\|}\right), \tag{10}$$

and Theorem 2 shows that the rows of $\mathbf{V}_1^T$ form an orthonormal basis for the row space $\mathcal{R}(\mathbf{X})$ of $\mathbf{X}$. This theorem is used in Theorem 3 to establish a geometric interpretation of the term $\mathbf{y}\mathbf{V}_1\mathbf{V}_1^T$, which allows the condition for which $\eta(\mathbf{X}, \mathbf{y}) \to \infty$ to be derived. In particular, it follows that even if $\kappa(\mathbf{X})$ is finite, the refined condition number $\eta(\mathbf{X}, \mathbf{y})$ may be infinite.

Theorems 2 and 3 are stated in terms of an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank $r \leqslant \min(m,\ n)$ because of their generality.

Theorem 2. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be an arbitrary matrix of rank $r \leqslant \min(m,\ n)$ and let $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ be its SVD. The first* r *rows of $\boldsymbol{V}^{\mathrm{T}}$ form an orthonormal basis for the row space $\mathcal{R}(\mathbf{A})$ of $\boldsymbol{A}$.*

Proof. If $\mathbf{t} \in \mathbb{R}^n$ lies in the row space of $\mathbf{A}$, there exists a vector $\mathbf{x} \in \mathbb{R}^m$ such that $\mathbf{t} = \mathbf{x}\mathbf{A}$, and thus

$$\mathbf{t} = \mathbf{x}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{x}[\mathbf{U}_1\ \ \mathbf{U}_2]\begin{bmatrix}\boldsymbol{\Sigma}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T\\ \mathbf{V}_2^T\end{bmatrix},$$

where $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{r\times r}$ is the diagonal matrix of the singular values of $\mathbf{A}$, $\mathbf{U}_1 \in \mathbb{R}^{m\times r}$, $\mathbf{U}_2 \in \mathbb{R}^{m\times(m-r)}$, $\mathbf{V}_1^T \in \mathbb{R}^{r\times n}$ and $\mathbf{V}_2^T \in \mathbb{R}^{(n-r)\times n}$. It follows that

$$\mathbf{t} = \mathbf{x}\mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^T = \mathbf{y}\mathbf{V}_1^T,\qquad \mathbf{y} = \mathbf{x}\mathbf{U}_1\boldsymbol{\Sigma}_1 \in \mathbb{R}^r,$$

and thus $\mathbf{t}$ is a linear combination of the rows of $\mathbf{V}_1^T$. It follows that the first r rows of $\mathbf{V}^T$ define an orthonormal basis for the row space of $\mathbf{A}$. □

Theorem 3. *Let $\mathbf{b} \in \mathbb{R}^n$, and let $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ be the components of $\boldsymbol{b}$ that lie in $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})^{\perp}$, respectively. Then*

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2,\quad \mathbf{b}_1 = \mathbf{b}\mathbf{V}_1\mathbf{V}_1^T \in \mathcal{R}(\mathbf{A}),\quad \mathbf{b}_2 = \mathbf{b}\mathbf{V}_2\mathbf{V}_2^T \in \mathcal{R}(\mathbf{A})^{\perp}, \tag{11}$$

where

$$\begin{aligned} \mathbf{V} &= [\mathbf{V}_1\ \ \mathbf{V}_2],\qquad \mathbf{V}_1 \in \mathbb{R}^{n\times r},\qquad \mathbf{V}_2 \in \mathbb{R}^{n\times(n-r)}, \\ \mathbf{V}_1^T\mathbf{V}_1 &= \mathbf{I}_r,\quad \mathbf{V}_2^T\mathbf{V}_2 = \mathbf{I}_{n-r},\quad \mathbf{V}_1^T\mathbf{V}_2 = \mathbf{0}_{r,n-r},\quad \mathbf{V}_2^T\mathbf{V}_1 = \mathbf{0}_{n-r,r}, \end{aligned} \tag{12}$$

and

$$\left\|\mathbf{b}\mathbf{V}_1\mathbf{V}_1^T\right\| = \|\mathbf{b}\mathbf{V}_1\| \qquad \text{and} \qquad \left\|\mathbf{b}\mathbf{V}_2\mathbf{V}_2^T\right\| = \|\mathbf{b}\mathbf{V}_2\|. \tag{13}$$

Proof. Since the rows of $\mathbf{V}_1^T$ and $\mathbf{V}_2^T$ form orthonormal bases for $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})^\perp$ respectively, there exist vectors $\mathbf{a}_1 \in \mathbb{R}^r$ and $\mathbf{a}_2 \in \mathbb{R}^{n-r}$ such that

$$\mathbf{b} = \mathbf{a}_1\mathbf{V}_1^T + \mathbf{a}_2\mathbf{V}_2^T.$$

It follows from (12) that the multiplication of this equation by $\mathbf{V}_1$ and then $\mathbf{V}_2$ yields $\mathbf{a}_1 = \mathbf{b}\mathbf{V}_1$ and $\mathbf{a}_2 = \mathbf{b}\mathbf{V}_2$ respectively, and thus

$$\mathbf{b} = \mathbf{b}\mathbf{V}_1\mathbf{V}_1^T + \mathbf{b}\mathbf{V}_2\mathbf{V}_2^T,$$

which establishes the result (11). □

The application of Theorems 2 and 3 to the LS problem (1) yields an inequality between $\eta(\mathbf{X}, \mathbf{y})$ and $\kappa(\mathbf{X})$, which follows from (10) and (13),

$$\eta(\mathbf{X}, \mathbf{y}) \leqslant \frac{\kappa(\mathbf{X})}{\cos\theta}, \qquad \cos\theta = \frac{\|\mathbf{y}\mathbf{V}_1\|}{\|\mathbf{y}\|}, \tag{14}$$

where, from (11) and (13), θ is the angle between $\mathbf{y}$ and its component that lies in $\mathcal{R}(\mathbf{X})$. This angle also arises in the expression for the square of the error r in the solution of the LS problem,

$$r^2 = \frac{\|\mathbf{w}_{\text{ls}}\mathbf{X} - \mathbf{y}\|^2}{\|\mathbf{y}\|^2} = 1 - \cos^2\theta.$$

Equation (14) and this expression for the square of the error address one of the aims of this paper that are stated in section 1. The relative magnitudes of $N+2$ and S determine the properties of $\mathbf{w}_{\text{ls}}$ because the conditions $N+2 < S$ and $N+2 > S$ lead to an overdetermined and an underdetermined LS problem, respectively. The condition $N+2 < S$ is proposed in [5, section 4] and [13, section 3], and the optimal value of N attains a compromise between the bias and variance of the model, such that a larger value of N can be used if the data have little or no noise. The condition $N+2 > S$ is proposed in [4, section 3] and [8, section 2.2] because the data $\{u_i\}_{i=1}^S$ are mapped to a higher dimensional space in which linear methods, for example, linear regression, can be used. The condition $N+2 < S$ yields $\cos\theta < 1$ and $r > 0$, but the condition $N+2 \geqslant S$ is preferred because it follows that $\cos\theta = 1$ and $r = 0$, and it therefore yields a lower upper bound for $\eta(\mathbf{X}, \mathbf{y})$, and the error in $\mathbf{w}_{\text{ls}}$ is zero.

The limiting cases are:

- $\cos\theta = 1$: $\mathbf{y} \in \mathcal{R}(\mathbf{X})$, and thus $\eta(\mathbf{X}, \mathbf{y}) \leqslant \kappa(\mathbf{X})$ and $r = 0$.
- $\cos\theta = 0$: $\mathbf{y} \in \mathcal{R}(\mathbf{X})^\perp$, and thus $\eta(\mathbf{X}, \mathbf{y}) \to \infty$ and $r = 1$.

Example 1. Let $\mathbf{X}$ and $\mathbf{y}$ be

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{y} = [-1 \;\; 1 \;\; 1].$$

The condition number of $\mathbf{X}$ is $\kappa(\mathbf{X}) = \sqrt{3}$ and it may therefore be thought that $\mathbf{w}_{\text{ls}}$ is stable with respect to a perturbation in $\mathbf{y}$. This is incorrect because $\mathbf{y} \in \mathcal{R}(\mathbf{X})^\perp$,

$$\mathbf{y}\mathbf{X}^T = [0 \;\; 0],$$

and thus $\mathbf{w}_{\text{ls}} = \mathbf{y}\mathbf{X}^\dagger = \mathbf{y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} = \mathbf{0}$, and hence $\eta(\mathbf{X}, \mathbf{y}) \to \infty$. □

3.2. The geometry of the effective condition number

This section considers the geometric conditions, in terms of the spaces spanned by the rows of $\mathbf{U}^T$ and $\mathbf{V}^T$ where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ is the SVD of $\mathbf{X}$, for which $\eta(\mathbf{X}, \mathbf{y})$ attains its minimum and maximum values. It is assumed, for simplicity, that $r = N+2 = S$, and thus the effective condition number of the LS problem is

$$\eta(\mathbf{X}, \mathbf{y}) = \frac{\|\mathbf{y}\|\,\|\mathbf{X}^{-1}\|}{\|\mathbf{w}_{\text{ls}}\|} = \frac{\|\mathbf{d}\|}{\sigma_r\|\mathbf{d}\boldsymbol{\Sigma}^{-1}\|}, \qquad \mathbf{d} = \mathbf{y}\mathbf{V}.$$

- Ill conditioned LS problem: $\eta(\mathbf{X}, \mathbf{y}) \approx \kappa(\mathbf{X}) = \sigma_1/\sigma_r \gg 1$
  $\|\mathbf{d}\|$ is dominated by the first few components of $\mathbf{d}$ such that

$$\frac{|d_i|}{\sigma_i} \to 0 \qquad \text{as} \qquad i \to r. \tag{15}$$

This condition is called the discrete Picard condition and it requires that the constants $|d_i|$ decay to zero faster than the singular values σ_i decay to zero [12]. It is shown in section 4 that the application of

regularisation to (1) requires that this condition be satisfied.

The dominant components of $\mathbf{y}$ lie in the first few rows of $\mathbf{V}^T$.

The dominant components of $\mathbf{w}_{\rm ls}$ lie in the first few rows of $\mathbf{U}^T$.

- Well conditioned LS problem: $\eta(\mathbf{X},\mathbf{y}) \approx 1$

  $\|\mathbf{d}\|$ is dominated by the last few components of $\mathbf{d}$ such that

$$\|\mathbf{d}\| \approx |d_r| \quad \text{and} \qquad \|\mathbf{d}\boldsymbol{\Sigma}^{-1}\| \approx \frac{|d_r|}{\sigma_r}.$$

The dominant components of $\mathbf{y}$ lie in the last few rows of $\mathbf{V}^T$.

The dominant components of $\mathbf{w}_{\rm ls}$ lie in the last few rows of $\mathbf{U}^T$.

It follows that the value of the effective condition number $\eta(\mathbf{X},\mathbf{y})$ is defined by the space spanned by the rows of $\mathbf{V}^T$ in which the dominant components of $\mathbf{y}$ lie.

4. Regularisation

Regularisation imposes stability on the solution of an ill conditioned set of linear algebraic equations, and its application to (1) is justified by the claim that it reduces overfitting [4, section 3], [5, section 4.2], [8, section 8.1.1]. The application of regularisation to (1) leads to the minimisation (6) whose solution is, assuming $N+2 \geqslant S$,

$$\mathbf{w}(\lambda) = \mathbf{y}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T = \sum_{i=1}^{S}\left(\left(\frac{d_i}{\sigma_i}\right)\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)\right)\mathbf{u}_i^T, \tag{16}$$

where $\mathbf{d} = \{d_i\}_{i=1}^S = \mathbf{y}\mathbf{V}$, the SVD of $\mathbf{X}$ is $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, the singular values of $\mathbf{X}$ are σ_i, $i = 1, \ldots, S$, and $\mathbf{u}_i^T$ is the ith row of $\mathbf{U}^T$. Regularisation assumes there are errors in $\mathbf{y}$ only and that $\mathbf{X}$ is exact. Its application requires that the discrete Picard condition (15) be satisfied, and it is based on a trade-off between the error and stability of $\mathbf{w}(\lambda)$:

- There exists an optimal value $\lambda^{\rm opt}$ of λ such that (i) the error in $\mathbf{w}(\lambda^{\rm opt})$ with respect to the theoretically exact solution $\mathbf{w}(0)$ of the LS problem is small, and (ii) $\mathbf{w}(\lambda^{\rm opt})$ is much more stable than $\mathbf{w}(0)$ with respect to a perturbation in $\mathbf{y}$.

This trade-off is the bias-variance trade-off in machine learning in which the bias and variance are, respectively, the error and stability of $\mathbf{w}(\lambda)$. The trade-off is satisfied if the discrete Picard condition (15) is satisfied:

- It is shown in [23, section 5.2] that $\mathbf{w}(\lambda^{\rm opt})$ is much more stable than $\mathbf{w}_{\rm ls} = \mathbf{w}(0)$ if (15) is satisfied.
- It is shown in [23, theorem 5.1] that the error in $\mathbf{w}(\lambda^{\rm opt})$ with respect to the theoretically exact solution of the LS problem is small if (15) is satisfied. If, however, (15) is not satisfied, then the error in the regularised solution is large and it increases as λ increases. It follows that regularisation is not benign and it must therefore be verified that the discrete Picard condition (15) is satisfied before it is applied.

Regularisation requires that the value of $\lambda^{\rm opt}$ be determined, and this issue is addressed in the next section.

4.1. The calculation of $\lambda^{\rm opt}$

This section considers the calculation of $\lambda^{\rm opt}$ using the GCV [19, 20] and the L-curve [21, section 4.6]. If $\eta(\mathbf{X},\mathbf{y}) \approx 1$, the LS problem is well conditioned and thus $\lambda^{\rm opt} = 0$, and it is therefore assumed in this section that the discrete Picard condition (15) is satisfied.

The L-curve, which is shown in figure 2, is a parametric plot of the magnitude $m(\lambda) = \log_{10}\|\mathbf{w}(\lambda)\|$ against the residual $r(\lambda) = \log_{10}\|\mathbf{w}(\lambda)\mathbf{X} - \mathbf{y}\|$, where $\mathbf{w}(\lambda)$ is defined in (16). The magnitude $m(\lambda)$ decreases and the residual $r(\lambda)$ is approximately constant as λ increases from zero to its value in the corner of the L. As λ increases from this value, $m(\lambda)$ is approximately constant and $r(\lambda)$ increases. The optimal value $\lambda^{\rm opt}$ of λ minimises, approximately, $m(\lambda)$ and $r(\lambda)$ simultaneously, and thus $\lambda^{\rm opt}$ is the value of λ in the corner of the L. It follows that $\lambda = \lambda^{\rm opt}$ balances the residual $r(\lambda)$ of the model and the satisfaction of the constraint on $m(\lambda)$.

The GCV is based on the principle that if an arbitrary component of $\mathbf{y}$ is deleted, an error in the predicted value of the deleted component by the regularised solution should be small. The optimal value of λ computed by the GCV minimises a function, but this minimum may be flat, which causes numerical difficulties. The GCV

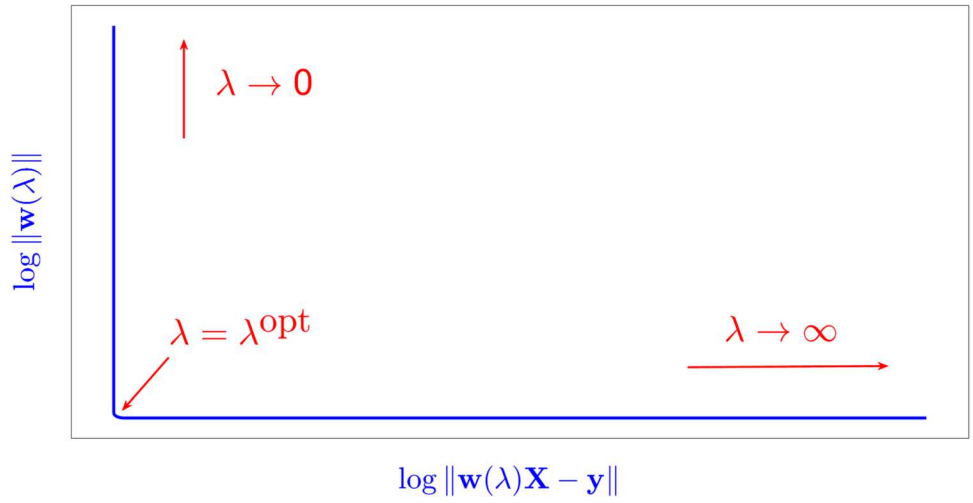

Figure 2. The L-curve when the discrete Picard condition is satisfied.

may also return a very small value of λ^{opt}, and other issues associated with the use of the GCV for computing the value of λ^{opt} are in [21, section 7.7].

5. Examples

This section contains examples that illustrate the theory in the previous sections. The determination of the requirement, or otherwise, to impose regularisation, as shown in (16), requires that the data are exact, and thus $\delta\mathbf{y} = \mathbf{0}$. The examples consider the dependence of the stability of $\mathbf{w}_{\text{ls}}$ on the parameters of the reservoir:

- The connections in the reservoir: Experiments show that the numerical condition of the LS problem is weakly dependent on the number of connections in the reservoir, that is, the sparsity of $\mathbf{W}$.
- The spectral radius $\mu(\mathbf{W})$ of $\mathbf{W}$: The condition $\mu(\mathbf{W}) < 1$ is often imposed because it guarantees satisfaction of the echo state property (ESP) [8, section 5.1]. The ESP states that the effect of a state $\mathbf{x}(n)$ of the reservoir and input sample $u(n)$ on a state $\mathbf{x}(n+k)$ of the reservoir decreases as $k \to \infty$. It can also be satisfied if $\mu(\mathbf{W}) > 1$ and thus the condition $\mu(\mathbf{W}) < 1$ is not necessary for the satisfaction of the ESP.
- The leakage α: Equation (2) shows that the contribution to $\mathbf{x}(n)$ of the data sample $u(n)$ decreases as $\alpha \to 0$. Experiments show that the condition number of the LS problem increases as α decreases.

The output of an ESN is dependent on many parameters that include, apart from the parameters listed above, the dimensions of $\mathbf{W}$, the number of neurons in the reservoir, the random seed for the computation of $\mathbf{W}$, and the activation function. A study that considers the effects of all these parameters on the performance of an ESN is necessarily large and cannot, therefore, be included in one paper. It is not the aim of this paper to determine optimal values for these parameters, but rather to consider the application of regularisation for time series analysis by considering several examples and the consequences of varying several parameters. Typical values for the parameters whose values are constant are used in order that the results be representative of results obtained with other values of these parameters.

All predictions in the examples are one step ahead predictions (5).

Example 2. The time series $u(n)$ of a non-linear communication channel,

$$u(n) = q(n) + 0.0036q^2(n) - 0.11q^3(n),$$

was formed [16], where

$$\begin{aligned} q(n) = {} & 0.08d(n+2) - 0.12d(n+1) + d(n) + 0.18d(n-1) \\ & - 0.1d(n-2) + 0.09d(n-3) - 0.05d(n-4) + 0.04d(n-5) \\ & + 0.03d(n-6) + 0.01d(n-7), \end{aligned}$$

and $d(n)$ is a sequence of random numbers drawn from the set $\{-3, -1, 1, 3\}$.

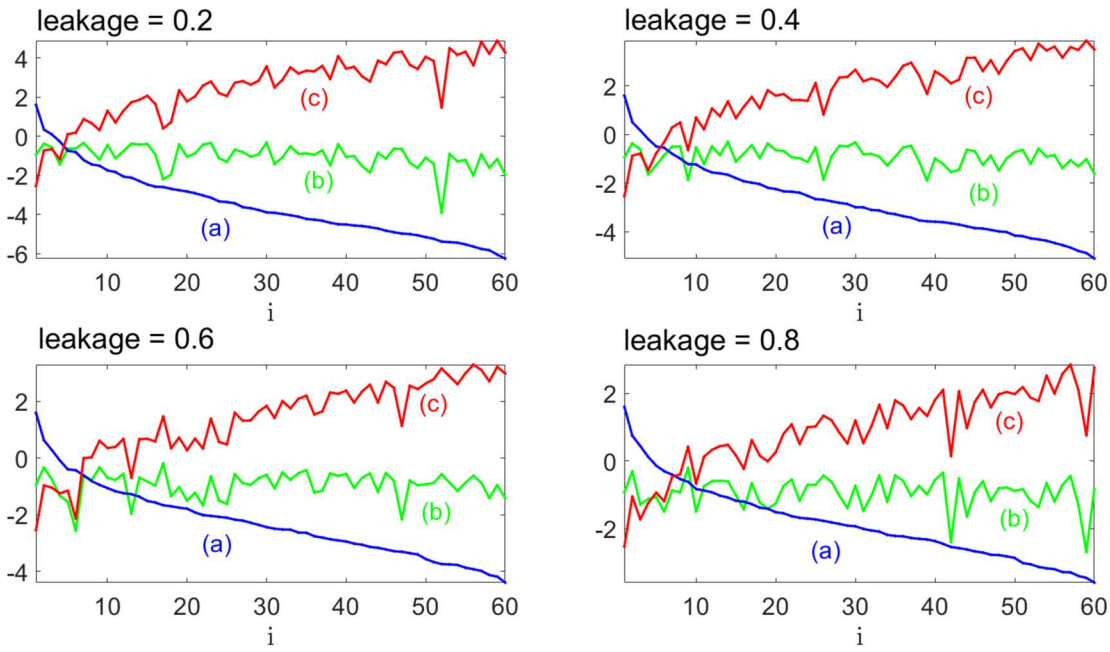

Figure 3. The variation of (a) the singular values $\log_{10}\sigma_i$ of $\mathbf{X}$, (b) the constants $\log_{10}|d_i|$, and (c) the ratios $\log_{10}|d_i|/\sigma_i$ with i, for four values of the leakage α, for Example 2.

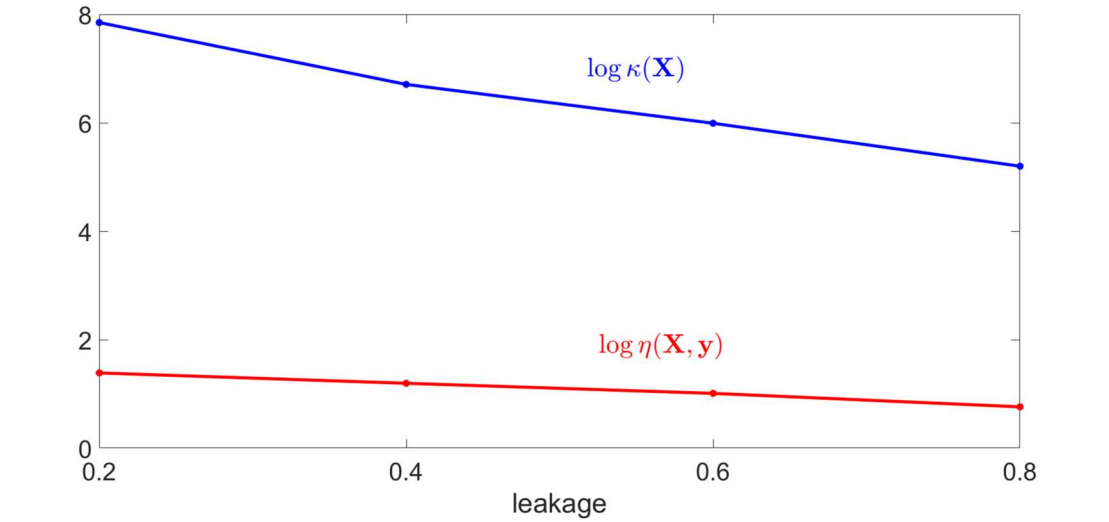

Figure 4. The effective condition number $\log_{10}\eta(\mathbf{X},\ \mathbf{y})$ and the condition number $\log_{10}\kappa(\mathbf{X})$ for four values of the leakage α, for Example 2.

The reservoir matrix $\mathbf{X}$ is 100×60, the sparsity of $\mathbf{W}$ is 20% and it is normalised by $\mu(\mathbf{W})$, the spectral radius of $\mathbf{W}$. Figure 3 shows the singular values σ_i of $\mathbf{X}$, the constants $|d_i|$ and the ratios $|d_i|/\sigma_i$ for four values of the leakage α. The constants $|d_i|$ are approximately constant and the dominant components of $|d_i|/\sigma_i$ are associated with large values of i, that is, the small singular values of $\mathbf{X}$. The regularised solution $\mathbf{w}(\lambda)$ is obtained by the deletion of the small singular values from $\mathbf{w}(0) = \mathbf{w}_{\rm ls}$ and thus regularisation leads to a large error in $\mathbf{w}(\lambda)$. Figure 4 shows the variation of the effective condition number $\log_{10}\eta(\mathbf{X},\ \mathbf{y})$ and the condition number $\log_{10}\kappa(\mathbf{X})$ with the leakage α. It is seen that $\kappa(\mathbf{X})$ is many orders of magnitude larger than $\eta(\mathbf{X}, \mathbf{y})$, and the large value of $\kappa(\mathbf{X})$ suggests that $\mathbf{w}_{\rm ls}$ is unstable with respect to a perturbation in $\mathbf{y}$, but the small value of $\eta(\mathbf{X}, \mathbf{y})$ shows that $\mathbf{w}_{\rm ls}$ is stable with respect to this perturbation.

Figure 5(a) shows the L-curves for four values of the leakage α. They have the same form, which is very different from the curve in figure 2. In particular, as λ increases from zero, $\|\mathbf{w}(\lambda)\|$ is constant and $\|\mathbf{w}(\lambda)\mathbf{X} - \mathbf{y}\|$ increases, and as λ increases further, $\|\mathbf{w}(\lambda)\|$ decreases rapidly and $\|\mathbf{w}(\lambda)\mathbf{X} - \mathbf{y}\|$ is constant. It follows there does not exist a value of λ that minimises simultaneously $\|\mathbf{w}(\lambda)\|$ and $\|\mathbf{w}(\lambda)\mathbf{X} - \mathbf{y}\|$, and thus

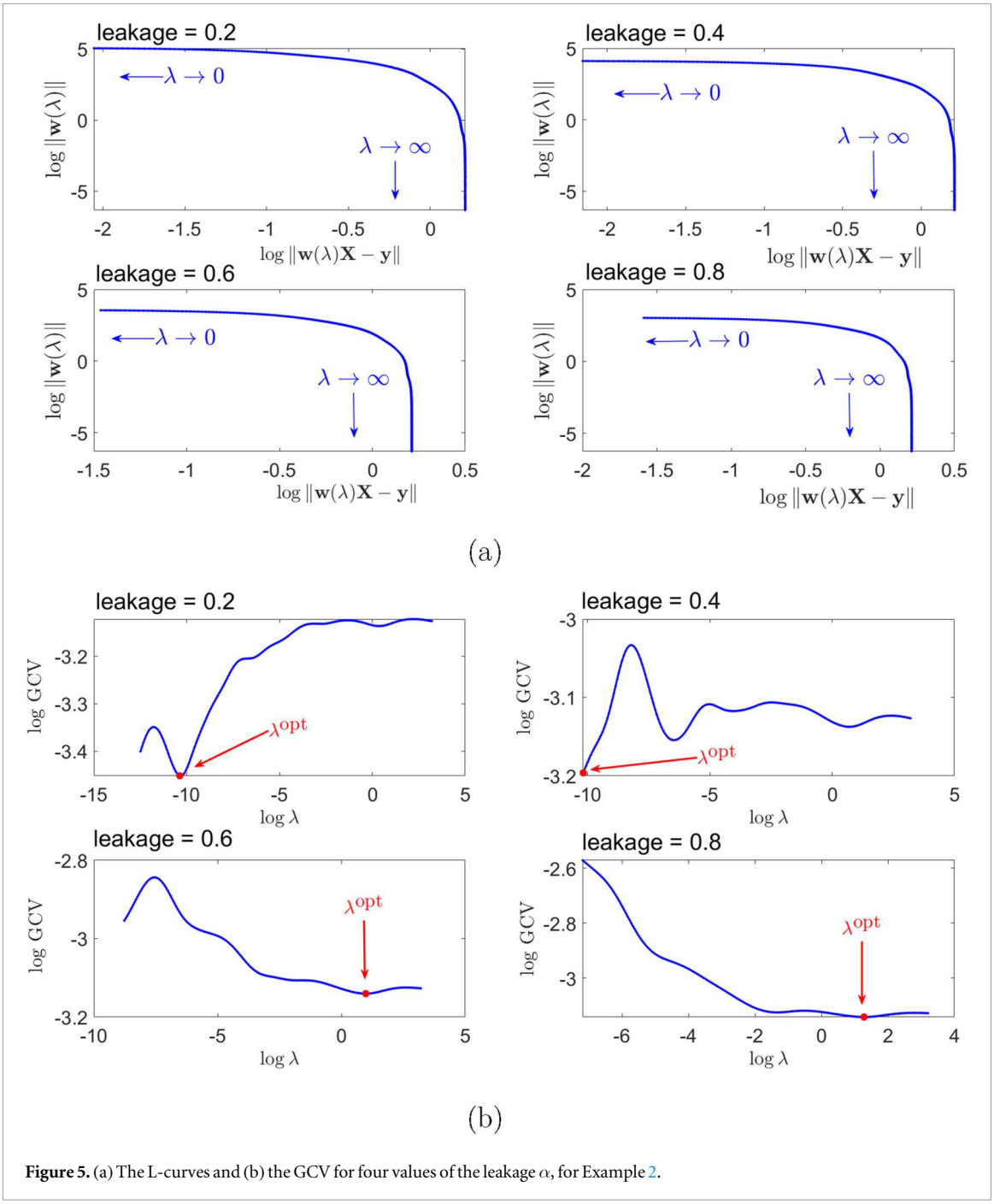

Figure 5. (a) The L-curves and (b) the GCV for four values of the leakage α, for Example 2.

$\lambda^{\text{opt}} = 0$, which is in accord with the small values of $\eta(\mathbf{X}, \mathbf{y})$ in figure 4. This absence of a minimum is in sharp contrast to the L-curve in figure 2, for which the minimum is attained in the corner of the L.

Figure 5(b) shows the GCV for the four values of α, and $\lambda^{\text{opt}} \approx 0$ for $\alpha = 0.2$ and $\alpha = 0.4$, but $\lambda^{\text{opt}} \approx 10$ for $\alpha = 0.6$ and $\alpha = 0.8$. These results suggest that regularisation is required if $\alpha = 0.6$ and $\alpha = 0.8$, which is incorrect because figures 3 and 5(a) show that $\lambda^{\text{opt}} = 0$. The L-curves in figure 5(a) are identical, but the curves of the GCV in figure 5(b) are different, which suggests that the L-curve is more effective than the GCV for determining the value of λ^{opt}. These problems with the GCV are mentioned in section 4.1. □

Example 3. The Lorenz system is a set of three first order differential equations that represent flow in three-dimensional space [14]. The Hénon map, which is a simple model that displays the same properties as the Lorenz system, is generated by the recurrence equation [24],

$$u(n) = 1 - 1.4u^2(n-1) + 0.3u(n-2) + z(n), \qquad n = 3, 4, \ldots, \tag{17}$$

where $u(1) = u(2) = 0$ and $z(n)$ is a normally distributed random variable that has zero mean and standard deviation 0.05.

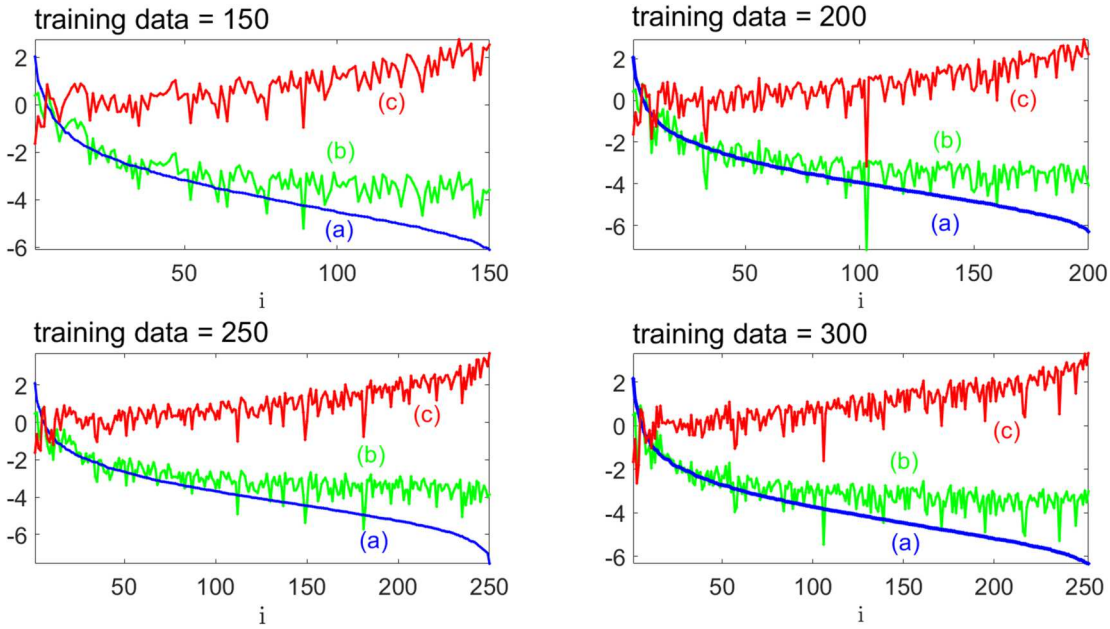

Figure 6. The variation of (a) the singular values $\log_{10}\sigma_i$ of $\mathbf{X}$, (b) the constants $\log_{10}|d_i|$, and (c) the ratios $\log_{10}|d_i|/\sigma_i$ with i, for four values of the length of the training data, for Example 3. The sparsity of $\mathbf{W}$ is 20% and it is normalised by $\mu(\mathbf{W})$.

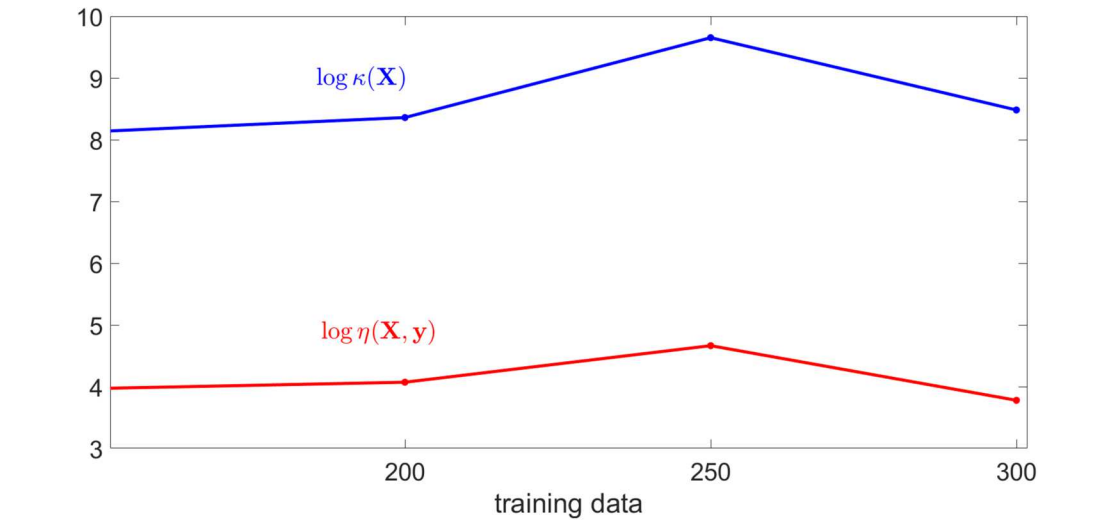

Figure 7. The effective condition number $\log_{10}\eta(\mathbf{X},\,\mathbf{y})$ and the condition number $\log_{10}\kappa(\mathbf{X})$, for four values of the length of the training data, for Example 3. The sparsity of $\mathbf{W}$ is 20% and it is normalised by $\mu(\mathbf{W})$.

The reservoir has 250 neurons, $\alpha = 0.2$, $\mathbf{W}$ is 20% sparse and normalised by $\mu(\mathbf{W})$, and the values $\{150,\ 200,\ 250,\ 300\}$ of the length of the training data were considered. Figure 6 shows that the singular values σ_i of $\mathbf{X}$ decay to zero slightly faster than the constants $|d_i|$ decay to zero and thus the discrete Picard condition (15) is not satisfied. The effective condition number is large, $\eta(\mathbf{X},\mathbf{y}) \approx 10^4$, but much smaller than $\kappa(\mathbf{X})$, as shown in figure 7. The L-curves are shown in figure 8(a) and each curve has a corner, albeit badly defined, at $\log_{10}\|\mathbf{w}(\lambda)\mathbf{X} - \mathbf{y}\| \approx -2$. Figure 8(b) shows that the GCV for the four values of the length of the training data is small, which is consistent with figure 6 because the discrete Picard condition is not satisfied.

The experiment was repeated, but $\mathbf{W}$ was dense and it was not normalised by $\mu(\mathbf{W})$. Figure 9 shows the variation of the singular values $\log_{10}\sigma_i$ of $\mathbf{X}$, the constants $\log_{10}|d_i|$, and the ratios $\log_{10}|d_i|/\sigma_i$ with i, for the four values of the length of the training data, and it differs from figure 6:

- The span of the singular values in figure 6 is larger than the span of the singular values in figure 9, and thus $\kappa(\mathbf{X})$ for $\mathbf{X}$ in figure 6 is much larger than $\kappa(\mathbf{X})$ for $\mathbf{X}$ in figure 9.

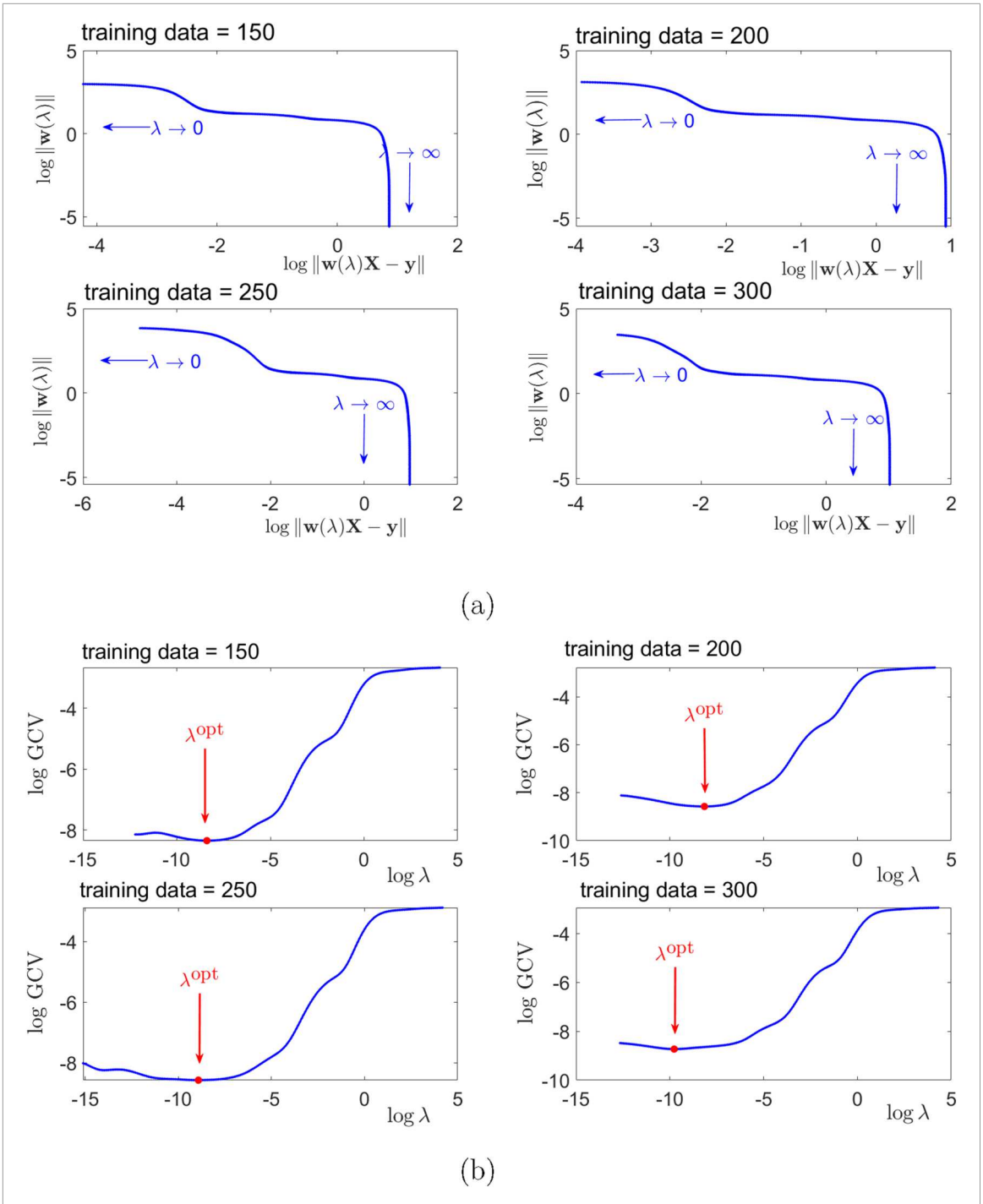

Figure 8. (a) The L-curves and (b) the GCV for four values of the length of the training data, for Example 3. The sparsity of **W** is 20% and it is normalised by $\mu(\mathbf{W})$.

- The constants $|d_i|$ span about six orders of magnitude in figure 6 and they span about three orders of magnitude in figure 9.

The condition numbers $\log_{10}\eta(\mathbf{X}, \mathbf{y})$ and $\log_{10}\kappa(\mathbf{X})$ for the four values of the length of the training data when **W** is dense and not normalised by $\mu(\mathbf{W})$ are shown in figure 10. Figures 7 and 10 suggest that a sparse form of **W** and/or its normalisation by $\mu(\mathbf{W})$ increase these condition numbers. The improved stability of **X** obtained by using the dense form of **W** and not normalising it by $\mu(\mathbf{W})$ is evident in figure 11(a), which shows the L-curve for the four values of the length of the training data. The curves show that $\lambda^{\text{opt}} = 0$, but it is difficult to draw a conclusion from figures 6 and 8(a) because figure 6 shows that regularisation must not be applied, and figure 8(a) shows that each L-curve has a corner, but the corners are badly defined. Figure 11(b) shows the GCV for the four values of the length of the training data and it is seen that $\lambda^{\text{opt}} \approx 10^{-1.5} = 0.03$. This result suggests

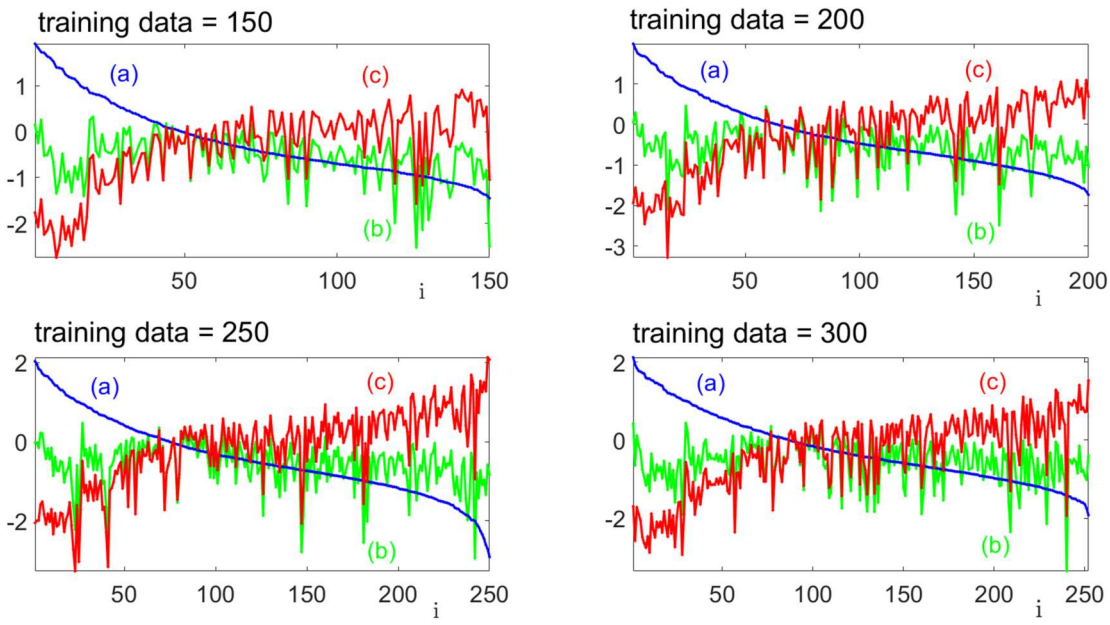

Figure 9. The variation of (a) the singular values $\log_{10}\sigma_i$ of $\mathbf{X}$, (b) the constants $\log_{10}|d_i|$, and (c) the ratios $\log_{10}|d_i|/\sigma_i$ with i, for four values of the length of the training data, for Example 3. The matrix $\mathbf{W}$ is dense and it is not normalised by $\mu(\mathbf{W})$.

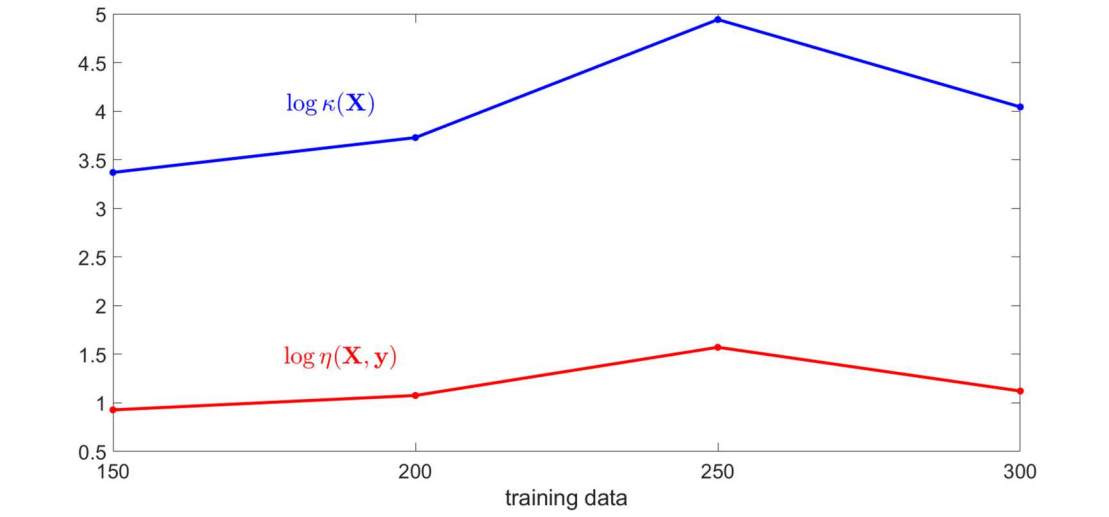

Figure 10. The effective condition number $\log_{10}\eta(\mathbf{X},\ \mathbf{y})$ and the condition number $\log_{10}\kappa(\mathbf{X})$, for four values of the length of the training data, for Example 3. The matrix $\mathbf{W}$ is dense and it is not normalised by $\mu(\mathbf{W})$.

that regularisation should be applied, but it is inconsistent with figures 9 and 11(a), which show that regularisation must not be applied. □

Example 4. Consider the series formed from the addition of several sine waves [25],

$$u(n) = \sum_{i=1}^{s} \sin(\beta_i n), \qquad n = 1,\ldots,10000, \tag{18}$$

where $s = 8$ and

$$\beta_1 = 0.20, \quad \beta_2 = 0.311, \quad \beta_3 = 0.42, \quad \beta_4 = 0.51,$$
$$\beta_5 = 0.63, \quad \beta_6 = 0.74, \quad \beta_7 = 0.85, \quad \beta_8 = 0.97.$$

The matrix $\mathbf{X}$ was of order 100×70, and the dependence of the condition numbers $\eta(\mathbf{X}, \mathbf{y})$ and $\kappa(\mathbf{X})$ on the leakage α and the sparsity of $\mathbf{W}$ were computed. Figure 12(a) shows the variation of $\log_{10}\kappa(\mathbf{X})$ with these parameters when $\mathbf{W}$ is normalised by $\mu(\mathbf{W})$. The smallest values of $\kappa(\mathbf{X})$ are larger than 10^5 and they occur when α is large, for all values of the sparsity. The matrix $\mathbf{W}$ is not normalised by $\mu(\mathbf{W})$ in figure 12(b), and the graphs

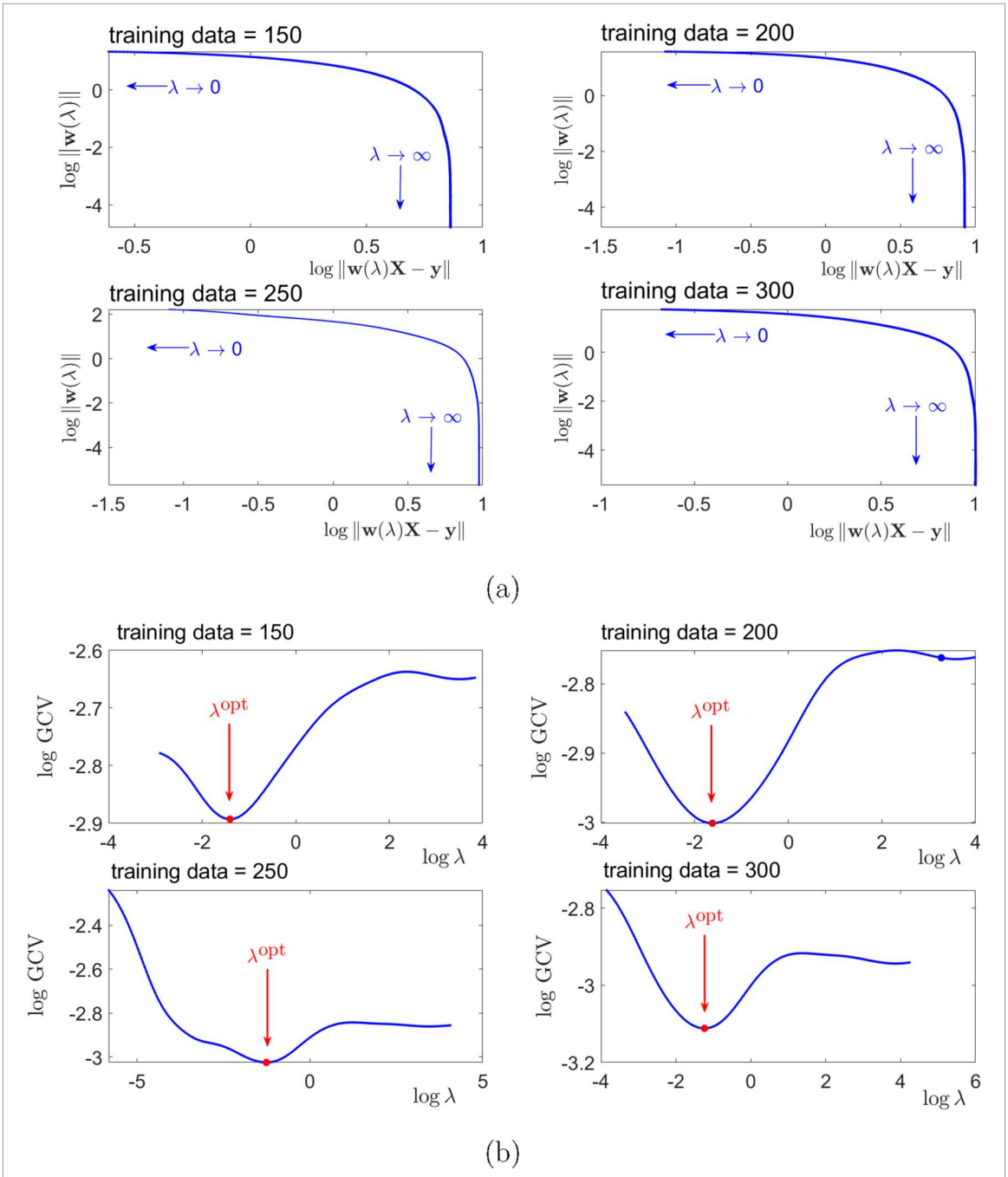

Figure 11. (a) The L-curves and (b) the GCV for four values of the length of the training data, for Example 3. The matrix **W** is dense and it is not normalised by $\mu(\mathbf{W})$.

in figures 12(a) and (b) are significantly different because the smallest values of $\kappa(\mathbf{X})$ are about 10 in figure 12(b). Figures 13(a) and (b) show the results for the effective condition number $\eta(\mathbf{X}, \mathbf{y})$ and they are similar to figures 12(a) and (b), respectively, because $\eta(\mathbf{X}, \mathbf{y})$ is, in general, smaller when **W** is not normalised by $\mu(\mathbf{W})$, and the smallest values occur when α is large and the sparsity is not large.

Figures 12 and 13 show that the stability of $\mathbf{w}_{\text{ls}}$ with respect to a perturbation in **y** increases rapidly as α increases, but this stability is weakly dependent on the sparsity of **W**. Furthermore, they show that normalisation of **W** by $\mu(\mathbf{W})$ leads to an increase, by several orders of magnitude, in the condition numbers $\eta(\mathbf{X}, \mathbf{y})$ and $\kappa(\mathbf{X})$ for large values of α. □

Examples 5, 6 and 7 differ from Examples 2, 3 and 4 because they consider the predictive ability of ESNs.

Example 5. Consider an ESN for which $\mathbf{X} \in \mathbb{R}^{100\times 40}$, $\alpha = 0.80$, and **W** is 5% sparse and normalised by $\mu(\mathbf{W})$, for the Mackey-Glass time series [15]. This series is the solution $x(t)$ of the differential equation,

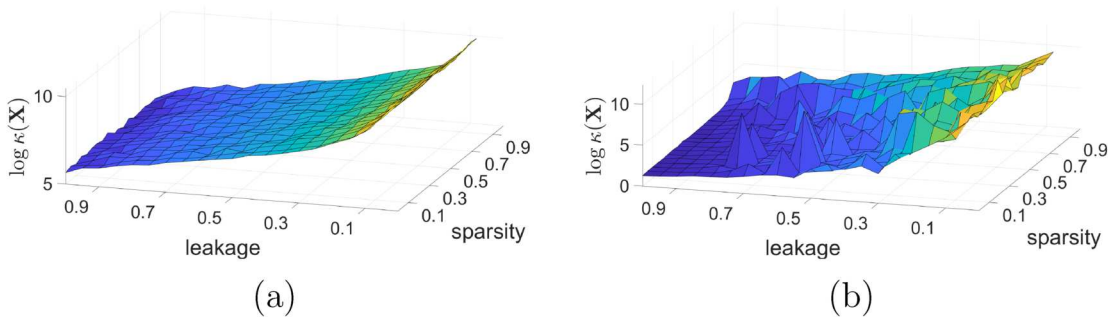

Figure 12. The variation of the condition number $\log_{10}\kappa(\mathbf{X})$ with the sparsity of $\mathbf{W}$ and leakage α when $\mathbf{W}$ is (a) normalised and (b) not normalised, by $\mu(\mathbf{W})$, for Example 4.

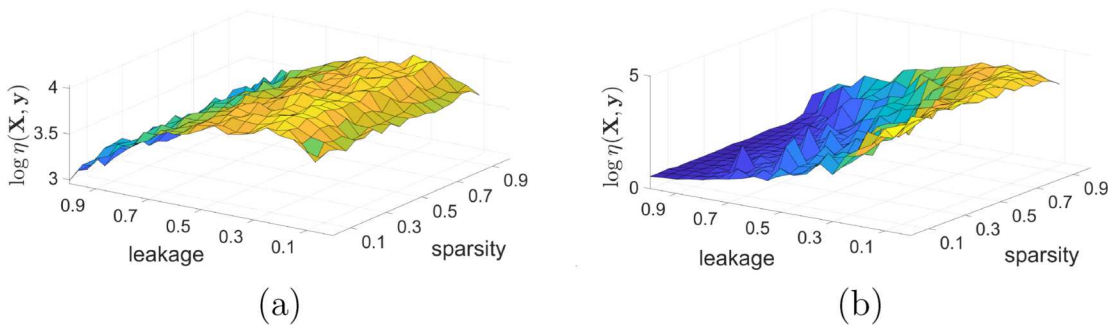

Figure 13. The variation of the effective condition number $\log_{10}\eta(\mathbf{X}, \mathbf{y})$ with the sparsity of $\mathbf{W}$ and leakage α when $\mathbf{W}$ is (a) normalised and (b) not normalised, by $\mu(\mathbf{W})$, for Example 4.

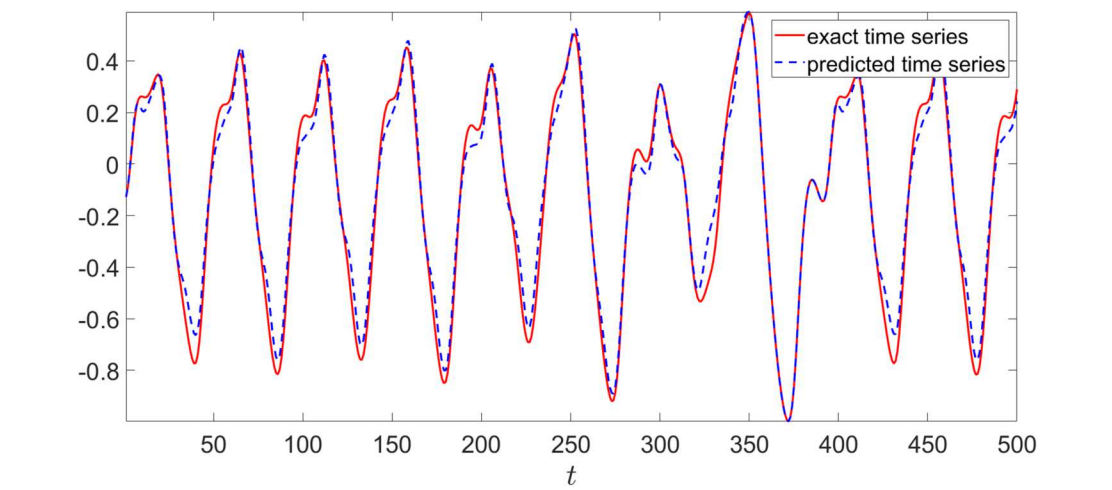

Figure 14. The Mackey-Glass time series $x(t)$, $t = 1, 2, \ldots, 500$, and its predicted form for Example 5.

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1 + x(t-\tau)^{10}} - bx(t),$$

where $a = 0.2$, $b = 0.1$ and $\tau = 17$. Figure 14 shows 500 samples of $x(t)$, $t = 1, 2, \ldots, 500,$ and the predicted time series from (5). The error in the predicted time series is small, but a bad prediction was obtained when $\mathbf{W}$ was not normalised by $\mu(\mathbf{W})$. □

Example 6. Predictions were made on the Hénon map (17) for which $\mathbf{X} \in \mathbb{R}^{120\times 40}$, $\alpha = 0.90$, and $\mathbf{W}$ is 5% sparse and normalised by $\mu(\mathbf{W})$. Figure 15 shows 300 samples $u(n)$, $n = 1, 2, \ldots, 300,$ using (5), and its predicted

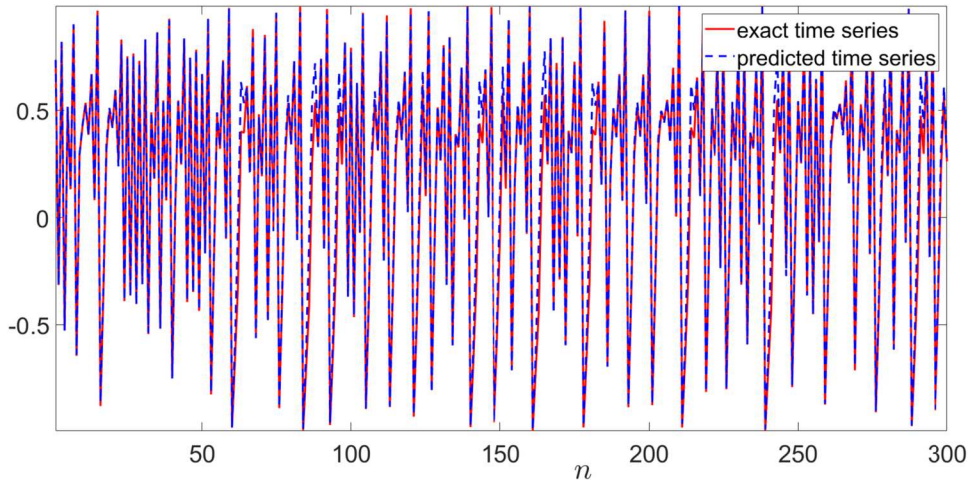

Figure 15. The Hénon map $u(n)$, $n = 1, 2, \ldots, 300$, and its predicted form for Example 6.

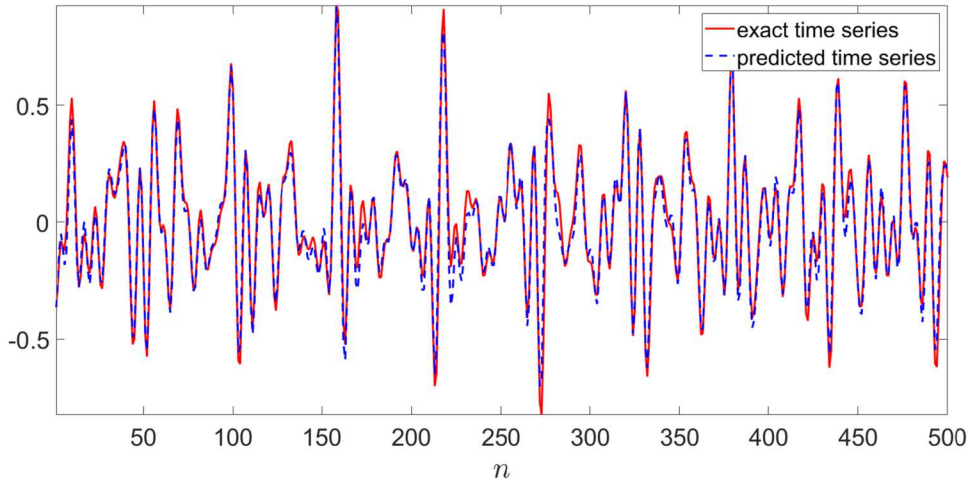

Figure 16. The time series $u(n)$, $n = 1, \ldots, 500$, defined in (18) and its predicted form for Example 7.

form, and it is seen that the error in the predicted values of the map is small. The experiment was repeated but $\mathbf{W}$ was not normalised by $\mu(\mathbf{W})$, and an unsatisfactory prediction was obtained. □

Example 7. An ESN for which $\mathbf{X}^{120\times 40}$, $\alpha = 0.70$, and $\mathbf{W}$ is 0.5% sparse and normalised by $\mu(\mathbf{W})$ was used to predict values of the time series (18), using (5). Figure 16 shows 500 samples of this time series and its predicted form. The error in the prediction is small, and as in Examples 5 and 6, bad predictions were obtained when $\mathbf{W}$ was not normalised by $\mu(\mathbf{W})$. □

6. Future work

This paper has considered methods of computational linear algebra for the determination of the numerical stability of an ESN. It is assumed in this work that $\delta\mathbf{X} = \mathbf{0}$, and thus its extension requires that this restriction be removed. This issue is addressed in [22], and the structured nature of $\mathbf{X}$, and therefore also of $\delta\mathbf{X}$, are included in this work.

It may be possible to extend the scale and class of problems that are currently considered by an ESN by combining quantum computers and ESNs, which yields quantum echo state networks (QESNs) [26, 27]. Large reservoirs in ESNs are associated with significant computational cost because of the complexity of the operations, for example, (i) the determination of the solution of the LS problem (1) when $\mathbf{X}$ is large and (ii) the

refined stability analysis of an ESN in [22], is high and it increases rapidly as the reservoir increases. This stability analysis is necessary because it provides a quantitative measure of the computational reliability of the output of an ESN. The reservoir in an ESN is replaced by a quantum system in a QESN, which allows greatly reduced execution times and thus the ability to model systems in new disciplines, for example, financial modelling, signal processing and the simulation of physical phenomena.

7. Summary

This paper has considered the application of regularisation to the LS problem in ESNs. Experiments on many time series and for a wide range of parameter values showed that regularisation should not be applied because $\mathbf{w}_{\rm ls} = \mathbf{y}\mathbf{X}^{\dagger}$ is stable with respect to a perturbation in $\mathbf{y}$. It must, however, be noted that the need, or otherwise, to apply regularisation is dependent upon many parameters of an ESN and thus each problem must be considered to determine if regularisation is necessary. This need to consider each problem is essential because the application of regularisation is not benign since its application when the discrete Picard condition is not satisfied leads to a large error between the theoretically exact and regularised solutions of the LS problem.

A refined condition number of the LS problem, called the effective condition number $\eta(\mathbf{X}, \mathbf{y})$, was introduced and it was shown that it provides a geometric interpretation of the relationships between $\mathbf{X}$ and $\mathbf{y}$ such that $\mathbf{w}_{\rm ls}$ is stable, and unstable, with respect to a perturbation in $\mathbf{y}$. It was shown that if $\mathbf{X}$ has full rank, then its condition number $\kappa(\mathbf{X})$ is finite, but $\eta(\mathbf{X}, \mathbf{y})$ may be infinite. Furthermore, there exist vectors $\mathbf{y}$ such that $\eta(\mathbf{X},\ \mathbf{y}) \approx 1$, even if $\kappa(\mathbf{X}) \gg 1$.

The paper did not consider the effect of a perturbation in $\mathbf{y}$ on the reservoir matrix $\mathbf{X}$, which also causes a perturbation in $\mathbf{w}_{\rm ls}$. A complete study of the numerical stability of an ESN must include both sources of errors, and the effect of these errors on the predictions on new data [22]. This study is essential in order that stable and unstable regimes of an ESN for the analysis of time series can be determined.

Data availability statement

The data are created to demonstrate the theory and they are fully described in the paper. The data that support the findings of this study are available upon reasonable request from the authors.

References

[1] Cabessa J, Hernault H, Kim H, Lamonato Y and Levy Y 2021 Efficient text classification with echo state networks *IEEE International Joint Conference on Neural Networks* (Shenzhen, China)
[2] Shrivastava H, Garg A, Cao Y, Zhang Y and Sainath T 2021 Echo state speech recognition *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, Canada)
[3] Viehweg J, Teutsch P and Mäder P 2025 A systematic study of echo state networks topologies for chaotic time series prediction *Neurocomputing* **618** 129032
[4] Bollt E 2021 On explaining the surprising success of reservoir computing forecaster of chaos? The universal machine learning dynamical system with contrasts to VAR and DMD *Chaos* **31** 013108
[5] Lukoševičius M 2012 A practical guide to applying echo state networks *Neural Networks: Tricks of the Trade* 2nd edn, ed G B Orr and K-R Müller (Springer) pp 659–86
[6] Aceituno P, Yan G and Liu Y-Y 2020 Tailoring echo state networks for optimal learning *iScience* **23** 101440
[7] Ehlers P, Nurdin H and Soh D 2025 arXiv:2312.15141v2
[8] Jaeger M 2009 Reservoir computing approaches to recurrent neural network training *Computer Science Review* **3** 127–49
[9] Bajaj C, McLennan L, Andeen T and Roy A 2023 Recipes for when physics fails: Recovering robust learning of physics informed neural networks *Machine Learning: Science and Technology* **4** 015013
[10] Winkler J R 2025 Overfitting, regularisation and condition estimation in regression (accepted)
[11] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2021 Understanding deep learning (still) requires rethinking generalization *Comm. ACM* **64** 107–15
[12] Hansen P C 1990 The discrete Picard condition for discrete ill–posed problems *BIT* **30** 658–72
[13] Jaeger H, Lukoševičius M, Popovici D and Siewert U 2007 Optimization and applications of echo state networks with leaky-integrator neurons *Neural Networks* **20** 335–52
[14] Lorenz E N 1963 Deterministic nonperiodic flow *J. Atmos. Sci.* **20** 130–41
[15] Mackey M C and Glass L 1977 Oscillation and chaos in physiological control systems *Science* **197** 287–9
[16] Rodan A and Tino P 2011 Minimum complexity echo state network *IEEE Trans. Neural Networks* **22** 131–44
[17] Yang C, Qiao J, Wang L and Zhu X 2019 Dynamical regularized echo state network for time series prediction *Neural Computing and Applications* **31** 6781–94
[18] Bollt E 2020 Regularized kernel machine learning for data driven forecasting of chaos *Annual Review of Chaos Theory, Bifurcations and Dynamical Systems* **9** 1–26
[19] Craven P and Wahba G 1978 Smoothing noisy data with spline functions *Numer. Math.* **31** 337–403
[20] Golub G H, Heath M and Wahba G 1979 Generalized cross-validation as a method for choosing a good ridge parameter *Technometrics* **21** 215–23
[21] Hansen P C 1998 *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* (Philadelphia: SIAM)

[22] Lu Y and Winkler J R 2025 Refined stability analysis of an echo state network (in preparation)
[23] Winkler J R and Mitrouli M 2020 Condition estimation for regression and feature selection *J. Comp. Appl. Maths.* **373** 112212
[24] Hénon M 1976 A two-dimensional mapping with a strange attractor *Comm. Math. Phys.* **50** 69–77
[25] Koryakin D, Lohmann J and Butz M 2012 Balanced echo state networks *Neural Networks* **36** 35–45
[26] Connerty E, Evans E, Angelatos G and Narayanan V 2024 arXiv:2412.07910v1
[27] Seddik S, Routaib H, Elmounadi A and El Haddadi A 2024 Enhancing African market predictions: Integrating quantum computing with echo state networks *Scientific African* **25** e02299