

Minimizing Stochastic Complexity with Ridge Regression

Antony Mizzi ^{*} , David M. Walker  and Michael Small 

Complex Systems Group, Department of Mathematics and Statistics, The University of Western Australia, Perth, WA 6009, Australia; david.walker@uwa.edu.au (D.M.W.); michael.small@uwa.edu.au (M.S.)

* Correspondence: antony.mizzi@research.uwa.edu.au

Abstract

We derive a penalty strength criterion for ridge regression using stochastic complexity, which is a refined variant of the minimum description length principle. Since stochastic complexity does not typically account for the effect of regularization on complexity, despite its ability to simplify models, we are required to make a slight modification to the underlying coding scheme. Our scheme makes use of a weighted ensemble of regularized model fits rather than a mixture of maximum likelihood estimates. Under this modification, regularization is interpreted as reducing model complexity by constraining flexibility. In the case of ridge regression, the complexity penalty term that we derive can be expressed analytically as the log determinant of the residual operator. We demonstrate the effect of this complexity penalty by fitting a linear readout to a reservoir computer, and by performing benchmark testing on publicly available datasets.

Keywords: model selection; coding theory and techniques; information theory; regularization

1. Introduction

We approach the problem of selecting a regularization strength for ridge regression with the intent of producing the simplest generalized linear model (GLM). This complexity-based approach is particularly relevant since the popularity of GLMs stems, in part, from their simplicity as a model class [1–5]. They are often used to identify and characterize relationships from data in fields like medicine [6,7], dietary science [8], economics [9,10], and sociology [11,12]. Such applications lie in accordance with Occam’s razor principle, which tells us to seek simple explanations, or in this case, simple model classes.

In these endeavours, it is common to use some form of regularization in the fitting process, like an L2 ridge penalty. In the case of ridge regularization, the penalty term stabilizes the matrix inverse under multicollinearity of input variables [3,13,14] and improves generalization to new data by reducing model variance at the expense of increased model bias [15–17]. The application of regularization also further reduces the complexity of linear models. This makes it possible to choose the regularization strength that minimizes model complexity.

In this paper, we measure complexity in terms of stochastic complexity [18], which is a refined variant of the minimum description length principle [19] (MDL). The stochastic complexity of a model is directly related to its flexibility. When used for model selection, it suggests that we should hesitate to choose a model class that fits our dataset well if it also provides good fits to unrelated datasets using the same predictors. While regularization constrains flexibility and produces simpler models, this is not directly accounted for by stochastic complexity. As such, the purpose of this paper, as well as deriving a ridge



Academic Editor: José F. F. Mendes

Received: 11 May 2026

Revised: 24 June 2026

Accepted: 26 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

strength criterion, is to demonstrate a method for determining the effect of regularization on the stochastic complexity of a model. Our method makes use of a novel interpretation of stochastic complexity's underlying coding scheme in order to directly account for the constraint imposed by regularization on model flexibility.

Our work is organized as follows. In Section 2, we provide an overview of stochastic complexity, including an explanation of the typical underlying coding scheme. In Section 3, we discuss our method for modifying this scheme to account for regularization. In Section 4, we derive an analytical form for the stochastic complexity of a linear model fitted by ridge regression using our modified scheme. We also present our ridge penalty selection criterion. In Section 5, we compare our criterion with several other information-theoretic criteria. In particular, we note that our work is closely related to that of Dwivedi et al. [20], who recently used luckiness normalized maximum likelihood [21] (LNML) for ridge strength selection. Their approach produces a different selection criterion, which stems from a different interpretation of regularization's effect on complexity. In Section 6, we discuss the effect of our modification on stochastic complexity's underlying probability distribution. We then visualize this effect using a two-dimensional toy example. In Section 7, we use our criteria to train linear readouts with the example of reservoir computing. We also perform benchmark testing in Section 8 using more than one hundred publicly available datasets from the Penn Machine Learning Benchmark (PMLB) repository [22]. In Section 9, we conclude with a discussion and summary.

2. Background into Stochastic Complexity and the Minimum Description Length Principle

The premise behind MDL is that models compress information. This is done by making use of a parametric model class to derive a coding scheme for possible data sets. The shorter the encoding of the dataset we wish to model under the scheme we derive, the better the compression achieved by the model. For example, in classical two-part MDL, the scheme is to first encode our model parameters and then to encode the errors that the model makes using these parameters. The best models, according to two-part MDL, minimize the combined lengths of their error and parameter codes.

The coding scheme assumed by stochastic complexity is more abstract. No one set of model parameters is used. Instead, a probability $P_0^*(Y; \mathcal{M}, X)$ is assigned to each potential dataset Y from the set of possibilities $Y \in \mathcal{Y}$ using a weighted ensemble of parameterizations from the same model class $\mathcal{M} : (X, \theta) \rightarrow Y$. If we denote the weighting associated with each parameterization θ in the ensemble as $\pi(\theta)$, then the probability assigned by the resulting ensemble would be equivalent to the marginal probability under π as a prior

$$P_0^*(Y; \mathcal{M}, X) = \sum_{\theta} P(Y|\theta; \mathcal{M}, X) \pi(\theta).$$

Importantly, the ensemble weightings are chosen deterministically based on the model class. This is what enables the resulting code to be deciphered and makes stochastic complexity a valid description length. Under the stochastic complexity scheme, ensemble weights $\pi(\theta)$ are chosen before the dataset Y is known in the way that minimizes regret for the worst possible scenario of Y . This regret $R_0(Y, P_0^*)$ is the difference between the length of code assigned to Y by the weighted ensemble (which we denote by $L_{sc}(Y)$) and the shortest length that could be used to encode Y with any other ensemble. Under the lower bound for efficiency given by Shannon's source coding theorem, these code lengths are negative log probabilities $L_{sc}(Y; \mathcal{M}, X) = -\log(P_0^*(Y))$. Since the best ensemble for

encoding any particular dataset Y is just its maximum likelihood estimate θ_Y^* , the worst-case regret which we seek to minimize by choice of weightings is the following

$$R_0^{\max}(Y, P_0^*) := \max_{Y \in \mathcal{Y}} (R_0(Y, P_0^*)), \quad (1)$$

$$\text{where } R_0(Y, P_0^*) := -\log(P_0^*(Y; \mathcal{M}, X)) + \log(P(Y|\theta_Y^*; \mathcal{M}, X)). \quad (2)$$

The weightings that minimize this worst-case regret are those that induce proportionality to the maximum likelihood estimate ($P_0^*(Y; \mathcal{M}, X) = \frac{1}{C_0} P(Y|\theta_Y^*; \mathcal{M}, X)$). This means that we can work directly with P_0^* without ever actually determining the optimal weightings. Under proportionality, regret is constant across all datasets $R_0(Y, P_0^*) = \log(C_0) \forall Y \in \mathcal{Y}$. The constant of proportionality C_0 must be the sum of all maximum likelihoods in order for it to induce a properly normalized distribution that integrates to one. For this reason, P_0^* is often called the normalized maximum likelihood (NML) distribution. The probability of Y under P_0^* is given as follows

$$P_0^*(Y; \mathcal{M}, X) = \frac{P(Y|\theta_Y^*; \mathcal{M}, X)}{\sum_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y'}^*; \mathcal{M}, X)}. \quad (3)$$

When we work with continuous data, and therefore an infinite set of possible datasets $\mathcal{Y}$, the ensemble weights $\pi(\theta)$ become difficult to determine. They have been shown [4,23], to approach the Jeffreys prior [24] for large datasets under certain conditions [25]. However, we are not required to determine the ensemble weights since proportionality still holds, and we can again work directly with P_0^* . In the case of continuous data, the stochastic complexity makes use of the integral version for P_0^*

$$L_{SC}(Y; \mathcal{M}, X) = -\log \frac{P(Y|\theta_Y^*; \mathcal{M}, X)}{\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y'}^*; \mathcal{M}, X) dY'}. \quad (4)$$

Notice that stochastic complexity can be broken into two terms, the first being the model's negative log likelihood. The other term measures the complexity of the parametric model class $\mathcal{M}$

$$\text{Penalty} = \log \left(\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y'}^*; \mathcal{M}, X) dY' \right),$$

and functions as the penalty term in model selection. This penalty directly measures flexibility and it lends stochastic complexity an important intuition beyond its meaning as a code length. It tells us not to trust models that fit the data well if they can also provide good fits to unrelated data using the same predictors.

3. The Effect of Regularization on Complexity

The role of model regularization on stochastic complexity can potentially be interpreted in different ways. Our interpretation is to directly consider the effect of regularization on the flexibility of the model class. This follows on from the intuitive meaning of stochastic complexity described in the previous section: that we should have more confidence in our regularized model if the regularization applied prevents the model from easily fitting unrelated datasets using the same predictors.

Rather than building a weighted ensemble of maximum likelihood estimates, we build a weighted ensemble of regularized fits. The weightings of this ensemble will depend on the regularization strength α . We select the weightings such that the probability assigned by the ensemble $P_\alpha^*(Y; \mathcal{M}, X)$ minimizes a modified form of regret in the worst case scenario. This modified regret, which we now denote by $R_\alpha(Y, P_\alpha^*)$, because of its dependence on the regularization strength α , is still the length of code that could have been saved

had the parameter fit for the dataset been known ahead of time. The only difference is that this parameter fit is now the set of regularized parameters and not the maximum likelihood estimate

$$R_{\alpha}^{\max}(Y, P_{\alpha}^*) := \max_{Y \in \mathcal{Y}} \{-\log(P_{\alpha}^*(Y; \mathcal{M}, X)) + \log(P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X))\}.$$

Assuming that the encoder and decoder agree on the regularization strength ahead of time, they can calculate the ensemble probability P_{α}^* that minimizes this modified regret in the worst case choice of the data set $Y \in \mathcal{Y}$. We demonstrate in Appendix A that the solution is to choose P_{α}^* proportional to the likelihood at the regularized parameter fit $P_{\alpha}^*(Y) \propto P(Y|\theta_{Y,\alpha}^*)$. The constant of proportionality is the normalization factor required to produce a distribution that integrates to unity. In the continuous case, the resulting NML distribution is defined as follows

$$P_{\alpha}^*(Y; \mathcal{M}, X) = \frac{P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X)}{\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) dY'}, \quad (5)$$

while in the discrete case, the integral is replaced with a sum. Because the normalization factor ensures that P_{α}^* is a valid probability mass, it follows from Kraft's inequality that a uniquely decodable scheme exists to describe the elements of $\mathcal{Y}$ with code word lengths between $-\log(P_{\alpha}^*(Y))$ and $-\log(P_{\alpha}^*(Y)) + 1$ nats (Ignoring fixed precision costs for continuous datasets). The scheme that achieves these code lengths is well defined for most model classes and types of regularization, which we discuss in Appendix B. The code lengths it assigns to datasets are given by the following

$$L_{SC}(Y; \mathcal{M}, X, \alpha) = -\log \frac{P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X)}{\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) dY'}. \quad (6)$$

One benefit of the model selection criterion that comes from this coding scheme ($L_{SC}(Y; \mathcal{M}, X, \alpha)$) is its generalizability. The criterion derived in Equation (6) is not specific to any particular model class or type of regularization. Another benefit is the fairness of the comparison it draws between regularization techniques. One technique can be compared with another based purely on how much each restricts the flexibility of the model class. The caveat, which is not unique to our approach, is that to derive an analytical value for stochastic complexity, we still need an analytical form for the model likelihood when regularization is applied.

4. Deriving a Ridge Penalty to Minimize Stochastic Complexity

In the case of regression with a known state matrix X , the parametric model class $\mathcal{M}$ is the multivariate normal distribution $\mathcal{N}(X\theta, \epsilon^2\mathbb{I})$. This is the model class for which the regression solution $\theta = \theta_Y^*$ provides the maximum likelihood estimate. From a Bayesian perspective, the ridge penalty comes from maximizing the posterior with the added assumption of a normal prior over the parameters. However, it is not strictly necessary to assume a prior over the parameters in order to calculate stochastic complexity. We only need to know the probability of each dataset $Y \in \mathcal{Y}$ given the regularized parameter fit $\theta_{Y,\alpha}^*$. According to our model class, that probability is the following

$$P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X) = (2\pi\epsilon)^{-n/2} \exp\left(-\frac{\|Y - X\theta_{Y,\alpha}^*\|^2}{2\epsilon^2}\right).$$

Fortunately, an analytical form for this probability exists. First, we can express the ridge regression solution $\theta_{Y,\alpha}^*$ in terms of the penalty α , the state matrix X and the dataset Y . It is given by $\theta_{Y,\alpha}^* = (X^T X + \alpha \mathbb{I})^{-1} X^T Y$. This means that an analytical form for the mean square error (MSE), which we denote by $\sigma_{Y,\alpha}^2$, also exists

$$\sigma_{Y,\alpha}^2 = \frac{1}{n} Y^T Q_\alpha Y, \text{ where } Q_\alpha = (M_\alpha - \mathbb{I})^T (M_\alpha - \mathbb{I})$$

$$\text{and } M_\alpha = X(X^T X + \alpha \mathbb{I})^{-1} X^T.$$

In the expression above, n refers to the length of Y . We can use the matrix form for mean square error to rewrite the NML probability $P_\alpha^*(Y)$ in Equation (5) for the context of ridge regression

$$P_\alpha^*(Y; \mathcal{M}) = \frac{(2\pi\epsilon)^{-n/2} \exp(-\frac{Y^T Q_\alpha Y}{2\epsilon^2})}{\int_{Y' \in \mathcal{Y}} (2\pi\epsilon)^{-n/2} \exp(-\frac{Y'^T Q_\alpha Y'}{2\epsilon^2}) dY'}. \quad (7)$$

The stochastic complexity then comes from taking the logarithm of this probability. The component corresponding to the complexity penalty is found by evaluating the integral in the denominator. For continuous datasets and linear models, it makes sense to take the space of possible datasets to be all real valued vectors of length n and set $\mathcal{Y}$ to $\mathbb{R}^n$

$$\log \int_{\mathbb{R}^n} (2\pi\epsilon^2)^{-n/2} \exp(-\frac{Y'^T Q_\alpha Y'}{2\epsilon^2}) dY'.$$

This integral then simplifies to the log determinant of $Q_\alpha^{-1/2}$

$$-\frac{1}{2} \log(\det(Q_\alpha)),$$

which can also be expressed in terms of the eigenvalues of the design matrix $X^T X$, denoted here by λ_i

$$-\frac{1}{2} \log(\det(Q_\alpha)) = \sum_{i=1}^m \log\left(1 + \frac{\lambda_i}{\alpha}\right).$$

This penalty describes how well we can fit various possible datasets using regression with a ridge constant of α and a gram matrix $X^T X$ that has the given spectrum $\{\lambda_1, \lambda_2, \dots, \lambda_m\}$. Lastly, we need the other component of the stochastic complexity, which is the negative log of the numerator in Equation (7), and also the log likelihood of Y under its own parameter fit $-\log P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X)$. Together, the stochastic complexity becomes the following function of the ridge penalty:

$$L_{SC}(Y; \mathcal{M}, X, \alpha) = \frac{n}{2\epsilon^2} \sigma_{Y,\alpha}^2 + \sum_{i=1}^m \log\left(1 + \frac{\lambda_i}{\alpha}\right) + \text{const}. \quad (8)$$

The constant term, which we ignore because it is irrelevant for our purpose, is $n \log(\sqrt{2\pi\epsilon})$.

The entries of the $m \times m$ gram matrix $X^T X$ scale proportionally to n , as do its eigenvalues λ_i . This means that in the asymptotic limit $n \rightarrow \infty$, the complexity penalty in Equation (8) can be written as follows

$$\sum_{i=1}^m \log\left(1 + \frac{\lambda_i}{\alpha}\right) \rightarrow \sum_{i=1}^m \log\left(\frac{\lambda_i}{\alpha}\right) = \sum_{i=1}^m \log(\lambda_i) - m \log(\alpha),$$

Which is equivalent to the Bayesian information penalty

$$\sum_{i=1}^m \log \left(1 + \frac{\lambda_i}{\alpha} \right) \text{ scales as } \mathcal{O}(1) + m \log(n).$$

This is a standard feature of MDL-derived selection criteria. We also demonstrate in Appendix C that the modified stochastic complexity in Equation (8) is universal in the expected sense [21]. This means that for any data-generating distribution from our parametric model class, the expected efficiency of our NML scheme is optimal in the asymptotic limit $n \rightarrow \infty$.

Calculating the stochastic complexity requires choosing a variance ϵ^2 for the multivariate Gaussian used in our coding scheme. This is also the case for various other ridge penalty criteria [26,27]. A typical method for doing so in the under-parameterized case ($m \ll n$) is to estimate ϵ in terms of the adjusted mean square error of the linear regression solution $\epsilon^2 = n\sigma_{Y,0}^2/(n-m)$. A more robust method, which is not limited to the under-parameterized case, is provided by Liu et al. [28]. They suggest approximating ϵ^2 by the following expression

$$\epsilon^2 = \frac{Y^T(\mathbb{I} - M_{\alpha'})Y}{n - \text{Tr}(M_{\alpha'})},$$

which makes use of the projection matrix $M_{\alpha'}$. We denote the ridge strength used in this projection matrix by α' to distinguish it from the strength α used in model training and the calculation of stochastic complexity.

Our formulation for optimizing the ridge penalty passes two sanity checks that we can construct from our understanding of ridge regression. Firstly, the optimal penalty scales correctly with the terms of the state matrix. If X is multiplied by a scalar $X \rightarrow c * X$, the optimal ridge penalty should scale proportionally to its square $\alpha \rightarrow c^2\alpha$, since this leaves the projection matrix $X(X^T X + \alpha\mathbb{I})^{-1}X^T$ unchanged. This is the case. The matrix Q_α is also invariant to the scaling $(X_{ij}, \alpha) \rightarrow (cX_{ij}, c^2\alpha)$, which means that the performance term $\frac{n}{2\epsilon}\sigma_{Y,\alpha}^2 = \frac{1}{2\epsilon}Y^T Q_\alpha Y$ and the complexity penalty $-\frac{1}{2}\log(\det(Q_\alpha))$ are both unaffected. Secondly, since the penalty term is a measure of model flexibility, it should only depend on the unique, non-zero columns in X . This is also true. Duplicate columns, or columns of zeros, add eigenvalues of zero to the spectrum and these do not contribute to the penalty term.

5. Alternative Approaches

In Section 3, we accounted for regularization by devising a coding scheme from an ensemble of regularized parameter fits. An alternative approach is to build the ensemble from maximum posterior likelihood estimates after assuming a prior distribution over parameter space. This is similar to the way that luckiness normalized maximum likelihood (LNML) assumes a so-called ‘luckiness function’ $p_{\text{luck}}(\theta)$ to derive a coding scheme which is optimized for the corresponding luckiness normalized probability density:

$$P_{\text{LNML}}(Y; \mathcal{M}, X) := \frac{\max_{\theta} [P(Y|\theta; \mathcal{M}, X)p_{\text{luck}}(\theta)]}{\int_{Y' \in \mathcal{Y}} \max_{\theta} [P(Y'|\theta; \mathcal{M}, X)p_{\text{luck}}(\theta)] dY'}. \quad (9)$$

Dwivedi et al. [20] follow this approach and choose p_{luck} to be Gaussian

$$p_{\text{luck}}(\theta) \propto \exp\left(-\frac{\alpha}{2\epsilon^2}\theta^T\theta\right).$$

This choice is meaningful because a Gaussian prior over the parameters produces a posterior likelihood which is maximal at the ridge regression solution. The numerator and

denominator therefore both measure posterior likelihoods. Their resulting luckiness normalized likelihood can be expressed as follows

$$P_{\text{Dwivedi}}(Y; \mathcal{M}, X) := \frac{P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X) p_{\text{luck}}(\theta_{Y,\alpha}^*)}{\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) p_{\text{luck}}(\theta_{Y',\alpha}^*) dY'}. \quad (10)$$

The description length associated with this probability density, which we denote by L_{Dwivedi} and provide below, is similar to our own, but includes an additional log parameter likelihood term

$$L_{\text{Dwivedi}}(Y; \mathcal{M}, X, \alpha) = \frac{n}{2\epsilon^2} \sigma_{Y,\alpha}^2 + \frac{\alpha(\theta_{Y,\alpha}^*)^T \theta_{Y,\alpha}^*}{2\epsilon^2} + \frac{1}{2} \sum_{i=1}^m \log\left(1 + \frac{\lambda_i}{\alpha}\right).$$

The integral in the denominator also evaluates to half the value of the complexity penalty that we found from the denominator in Equation (7). This can be seen from the way that the prior and posterior probability factorize

$$\exp\left(-\frac{1}{2\epsilon} Y^T Q_\alpha Y + \alpha(\theta_{Y,\alpha}^*)^T \theta_{Y,\alpha}^*\right) = \exp\left(-\frac{1}{2\epsilon} Y^T Q_\alpha^{1/2} Y\right).$$

The approach taken by Dwivedi et al. can be generalized to any regularization techniques that offer a Bayesian interpretation. The luckiness function would simply be replaced with whatever prior distribution is implicitly assumed over the parameters.

Another MDL-inspired approach taken by Silhavy et al. [29] makes use of a two-part coding scheme. They take the description length found for linear regression by Giurcaneanu et al. [30] and consider the additional cost of encoding the ridge parameter (or Lasso parameter). It should be noted that Giurcaneanu et al. found the description length for linear regression under a different framing of the model class. They encode the error variance $\sigma_{Y,\alpha=0}^2$ and then use the variance to encode the errors $Y \sim \mathcal{N}(X\theta, \sigma_{Y,\alpha=0}^2 \mathbb{I})$. As such, Silhavy et al.'s method for choosing an MDL ridge parameter is quite different than ours. They seek to minimize the following criterion

$$L_{\text{Silhavy}}(Y; \mathcal{M}, X, \alpha) := \frac{n-k}{2} \log(\sigma_{Y,\alpha}^2) + \frac{1}{2} \log(\alpha) + \frac{k}{2} \log((\theta_{Y,\alpha}^*)^T X^T X \theta_{Y,\alpha}^*) + \text{const.} \quad (11)$$

It should also be noted that the major objective of Silhavy et al.'s work was to apply MDL to multi-criteria decision analysis. This means seeking a tradeoff between model complexity and a combination of various performance measures. In this paper, we only deal with a single objective, which is model log likelihood (effectively the mean square model error).

Other information criteria, like the Akaike [31] and Bayesian [32] Information Criteria (AIC and BIC, respectively), can also be used to choose a ridge strength if the number of model parameters is replaced with the number of effective degrees of freedom, which has dependence on α [33,34]. The commonly used measure of effective degrees of freedom is the trace of the projection matrix $M_\alpha = X(X^T X + \alpha \mathbb{I})^{-1} X^T$ [35,36]. This results in the modified AIC and BIC criteria provided below

$$\text{AIC}_{\text{eff}}(Y; X, \alpha) = \frac{n}{\epsilon^2} \sigma_{Y,\alpha}^2 + 2 \sum_{i=1}^m \left(\frac{\lambda_i}{\lambda_i + \alpha} \right), \text{ and}$$

$$\text{BIC}_{\text{eff}}(Y; X, \alpha) = \frac{n}{\epsilon^2} \sigma_{Y,\alpha}^2 + \log(n) \sum_{i=1}^m \left(\frac{\lambda_i}{\lambda_i + \alpha} \right).$$

6. A Visualization of Probability Density on a Toy Problem

By modifying regret to account for regularization in Section 3, we alter the NML probability density P_α^* used to assign description lengths to datasets. This redistributes density towards the datasets whose likelihoods are reduced the least when switching from MLE parameters $P(Y|\theta_Y^*)$ to regularized fits $P(Y|\theta_{Y,\alpha}^*)$. In Figure 1, we visualize P_α^* on a toy example. We also visualize the LNML density used by Dwivedi et al. [20] for comparison. The example provided comes from fitting a linear model that maps the following 2×2 input matrix

$$X = \begin{pmatrix} -0.64 & -0.61 \\ 0.32 & 0.01 \end{pmatrix},$$

to possible output datasets (Y_1, Y_2) in $\mathbb{R}^2$. The values in this example were generated by random sampling of the unit Gaussian and truncated to two decimal places.

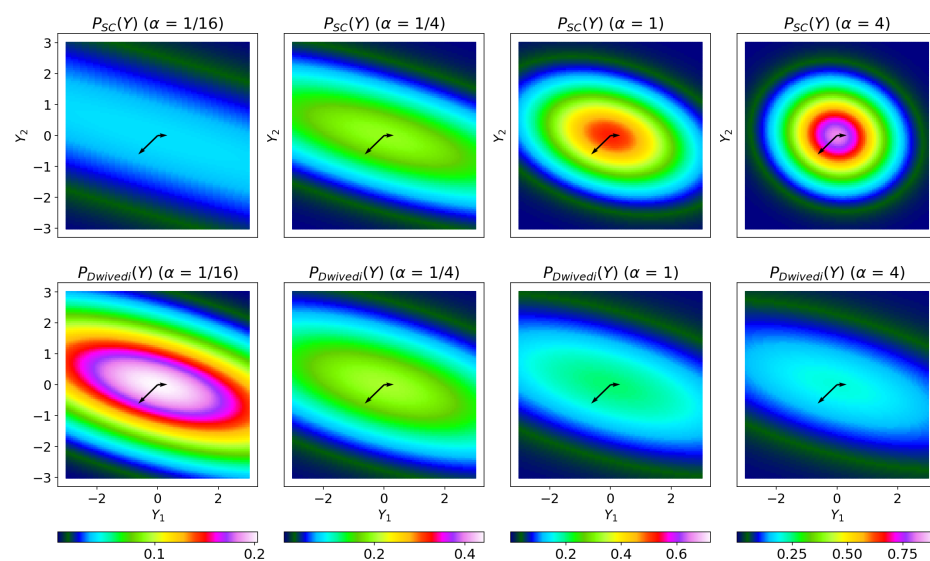


Figure 1. The probability density $P_\alpha^*(Y)$ (top row) over the space of possible datasets $Y \in \mathbb{R}^2$ with varying strengths of ridge regularization (increasing from top left to top right). Density in each column of subplots is indicated by the color scheme beneath. The elliptic structures are level curves of the quadratic form induced by the matrix Q_α . For comparison, the luckiness normalized density used by Dwivedi et al. is graphed for the same regularization strengths in the bottom row. The elliptic structures in the bottom row are level curves of posterior likelihood, which are induced by the matrix $Q_\alpha^{1/2}$ rather than Q_α . In both cases, probability density is only graphed over a square centered at the origin, but we calculated it by considering the space of datasets to be the entire two-dimensional plane $\mathcal{Y} = \mathbb{R}^2$. The black arrows are position vectors of the rows in X .

When accounting for ridge regularization on linear models, P_α^* assigns equal probability density to the surfaces of the ellipsoids formed by the matrix Q_α (meaning the level sets of the quadratic form $Y^T Q_\alpha Y = \text{Const}$). These are also surfaces of equal description length in dataset space. As the ridge penalty α is increased, these ellipsoids become smaller and more spherical. The density of P_α^* is increased close to the origin and decreased for datasets with large entries. However, more generally speaking, the geometry of P_α^* depends on the model class and the type of regularization performed.

7. Fitting a Reservoir Computer

As a proof of concept, we use our ridge penalty criterion to select a penalty strength for fitting reservoir computer readouts. Reservoir computers are high-dimensional dynamical systems that become conditionally stable when driven by an input signal. Their dynamic response is mapped to the future state of the input with a linear readout. This mapping

is possible because the reservoir state-space becomes a time-delay embedding-space of the driving time series [37] according to the echo-state property [38,39], which reservoirs must satisfy. Reservoir readouts are often fitted with ridge regression [40–42] and their complexity is a subject of interest [5,43–45]. In particular, we use echo-state networks [38] (ESNs) as our reservoirs and perform time series prediction on Lorenz-96 [46]. Our exact method is detailed in Appendix D along with a description of echo-state networks.

The average generalization error for one-step prediction of Lorenz-96 is compared with network size in Figure 2. The ridge penalty chosen according to stochastic complexity (in black) produces smaller generalization error on average than any fixed penalty magnitude. This is true across a range of network sizes and suggests that our method leads to sensible choices for the ridge parameter. Our criterion also achieves comparable results to some of the alternative approaches mentioned in Section 5, and performs comparatively well on large networks. This is demonstrated in Figure 3. It should be noted that the relative strength of the model complexity term is dependent on the value of ϵ for these methods, which we choose to be the smaller of the two values described in Section 4

$$\epsilon^2 = \min\left(\frac{n\sigma_{Y,0}^2}{n-k'}, \frac{Y^T(\mathbb{I} - M_{\alpha'})Y}{n - \text{Tr}(M_{\alpha'})}\right).$$

The first of these values is always positive and defined, since the length of the training time series is 1200 points and the largest network has 1096 nodes. The latter is calculated using a fixed value of $\alpha' = \exp(-14)$, which we select because the resulting variance estimate ϵ^2 tends to be stable for perturbations around this value, and a regularization strength of $\alpha = \exp(-14)$ tends to produce well-performing echo-state network readouts.

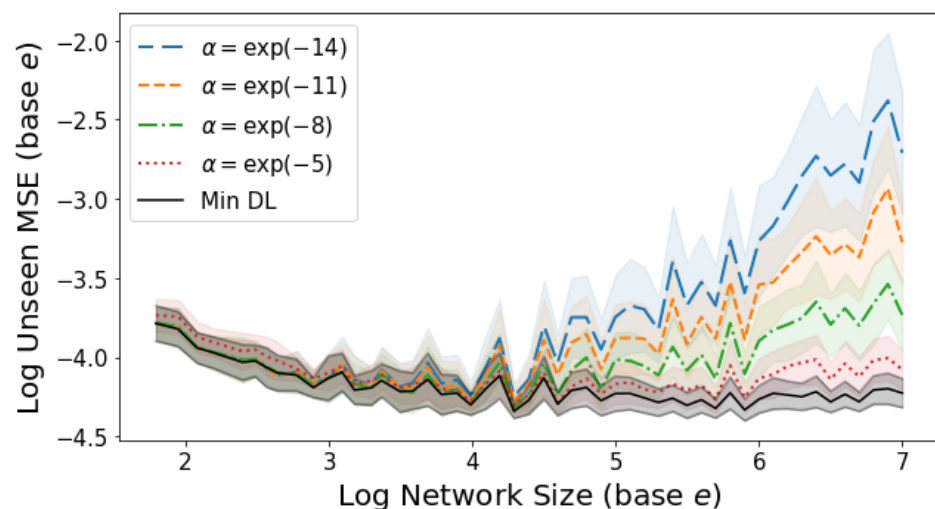


Figure 2. Reservoir performance compared with network size for different choices of regularization strength. The black line represents the performance (as measured by average log mean square testing error) achieved by the ridge strength that minimizes our selection criterion. The colored lines represent the performances achieved by using various fixed ridge penalty strengths. results are averaged over 200 samples and the error bounds indicate twice the standard deviation of bootstrap means.

Qualitative comparisons of reconstruction ability are also provided in Figures 4 and 5. This time, echo-state networks are trained for one-step-ahead prediction of the Lorenz system. Individual trajectories in each subplot are generated by continuously driving echo-state networks with their own past predictions. These predictions depend on output coefficients, which in turn depend on the ridge strength selected by each criterion. All criteria tend to produce reconstructions that are qualitatively similar to the Lorenz attractor when networks and prediction step sizes are small. This indicates that the ridge strengths

they select tend to be appropriate for capturing the underlying dynamics. For example, the reconstruction composites in Figure 4 are made using 200 node networks trained at a sampling rate of 0.02 s. On the other hand, ridge strengths selected by stochastic complexity appear to produce reasonable reconstructions more often when networks and step sizes are larger. This is the case in Figure 5, where predictions are made using 800 node networks and a step size of 0.1 s.

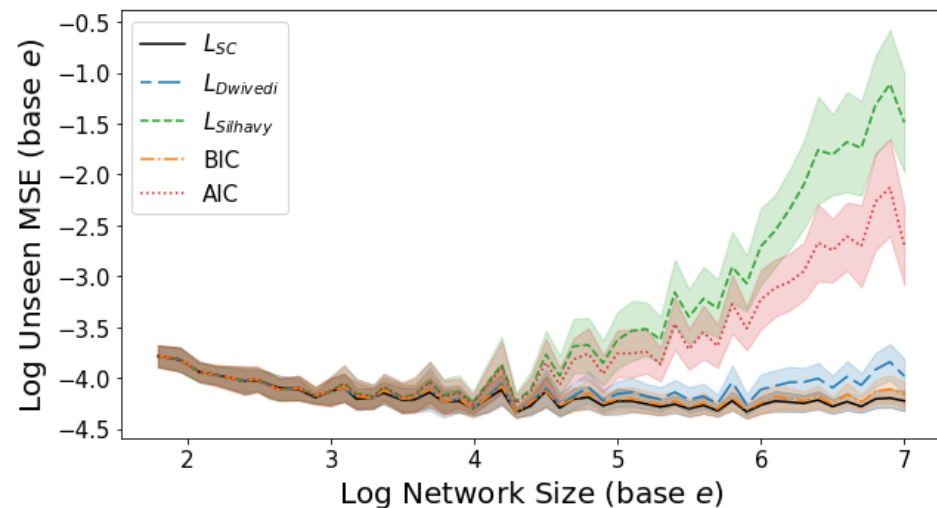


Figure 3. Reservoir performance compared with network size for different choices of regularization strength. The black line represents the performance (as measured by average log mean square testing error) achieved by the ridge strength that minimizes our selection criterion. The colored lines represent the performances achieved using the ridge strengths selected by the other information theoretic criteria discussed in Section 5. Results are averaged over 200 samples and the error bounds indicate twice the standard deviation of bootstrap means.

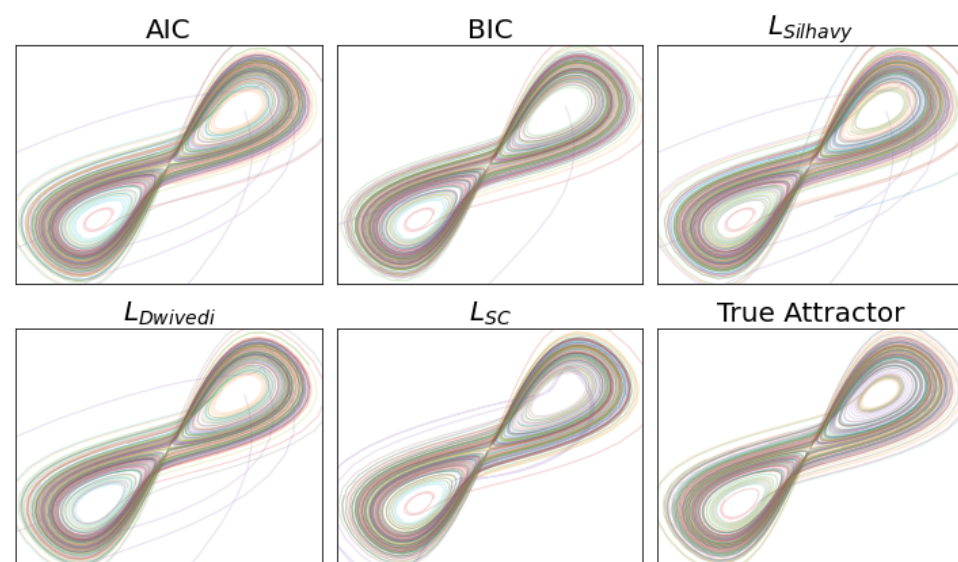


Figure 4. Composite time-delay embedding plots made from 50 trajectories autonomously reconstructed by echo-state networks. Each trajectory (represented in a different color) consists of 200 points predicted by a different network trained on a different integration of Lorenz. The same 50 networks, which each possess 200 nodes, are used to create each subplot. The difference between subplots, except for the bottom right sub-plot, is the ridge strength used to train the ESN readouts. In each case, these strengths were selected by the criterion stated in the title. ESNs were trained to predict the state of the Lorenz attractor 0.02 s into the future. The plot in the bottom right corner contains 50 trajectories numerically integrated from the governing Lorenz equations and represents the set of ideal reconstructions for comparison.

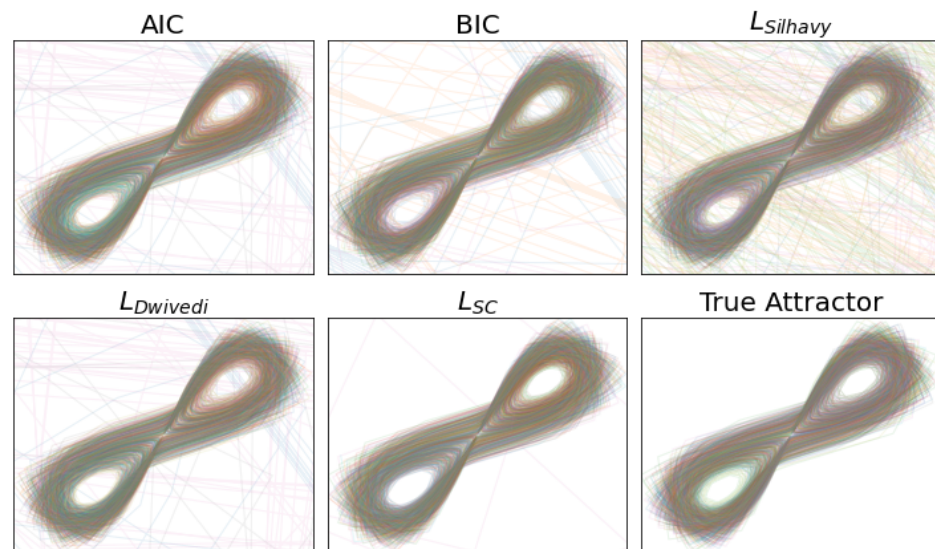


Figure 5. Composite time-delay embedding plots made from 50 trajectories autonomously reconstructed by echo-state networks, like in Figure 4. Each trajectory (represented by a different color) consists of 200 points predicted by a different network trained on a different integration of Lorenz. The same 50 networks are to create each subplot. However, networks are larger than those used in Figure 4, this time containing 800 nodes. The difference between subplots, except for the bottom right, is the ridge strength used to train the ESN readouts. The strengths are selected by the criterion stated in the title. ESNs were trained to predict the state of the Lorenz attractor further into the future (0.1 s) than those in Figure 4. The plot in the bottom right corner contains 50 trajectories numerically integrated from the governing Lorenz equations and represents the set of ideal reconstructions for comparison.

8. Results on PMLB Datasets

In this section, we replicate the benchmark testing performed by Dwivedi et al. [20] on the publicly available real-world datasets from the PMLB GitHub repository [22] (v1.0.2). Of the 122 datasets in the repository that are intended for regression, we ignore those that have fewer than 5 features, and test the information theoretic criteria discussed in this paper on the remaining 109 datasets. These datasets are sourced from a wide range of research areas.

The original experiments performed by Dwivedi et al. [20] use cross-validation for benchmarking and establish that their method performs well in comparison. This is the case in particular when data are scarce. For the sake of introducing a second point of comparison, we instead perform benchmarking with Stein's unbiased risk estimate [47,48] (SURE).

In Figure 6, we plot the average testing MSE achieved with each information theoretic criteria against the average achieved using SURE. The results in each subplot are achieved using a different ratio of training data points n to the number of dataset features m (which is also the number of tunable parameters). These ratios range from over-parameterized ($n = \lceil m/2 \rceil$) to moderately parameterized ($n = 4m$). In each of the 50 samples that we average across to produce the results in Figure 6, a random subset of rows is selected from each dataset for training, and the rest are reserved for testing. We estimate the variance hyper-parameter ϵ^2 using Liu's method [28] whenever the ratio of n to m is less than one. Otherwise, like in the previous section, we take the smaller of the estimates provided by the methods discussed in Section 4. However, because the typical scale of data varies widely between datasets in the PMLB repository, we adjust the value of α' used in Liu's method by the average squared magnitude $\bar{x}^2$ of the entries in each dataset ($\alpha' = \exp(-14)\bar{x}^2$).

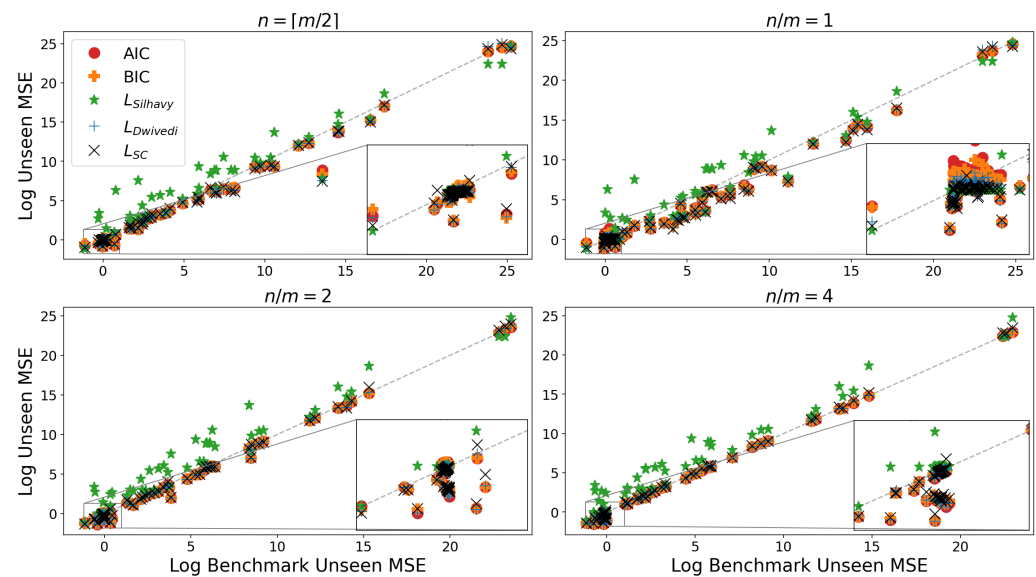


Figure 6. Average log testing MSE (base e) using various information theoretic ridge strength selection criteria compared with the average log testing MSE achieved using Stein’s unbiased risk estimate (SURE). Every point in each sub plot represents the average results achieved on a different dataset from the PMLB repository across 50 repetitions of the experiment. On the top left, models are trained using half as many data points n as the number of features each dataset contains m (with rounding up for datasets with an odd number of features). On the top-right, the ratio of training data points to features is 1, and on the bottom left and right, the ratio is 2 and 4, respectively. Points below the dashed line in each subplot represent datasets on which better average testing performance was achieved by the criterion than by SURE.

In the highly-parameterized settings ($1/2 \leq n/m \leq 2$) the information theoretic criteria discussed in this paper tend to perform well in comparison to the benchmark performance provided by SURE and similarly to one another. This aligns with the findings of Dwivedi et al. [20] and the comparison they drew with cross-validation. As the number of data points used for training is increased beyond the highly-parameterized settings, performances across all methods, including the benchmark, tend to become more similar.

9. Summary and Comments

We have proposed a selection criterion for the choice of penalty strength in ridge regression. In the process, we have demonstrated an application of stochastic complexity to a common problem. This required that we devise a model encoding scheme in accordance with the typical stochastic complexity scheme, but which is sensitive to the application of regularization. Our scheme suggests that we can consider the reduction in model complexity induced by regularization as a constraint on model flexibility. We believe that this result is consistent with the intuitive message espoused by stochastic complexity—that we should be hesitant in choosing a model class that fits our data well if it provides good fits to unrelated datasets using the same predictors. Regularization should ease our hesitancy by preventing our model from easily fitting these unrelated datasets.

The ridge penalty selection criterion that we derived produced promising results when used to fit linear readouts for echo-state networks. It also performed well in highly-parameterized settings on real-world datasets from the PMLB repository when compared to the benchmark provided by Stein’s unbiased risk estimate. However, this performance was well aligned with other information-theoretic criteria.

Author Contributions: Conceptualization, A.M.; methodology, A.M.; software, A.M.; validation, A.M., D.M.W. and M.S.; formal analysis, A.M.; investigation, A.M.; resources, A.M.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.M., D.M.W. and M.S.; visualization, A.M.; supervision, D.M.W. and M.S.; project administration, A.M., D.M.W. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: Provided by the Australian Research Council (ARC) Discovery Grant (No. DP200102961), funded by the Australian Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in the github repository Stochastic_Complexity_for_RR, accessible at https://github.com/antony-mizzi/Stochastic_Complexity_for_RR (accessed on 5 June 2026).

Acknowledgments: A.M. was supported by the Australian Government RTP scholarship at the University of Western Australia. M.S. and D.M.W. are supported by the ARC Discovery Grant (No. DP200102961), funded by the Australian Government.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GLM	Generalized Linear Model
MDL	Minimum Description Length
NML	Normalized Maximum Likelihood
LNML	Luckiness Normalized Maximum Likelihood
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
ESN	Echo-State Network
MSE	Mean Square Error
SURE	Stein's Unbiased Risk Estimate

Appendix A. Demonstration of Minimax Regret Under Proportionality

In the original formulation of stochastic complexity, a coding scheme for possible datasets $Y \in \mathcal{Y}$ is optimized according to the probabilities assigned by the distribution P_0^* that uniquely minimizes worst-case regret as defined in Section 2. It was first proven by Shtarkov [49] that worst-case regret is minimal when P_0^* is proportional to maximum likelihood $P_0^*(Y) \propto P(Y|\theta_Y^*)$. The simpler argument that we use in this demonstration is borrowed from Grunwald [21] (Chapter 6.1.2.)

Recall the modified definitions of regret

$$R_\alpha(Y, P_\alpha^*) := \log(P(Y|\theta_{Y,\alpha}^*)) - \log(P_\alpha^*(Y)) \equiv -\log\left(\frac{P_\alpha^*(Y)}{P(Y|\theta_{Y,\alpha}^*)}\right),$$

and worst case regret

$$R_\alpha^{\max}(\mathcal{Y}, P_\alpha^*) := \max_{Y \in \mathcal{Y}} (R_\alpha(Y, P_\alpha^*)),$$

provided in Section 3. We first establish that regret is constant across datasets (meaning $R_\alpha(Y_1, P_\alpha^*) = R_\alpha(Y_2, P_\alpha^*) \forall Y_1, Y_2 \in \mathcal{Y}$) if and only if P_α^* is proportional to the likelihood of regularized fits ($P_\alpha^* \propto P(Y|\theta_{Y,\alpha}^*)$). This is because it follows from our definition of regret

that equivalence for different datasets requires equal ratios of their probabilities to the likelihoods of their regularized fits

$$R_{\alpha}(Y_1, P_{\alpha}^*) = R_{\alpha}(Y_2, P_{\alpha}^*) \Rightarrow \frac{P_{\alpha}^*(Y_1)}{P(Y_1|\theta_{Y_1,\alpha}^*)} = \frac{P_{\alpha}^*(Y_2)}{P(Y_2|\theta_{Y_2,\alpha}^*)} := \frac{1}{C_{\alpha}}.$$

For any distribution P' defined on $\mathcal{Y}$ that is different to P_{α}^* , there must be some dataset $Y \in \mathcal{Y}$ to which it assigns a lesser probability than P_{α}^*

$$\exists Y \in \mathcal{Y} \text{ s.t. } P'(Y) < P_{\alpha}^*(Y).$$

If no such dataset exists, then P' does not integrate to one and is not a valid distribution. The regret attained by P' on this dataset must be greater than the worst case regret attained by P_{α}^*

$$\log(P(Y|\theta_{Y,\alpha}^*)) - \log(P'(Y)) > \log(P(Y|\theta_{Y,\alpha}^*)) - \log(P_{\alpha}^*(Y)) = \log(C_{\alpha}).$$

It follows that the worst case regret under P' is greater than under P_{α}^* . This is true for any arbitrary distribution P' that is different to P_{α}^* , so P_{α}^* is the unique distribution that achieves minimax regret.

Appendix B. Conditions for a Well-Defined Coding Scheme

As discussed in Section 3 and illustrated in Section 6, by adjusting the notion of regret to account for regularization, we change the probability density of the resulting NML distribution P_{α}^* . The coding scheme produced by this new distribution will remain well-defined in all contexts where the scheme produced without regularization is well-defined, so long as it stays non-zero over the original support $P_{\alpha}^*(Y) > 0 \forall Y \in \mathcal{Y}$. This is the case if the following two requirements are satisfied. Firstly, the normalization factor must not be made zero or infinite by regularization,

$$\begin{aligned} 0 < \int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) dY' < \infty \text{ in all cases where} \\ 0 < \int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) dY' < \infty. \end{aligned}$$

Secondly, model likelihoods at regularized parameter fits must remain non-zero over the full support of P_0^* , which we can write as follows,

$$\theta_{Y,\alpha}^* \text{ exists, and } P(Y|\theta_{Y,\alpha}^*) > 0 \forall Y \in \mathcal{Y}.$$

Together, these requirements ensure that codeword lengths defined by our modified form of stochastic complexity

$$L_{SC}(Y; \mathcal{M}, X, \alpha) = -\log \frac{P(Y|\theta_{Y,\alpha}^*; \mathcal{M}, X)}{\int_{Y' \in \mathcal{Y}} P(Y'|\theta_{Y',\alpha}^*; \mathcal{M}, X) dY'}$$

are finite for every possible dataset, even when the space of possibilities is the full support of P_0^* .

The first of the two requirements is always satisfied. Probability is greatest at the maximum likelihood estimate $P(Y|\theta_Y^*) \geq P(Y|\theta_{Y,\alpha}^*)$, which means that the modified normalization factor (Equation (6)) is never more than the original normalization factor (Equation (4)). It also does not make sense for the normalization factor to be reduced to zero. On the contrary, there are combinations of model classes and regularization techniques, like linear

models and ridge regression, that together produce a finite normalization factor despite the normalization factor being infinite without regularization.

The second requirement is satisfied for most conceivable model classes and types of regularization. For model likelihoods to be reduced to zero by regularization, the distribution used to assign model likelihood must have restricted support. We used a multivariate Gaussian to assign likelihoods to linear models, which means that likelihood is positive no matter how poor the regularized fit ($P(Y|\theta) > 0 \forall Y, \theta$). Even if the distribution used to assign likelihood has restricted support, regularization still needs to make it possible to assign a best-fit likelihood of zero. For example, no strength of L2 or L1 regularization could produce zero likelihood fits ($P(Y|\theta_{Y,\alpha}^*) = 0$) in any scenario where maximum likelihood is non-zero ($P(Y|\theta_Y^*) > 0$). This is because they both maximize posterior likelihood after assuming priors $\pi(\theta)$ that have universal support. Assigning a zero-likelihood regularized fit when a non-zero maximum likelihood exists would require the following impossibility

$$\max_{\theta \in \Theta} (P(Y|\theta)\pi(\theta)) = 0 \text{ despite } P(Y|\theta_Y^*), \pi(\theta_Y^*) > 0 \text{ for some } \theta_Y^* \in \Theta.$$

In classification tasks, models typically assign probabilities to each category in such a way that probabilities are strictly greater than zero and less than one. This makes it impossible for regularized parameter fits to ever assign dataset likelihoods of zero and violate the second requirement. Stochastic complexity is also reduced to its discrete form in categorization tasks, which means that the normalization factor is always bounded above by the finite number of possible datasets, and the first requirement is always satisfied.

Appendix C. Universality in Expectation

For the NML distribution derived in Section 4 to constitute a universal model in the expected sense, its efficiency must converge on the optimal efficiency for any possible data-generating process as the number of data points tends to infinity [21]. The data-generating processes in consideration are those that can be generated by any parameterization of the model class used to derive the coding scheme. In our case, these are the multivariate normal distributions with variance ϵ^2 and mean $X\theta$. The requirement for universality is satisfied if expected regret (R_0) per symbol tends to zero in the asymptotic limit for dataset size. Expectation is taken with respect to the probability of each dataset y^n given the data generating process. Here we denote datasets by y^n to indicate their length. We also denote data-generating distributions by P_θ , under which the probability of each dataset is conditional on some true parameter vector $P_\theta(y^n) = P(y^n|\theta)$. As such, we can write the requirement for universality in the expected sense using the definition of R_0 from Section 2

$$\frac{1}{n} \mathbb{E}_{P_\theta} [R_0(y^n, P_\alpha^*)] := \frac{1}{n} \int R_0(y^n, P_\alpha^*) P(y^n|\theta) dy^n \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\text{where } R_0(y^n, P_\alpha^*) = \log(P(y^n|\theta_{y^n}^*)) - \log(P_\alpha^*(y^n)) = \log \frac{P(y^n|\theta_{y^n}^*)}{P_\alpha^*(y^n)}.$$

We can also re-write regret using the expression derived for P_α^* in Section 4

$$R_0(y^n, P_\alpha^*) = \log \frac{\exp(-\frac{1}{2\epsilon^2}(y^n)^T Q_0 y^n)}{\exp(-\frac{1}{2\epsilon^2}(y^n)^T Q_\alpha y^n)} + \sum_{i=1}^m \log(1 + \frac{\lambda_i}{\alpha}),$$

and further simplify it to the following

$$R_0(y^n, P_\alpha^*) = \frac{1}{2\epsilon^2} (y^n)^T (Q_\alpha - Q_0) y^n + \sum_{i=1}^m \log(1 + \frac{\lambda_i}{\alpha}). \quad (\text{A1})$$

The complexity penalty $\sum_{i=1}^m \log(1 + \frac{\lambda_i}{\alpha})$ has no dependence on the dataset y^n . As such, we can exclude it from the expectation and re-write our requirement

$$\frac{1}{n} \mathbb{E}_{P_\theta} \left[\frac{1}{2\epsilon^2} (y^n)^T (Q_\alpha - Q_0) y^n \right] + \frac{1}{n} \log(1 + \frac{\lambda_i}{\alpha}) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (\text{A2})$$

As discussed in Section 4, the complexity penalty is asymptotically equivalent to the BIC penalty and scales as $\log(n)$. This means that the second term in Equation (A2) vanishes as n is taken to infinity at a rate of $\log(n)/n$. We are left with only the requirement that the first term in Equation (A2) vanishes. Given that P_θ is the multivariate normal density function, we can evaluate the expectation value and write this requirement as follows

$$\frac{1}{n} \left(\frac{(\mu^n)^T (Q_\alpha - Q_0) \mu^n}{2\epsilon^2} + \frac{1}{4\epsilon^2} \text{Tr}(Q_\alpha - Q_0) \right) \rightarrow 0, \quad (\text{A3})$$

where μ^n is the expected dataset given the generating distribution ($\mu^n = X\theta$).

To show that Equation (A3) holds in the asymptotic limit for n , we can factorize the matrix $Q_\alpha - Q_0$ using the definition of Q_α provided in Section 4

$$Q_\alpha - Q_0 = (M_\alpha - M_0)(M_\alpha + M_0 - 2\mathbb{I}),$$

and then use the matrix identity $A^{-1} - (A + \alpha\mathbb{I})^{-1} = \alpha A^{-1}(A + \alpha\mathbb{I})^{-1}$

$$Q_\alpha - Q_0 = \alpha X(X^T X)^{-1}(X^T X + \alpha I)^{-1} X^T (M_\alpha + M_0 - 2\mathbb{I}).$$

The entries of the $m \times m$ gram matrix $X^T X$ grow on the order n . This means that the entries of the $n \times n$ matrix $M_\alpha - M_0$ that we expanded grow on the order of $1/n^2$. Meanwhile, The diagonal entries of $M_\alpha + M_0$ are bounded between 0 and 2, and the off-diagonal entries are bounded between -2 and 2 . This means that the magnitude of all entries in the matrix $M_\alpha + M_0 - 2\mathbb{I}$ are bounded above by 2. Since the entries of the matrix $Q_\alpha - Q_0$ are obtained from the product of these matrices

$$(Q_\alpha - Q_0)_{ij} = \sum_{k=0}^n (M_\alpha - M_0)_{ik} (M_\alpha + M_0 - 2\mathbb{I})_{kj},$$

they adhere to an upper bound

$$|(Q_\alpha - Q_0)_{ij}| \leq 2 \sum_{k=0}^n |(M_\alpha - M_0)_{ik}|,$$

which grows on the order of $1/n$. This means that the quadratic form $(\mu^n)^T (Q_\alpha - Q_0) \mu^n$ and the trace $\text{Tr}(Q_\alpha - Q_0)$, which require an additional summation of n terms, are scale invariant. It follows that the expectation value in Equation (A2) also vanishes in the asymptotic limit and P_α^* achieves universality in the expected sense. This remains true even if the regularization strength α is allowed to grow in conjunction with the number of data-points, so long as the growth rate is sub-linear.

Appendix D. Echo-State Networks and Time-Series Prediction

Echo-state networks are a specific type of reservoir. They consist of a network and a set of input weights and behave like the recurrent part of a recurrent neural network. At each time step, the network is driven according to the value of the time series $u(t)$. The network state $x(t)$ updates according to the governing equation

$$x(t+1) = f(Ax(t) + W^{in}u(t) + b).$$

Unlike recurrent neural networks, the adjacency matrix A , input connections W^{in} and node biases b all remain untrained. The activation function f , which in our case is the hyperbolic tangent $f(\cdot) = \tanh(\cdot)$, introduces non-linearity. Predictions at each time step are given by the linear combination of node values $\theta^T x(t)$ with parameters θ determined from ridge regression $\theta_{Y,\alpha}^* = (X^T X + \alpha \mathbb{I})^{-1} X^T Y$. We train echo-state networks for one-step-ahead prediction. In this case, X is the matrix with a t th row of $x(t)$ and Y is the vector with a t th entry of $u(t+1)$.

In the experiments performed in Section 7, every sample involves a new ESN architecture and a newly generated time series. We generate the network adjacency A as a weighted Watts-Strogatz network with a connectivity of 6 weightings drawn from the standard normal. The adjacency matrix is then normalized to have a spectral radius of ρ , which is sampled uniformly between $3/4$ and $3/2$. Input connection strengths are sampled from a Gaussian with zero mean and a standard deviation of 0.3. We do not use node biases ($b = \vec{0}$).

For Figures 2 and 3 the driving time series consists of 1300 points generated from a Runge-Kutta (RK4) integration of Lorenz-96 with a sampling rate of 0.1 s and an integration step length of 0.01 s. However, ESNs are only trained according to the final 1200 of these time series points. The first 99 are disregarded to establish conditional dependence of the state of the network on the time series and predictions are trained 1 step ahead. Lorenz96 is a high-dimensional dynamical system with a Lyapunov time that we calculate to be about 0.68 s using the method in [50]. It has a governing set of equations detailed below in Equation (A4), and we use the hyperparameters $(R, n) = (8, 15)$.

$$\dot{x}_i = (x_{i+1} - x_{i-2})x_{i-1} - x_i + R, \quad (\text{A4})$$

Note that in this equation, x_0 is the same as x_n .

For the time delay embedding plots (Figures 4 and 5), we instead train the reservoir at one-step prediction using 1200 points from the Lorenz system integrated using an RK4 scheme with a step length of 0.01 s. The driving time series again contains 1300 points but the first 99 are disregarded when training the readout. The governing equations of this system are provided below. We use standard hyperparameters $(\sigma, \rho, \beta) = (10, 28, \frac{8}{3})$ for which we calculate the Lyapunov time to be around 1.1 s.

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z - y) \\ \dot{z} &= xy - \beta z \end{aligned}$$

The trajectories shown in Figures 4 and 5 are generated autonomously. This means that rather than driving ESNs by an external signal, we instead drive them at each time step by their previous prediction. Effectively, their predictions are an Eulerian integration of the following system of equations

$$\begin{aligned} \hat{u}(t) &= \theta^T x(t), \\ x(t+1) &= f(Ax(t) + W^{in}\hat{u}(t) + b). \end{aligned}$$

This autonomous prediction is continued for a fixed duration of 200 time steps. For each predicted trajectory $\hat{u}$, We graph $\hat{u}(t)$ against $\hat{u}(t+5)$ in Figure 4 and $\hat{u}(t)$ against $\hat{u}(t+1)$ in Figure 5. Both plots are comprised of 50 different trajectories, each of which is produced by a different ESN architecture trained on a different time series. The same 50 architectures and time series are used in each subplot.

References

- Judd, K.; Mees, A. On selecting models for nonlinear time series. *Phys. D. Nonlinear Phenom.* **1995**, *82*, 426–444. [\[CrossRef\]](#)
- Myers, R.H.; Montgomery, D.C.; Vining, G.G.; Robinson, T.J. *Generalized Linear Models: With Applications in Engineering and the Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
- Conniffe, D.; Stone, J. A critical view of ridge regression. *J. R. Stat. Soc. Ser. D (Stat.)* **1973**, *22*, 181–187. [\[CrossRef\]](#)
- Clarke, B.S.; Barron, A.R. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory* **2002**, *36*, 453–471.
- Abdalla, M.; Zrounba, C.; Cardoso, R.; Jimenez, P.; Ren, G.; Boes, A.; Mitchell, A.; Bosio, A.; O'Connor, I.; Pavanello, F. Minimum complexity integrated photonic architecture for delay-based reservoir computing. *Opt. Express* **2023**, *31*, 11610–11623. [\[CrossRef\]](#) [\[PubMed\]](#)
- Schneider, A.; Hommel, G.; Blettner, M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. *Dtsch. Ärztebl. Int.* **2010**, *107*, 776–782. [\[PubMed\]](#)
- Nozari, E.; Bertolero, M.A.; Stiso, J.; Caciagli, L.; Cornblath, E.J.; He, X.; Mahadevan, A.S.; Pappas, G.J.; Bassett, D.S. Macroscopic resting-state brain dynamics are best described by linear models. *Nat. Biomed. Eng.* **2024**, *8*, 68–84. [\[PubMed\]](#)
- Zhang, F.; Tapera, T.M.; Gou, J. Application of a new dietary pattern analysis method in nutritional epidemiology. *BMC Med. Res. Methodol.* **2018**, *18*, 119. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hellwig, Z. *Linear Regression and Its Application to Economics*; Elsevier: Amsterdam, The Netherlands, 2014.
- Peters, G. A linear forecasting model and its application to economic data. *J. Forecast.* **2001**, *20*, 315–328. [\[CrossRef\]](#)
- Torres, A.F.C.; Akbaritabar, A. The hegemonic use of linear models in quantitative social sciences. *Bull. Sociol. Methodol. Bull. De Méthodol. Sociol.* **2025**, *167*, 78–95. [\[CrossRef\]](#)
- Mize, T.D. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociol. Sci.* **2019**, *6*, 81–117. [\[CrossRef\]](#)
- Luqman, M.; Bhatti, S.H.; Aydin, D.; Jamil, M. Addressing multicollinearity in general linear model: A novel approach for ridge parameter with performance comparison. *PLoS ONE* **2025**, *20*, e0335072. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hoerl, A.E.; Kennard, R.W. Ridge regression—1980: Advances, algorithms, and applications. *Am. J. Math. Manag. Sci.* **1981**, *1*, 5–83. [\[CrossRef\]](#)
- Wu, D.; Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10112–10123.
- Gillariose, J.; Joseph, J.; Chesneau, C. Lasso and ridge regression: A comprehensive review of applications and developments in machine learning. *Int. J. Data Sci. Anal.* **2026**, *21*, 7.
- Reid, S.; Grudic, G. Regularized linear models in stacked generalization. In *Proceedings of the International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 112–121.
- Rissanen, J. Stochastic complexity. *J. R. Stat. Soc. Ser. B (Methodol.)* **1987**, *49*, 223–239. [\[CrossRef\]](#)
- Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [\[CrossRef\]](#)
- Dwivedi, R.; Singh, C.; Yu, B.; Wainwright, M. Revisiting minimum description length complexity in overparameterized models. *J. Mach. Learn. Res.* **2023**, *24*, 1–59.
- Grünwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
- Olson, R.S.; La Cava, W.; Orzechowski, P.; Urbanowicz, R.J.; Moore, J.H. PMLB: A large benchmark suite for machine learning evaluation and comparison. *BioData Min.* **2017**, *10*, 36. [\[CrossRef\]](#) [\[PubMed\]](#)
- Grünwald, P. The Minimum Description Length Principle and Reasoning Under Uncertainty. Ph.D. Thesis, University of Amsterdam, Amsterdam, The Netherlands, January 1998.
- Jeffreys, H. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, London, UK, 24 September 1946; The Royal Society: London, UK, 1946; Volume 186, pp. 453–461.
- Kontkanen, P.; Buntine, W.; Myllymäki, P.; Rissanen, J.; Tirri, H. Efficient computation of stochastic complexity. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, Key West, FL, USA, 21 March 2003; Society for Artificial Intelligence and Statistics: Key West, FL, USA, 2003; pp. 233–238.
- Kibria, B.G. More than hundred (100) estimators for estimating the shrinkage parameter in a linear and generalized linear ridge regression models. *J. Econom. Stat.* **2022**, *2*, 233–252.
- Cule, E.; Do Iorio, M. Ridge regression in prediction problems: Automatic choice of the ridge parameter. *Genet. Epidemiol.* **2013**, *37*, 704–714. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liu, X.; Zheng, S.; Feng, X. Estimation of error variance via ridge regression. *Biometrika* **2020**, *107*, 481–488. [\[CrossRef\]](#)
- Silhavy, P.; Hlaváčková-Schindler, K.; Silhavy, R. Minimum description length and multi-criteria decision analysis in predictive modeling. *IEEE Access* **2025**, *13*, 19388–19407. [\[CrossRef\]](#)
- Giurcăneanu, C.D.; Razavi, S.A.; Liski, A. Variable selection in linear regression: Several approaches based on normalized maximum likelihood. *Signal Process.* **2011**, *91*, 1671–1692. [\[CrossRef\]](#)
- Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [\[CrossRef\]](#)

32. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [\[CrossRef\]](#)
33. Van Le, C. How to choose tuning parameters in lasso and ridge regression? *Asian J. Econ. Bank.* **2020**, *4*, 61–76.
34. Flynn, C.J.; Hurvich, C.M.; Simonoff, J.S. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *J. Am. Stat. Assoc.* **2013**, *108*, 1031–1043. [\[CrossRef\]](#)
35. Dijkstra, T.K. Ridge regression and its degrees of freedom. *Qual. Quant.* **2014**, *48*, 3185–3193.
36. Boonstra, P.S.; Mukherjee, B.; Taylor, J.M. A small-sample choice of the tuning parameter in ridge regression. *Stat. Sin.* **2015**, *25*, 1185–1206. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Duan, X.-Y.; Ying, X.; Leng, S.-Y.; Kurths, J.; Lin, W.; Ma, H.-F. Embedding theory of reservoir computing and reducing reservoir network using time delays. *Phys. Rev. Res.* **2023**, *5*, L022041. [\[CrossRef\]](#)
38. Jaeger, H. *The “Echo State” Approach to Analysing and Training Recurrent Neural Networks—With an Erratum Note*; Technical Report; Jacobs University Bremen: Bremen, Germany, 2010; Volume 148, p. 13.
39. Yildiz, I.B.; Jaeger, H.; Kiebel, S.J. Revisiting the echo state property. *Neural Netw.* **2012**, *35*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Panahi, S.; Zhai, Z.-M.; Haile, M.; Lai, Y.-C. Neuromorphic reservoir computing. *Chaos Interdiscip. J. Nonlinear Sci.* **2025**, *35*, 123144. [\[CrossRef\]](#)
41. Cucchi, M.; Abreu, S.; Ciccone, G.; Brunner, D.; Kleemann, H. Hands-on reservoir computing: A tutorial for practical implementation. *Neuromorphic Comput. Eng.* **2022**, *2*, 032002. [\[CrossRef\]](#)
42. Kent, R.; Lienhard, B.; Lafyatis, G.; Gauthier, D.J. Superconducting qubit readout using next-generation reservoir computing. *Phys. Rev. Appl.* **2026**, *25*, 044009. [\[CrossRef\]](#)
43. Dale, M.; Dewhurst, J.; O’Keefe, S.; Sebald, A.; Stepney, S.; Trefzer, M.A. The role of structure and complexity on reservoir computing quality. In *Proceedings of the International Conference on Unconventional Computation and Natural Computation, Tokyo, Japan, 3–7 June 2019*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 52–64.
44. Zeng, Y.; Li, Z.; Gao, X.; Cheng, M. Optical chaos shift keying communication system via neural network-based signal reconstruction. *Nonlinear Dyn.* **2025**, *113*, 32849–32863. [\[CrossRef\]](#)
45. Fong, R.S.; Li, B.; Tino, P. Universality of real minimal complexity reservoir. In *Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 16622–16629.
46. Lorenz, E.N. Predictability: A problem partly solved. In *Proceedings of Seminar on Predictability, Shinfield Park, Reading, UK, 4–8 September 1996*; ECMWF: Reading, UK, 1996; Volume 1, pp. 1–18.
47. Stein, C.M. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* **1981**, *9*, 1135–1151. [\[CrossRef\]](#)
48. Donoho, D.L.; Johnstone, I.M. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **1995**, *90*, 1200–1224. [\[CrossRef\]](#)
49. Shtar’kov, Y.M. Universal sequential coding of single messages. *Probl. Peredachi Informatsii* **1987**, *23*, 3–17.
50. Benettin, G.; Galgani, L.; Giorgilli, A.; Strelcyn, J.M. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. *Meccanica* **1980**, *15*, 9–20. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.