Efficient Model Selection Using Penalized Methods

Jie Ding, Enmao Diao, Jiawei Zhou, and Vahid Tarokh
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, 02138 USA
Email: jieding@fas.harvard.edu

***Abstract*—Machine learning systems learn from and make predictions by building models from observed data. Because large models tend to inflate the variance (e.g. overfitting) while small models tend to cause biases (e.g. underfitting) for a given fixed dataset, a critical challenge is to select an appropriate model (e.g. set of variables/features). Model selection aims to strike a balance between the goodness of fit and model complexity, and thus to gain reliable predictive power. In this paper, we study a penalized model selection technique that asymptotically achieves the optimal expected prediction loss (referred to as the limit of learning) offered by a set of candidate models (model class). We prove that the proposed procedure is both statistically efficient in the sense that it asymptotically approaches the limit of learning, and computationally efficient in the sense that it can be much faster than cross validation methods. Our theory applies for a wide variety of model classes, loss functions, and high dimensions (i.e. the models' complexity can grow with data size). As we will show, allowing high dimensions is necessary if "all models are wrong" and data are sequentially observed (which often happens in practice). We demonstrate the proposed methodology by applying it to generalized linear models and neural networks, showing its superior predictive power and/or smaller computational cost compared to various types of cross validations. We also released a python package for easy implementation when the candidate models under consideration are neural networks.**

***Keywords*-Cross-validation; Computational efficiency; Feature selection; High dimension; Limit of learning;**

I. Introduction

How much knowledge can we learn from a given set of data? Statistical modeling provides a simplification of real world complexity. It can be used to learn the key *patterns* or *relationships* from available data and to predict the future data. In order to model the data, typically the first step in data analysts is to narrow the scope by specifying a set of candidate parametric models (referred to as model class). The model class can be determined by exploratory studies or scientific reasoning. For data with specific types and sizes, each postulated model may have its own advantages. In the second step, data analysts estimate the parameters and "goodness of fit" of each candidate model. An illustration of a typical learning procedure is plotted in Fig. 1, where the true data generating model may or may not be included in the model class. Simply selecting the model with the best fitting performance usually leads to suboptimal results. For example, the largest model always fits the best in a nested model class. But too large a model can lead to inflated variance and thus severe overfitting. Therefore, the third step is to apply a model selection procedure. State-of-art selection procedure can be roughly categorized into two classes, the penalized selection and cross-validation. We shall elaborate on those in the next section.

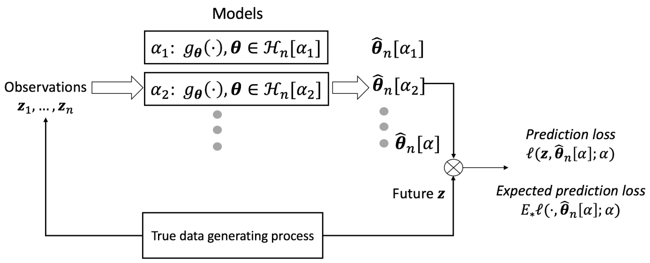

Models
α_1: $g_{\boldsymbol\theta}(\cdot), \boldsymbol\theta \in \mathcal{H}_n[\alpha_1]$
α_2: $g_{\boldsymbol\theta}(\cdot), \boldsymbol\theta \in \mathcal{H}_n[\alpha_2]$
$\hat{\boldsymbol\theta}_n[\alpha_1]$
$\hat{\boldsymbol\theta}_n[\alpha_2]$
$\hat{\boldsymbol\theta}_n[\alpha]$
Observations
$\boldsymbol z_1, \dots, \boldsymbol z_n$
Prediction loss
$\ell(\boldsymbol z, \hat{\boldsymbol\theta}_n[\alpha]; \alpha)$
Future $\boldsymbol z$
Expected prediction loss
$E_*\ell(\cdot, \hat{\boldsymbol\theta}_n[\alpha]; \alpha)$
True data generating process

Figure 1: Illustration of a typical learning procedure, where each candidate model α_j is trained in terms of $\hat{\boldsymbol\theta}_n[\alpha_j]$ in its parameter space $\mathcal{H}_n[\alpha_j]$, and then used to evaluate future data under some loss function $\ell(\cdot)$.

Example 1 (Generalized linear models). *In a generalized linear model (GLM), each response variable Y is assumed to be generated from a particular distribution (e.g. Gaussian, Binomial, Poisson, Gamma), with its mean $\boldsymbol\mu$ linked with potential covariates $X_1, X_2, \dots$ through $E_*(Y) = \mu = g^{-1}(\beta_1 X_1 + \beta_2 X_2 + \cdots)$. In this example, data $Z = [Y, X_1, X_2, \dots]^{\mathrm{T}}$, unknown parameters are $\boldsymbol\theta = [\beta_1, \beta_2, \dots]^{\mathrm{T}}$, and models are subsets of $\{\beta_1, \beta_2, \dots\}$. We may be interested in the most appropriate distribution form (pattern) as well as the most significant variables X_j's (relationships).*

Example 2 (Neural networks). *In establishing a neural network (NN) model, we need to choose the number of neurons and hidden layers, activation function (relationships), and the configuration of their connectivity (pattern). In this example, data are similar to that of the above example, and unknown parameters are the weights on connected edges. Clearly, with larger number of neurons and connections, more complex functional relationships can be modeled. But* *selecting models with too large of dimensions may result in overfitting and more computational complexity.*

How can we quantify the theoretical limits of learning procedures? We first introduce the expected prediction loss that quantifies the predictive power of each candidate model.

Definition 1 (Expected prediction loss). *The loss function for each data size n and $\alpha \in \mathcal{A}_n$ (model class) is a map $l_n : \mathcal{Z} \times \mathcal{H}_n[\alpha] \rightarrow \mathbb{R}$, usually written as $l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha)$, where $\mathcal{Z}$ is the data domain, $\mathcal{H}_n[\alpha]$ is the parameter space associated with model α, and α is included to emphasize the model under consideration. As Fig. 1 shows, for a loss function and a given dataset $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ which are independent and identically distributed (i.i.d.), each candidate model α produces an estimator $\hat{\boldsymbol{\theta}}_n[\alpha]$ (referred to as the minimum loss estimator) defined by*

$$\hat{\boldsymbol{\theta}}_n[\alpha] \triangleq \underset{\theta \in \mathcal{H}_n[\alpha]}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} l_n(z_i, \boldsymbol{\theta}; \alpha). \tag{1}$$

Moreover, the expected prediction loss given by candidate model α, denoted by $\mathcal{L}_n(\alpha)$, is defined by

$$\mathcal{L}_n(\alpha) \triangleq E_* l_n\big(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big) = \int_{\mathcal{Z}} p(\boldsymbol{z}) l_n\big(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big) d\boldsymbol{z}. \tag{2}$$

Here, E_* denotes the expectation with respect to the distribution of a future (unseen) random variable $\boldsymbol{z}$. We also define the risk by $\mathcal{R}_n[\alpha] = E_* \mathcal{L}_n[\alpha]$, where the expectation in $\mathcal{R}_n[\alpha]$ is taken with respect to the observed data.

Typically $\boldsymbol{z}$ consists of response $\boldsymbol{y}$ and covariates $\boldsymbol{x}$, and only the entries of $\boldsymbol{x}$ associated with α are involved in the evaluation of l_n. Throughout the paper, we consider loss functions $l_n(\cdot)$ such that $\mathcal{L}_n[\alpha]$ is always nonnegative. A common choice is to use negative log-likelihood of model α minus that of the true data generating model. Table I lists some other loss functions widely used in machine learning. Based on Definition 1, a natural way to define the limit of learning is by using the optimal prediction loss.

Definition 2 (Limit of learning). *For a given data (of size n) and model class $\mathcal{A}_n$, the limit of learning (LoL) is defined as $\arg\min_{\alpha \in \mathcal{A}_n} \mathcal{L}_n(\alpha)$, the optimal expected prediction loss offered by candidate models.*

We note that the LoL is associated with three key elements: data, loss function, and model class. Motivated by the original derivation of Akaike information criterion (AIC) [1], [2] and Takeuchi's information criterion (TIC) [3], we propose a penalized selection procedure and prove that it can approach the LoL under reasonable assumptions. Those assumptions allow a wide variety of loss functions, model classes (i.e. nested, non-overlapping or partially-overlapping), and high dimensions (i.e. the models' complexity can grow with data size). Our theoretical results extend the classical statistical theory on AIC for linear (fixed-design) regression models. Moreover, we also review the conceptual and technical connections between cross validation and information theoretical criteria. In particular, we show that the proposed procedure can be much more computationally efficient than cross validation (with the same level of predictive power).

Why is it necessary to consider a high dimensional model class, e.g., to let the number of candidate models or each model's complexity grow with data size? In the context of fixed design regression, technical discussions that address the question have been elaborated in [4], [5]. Here, we give an intuitive explanation for a general setting. We let $\boldsymbol{\theta}_n^*[\alpha]$ denote the minimum loss parameter defined by

$$\boldsymbol{\theta}_n^*[\alpha] \triangleq \underset{\theta \in \mathcal{H}_n[\alpha]}{\arg\min} E_* l_n(\cdot, \boldsymbol{\theta}; \alpha). \tag{3}$$

We shall show in the appendix that $\mathcal{L}_n[\alpha]$ may be expressed as

$$\begin{aligned}\mathcal{L}_n[\alpha] =& E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}_n^*[\alpha]; \alpha) + \frac{1}{2}\big(\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\big)^{\mathrm{T}} V_n(\boldsymbol{\theta}_n^*; \alpha) \cdot \\ & \big(\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\big) \times \{1 + o_p(1)\}\end{aligned} \tag{4}$$

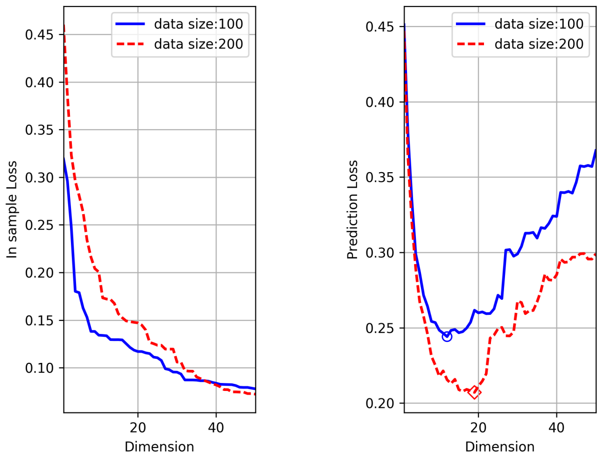
under some regularity conditions, where $V_n(\boldsymbol{\theta}_n^*; \alpha) \triangleq E_* \nabla_{\boldsymbol{\theta}}^2 l_n(\cdot, \boldsymbol{\theta}; \alpha)$, and $o_p(1)$ is a sequence of random variables that converges to zero in probability. The expected prediction loss consists of two additive terms: the first being the bias term, and the second being the variance term. Large models tend to reduce the bias but inflate the variance (*overfitting*), while small models tend to reduce the variance but increase the bias (*underfitting*) for a given fixed dataset. Suppose that "all models are wrong", meaning that the data generating model is not included in the model class. Usually, the bias is non-vanishing (with n) for a fixed model complexity (say d), and it is approximately a decreasing function of d; while on the other hand, the variance vanishes at rate n^{-1} for a fixed d, and it is an increasing function of d. Suppose for example that the bias and variance terms are approximately $c_1 \gamma^{-d}$ and $c_2 d/n$, respectively, for some positive constants c_1, c_2, γ. Then the optimal d is at the order of $\log(n)$.

In view of the above arguments, as more data become available, the model complexity need to be enlarged in order to strike a balance between bias and variance (or *approach the LoL*). To illustrate, we generated $n = 100, 200$ data from a logistic regression model, where coefficients are $\beta_i = 10/i$ and covariates x_i's are independent standard Gaussian (for $i = 1, \ldots, 100$). We consider the nested model class $\mathcal{A}_n = \{\{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, 50\}\}$, and the loss function is chosen to be the negative log-likelihood. We summarize the results in Fig. 2. As model complexity increases, the "goodness of fit" (measured by in-sample loss) improves (Fig. 2a), while the "predictive power" (measured by expected prediction loss) first improves and

Table I: Some common loss functions in addition to negative log-likelihood

Name	quadratic	exponential	hinge	perceptron	logistic
Formula	$(y-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})^2$	$e^{-y\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}}$	$\max\{0, 1-y\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\}$	$\max\{0, -y\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\}$	$\log(1+e^{-y\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}})$
Domain	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \{0, 1\}$

then deteriorates after some "optimal dimension" (Fig. 2b). Moreover, the optimal dimension becomes larger as sample size increases. It means that better fitting does not mean better predictive power, and large data sizes requires the search over a larger model class.



(a) The "goodness of fit" of each model under data size $n = 100$ (solid blue) and $n = 200$ (dash red).

(b) The expected prediction loss (numerically computed using independently generated data) of each model under data size $n = 100$ (solid blue) and $n = 200$ (dash red).

Figure 2: Experiment showing the "bigger models for bigger data" phenomena that is almost ubiquitous in statistical prediction and machine learning tasks.

The outline of the paper is given as follows. In Section II, we propose a computationally efficient method that determines the most appropriate learning model as more data become available. We prove that the LoL can be asymptotically approached under some regularity assumptions. In Section III, we demonstrate the applications of the proposed methodology to generalized linear models and neural networks, in order to select the variables/neurons with optimal predictive power and low computational cost.

II. Limit of Learning

A. Notation

Let $\mathcal{A}_n$, α, $d_n[\alpha]$, $\mathcal{H}_n[\alpha] \subset \mathbb{R}^{d_n[\alpha]}$ denote respectively a set of candidate models, a candidate model, its dimension, its associated parameter space. Let $d_n \triangleq \max_{\alpha\in\mathcal{A}_n} d_n[\alpha]$ denote the dimension of the largest candidate model. We shall frequently use subscript n to emphasize the dependency on n, and include an α in the arguments of many variables or functions in order to emphasize their dependency on the model (and parameters space) under consideration. For a measurable function $f(\cdot)$, we define $E_n f(\cdot) = n^{-1}\sum_{i=1}^{n} f(\boldsymbol{z}_i)$. For example, $E_n \boldsymbol{l_n}(\cdot, \boldsymbol{\theta}; \alpha) = n^{-1}\sum_{i=1}^{n} \boldsymbol{l_n}(\boldsymbol{z}_i, \boldsymbol{\theta}; \alpha)$. We let $\boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) \triangleq \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha)$, and $\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) \triangleq \nabla^2_{\boldsymbol{\theta}} l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha)$, which are respectively measurable vector-valued and matrix-valued functions of $\boldsymbol{\theta}$. We define the matrices

$$V_n(\boldsymbol{\theta}; \alpha) \triangleq E_* \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)$$
$$J_n(\boldsymbol{\theta}; \alpha) \triangleq E_* \bigl\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha) \times \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)^{\mathrm{T}}\bigr\}$$

Recall the definition of $\mathcal{L}_n[\alpha]$. Its sample analog (also referred to as the *in-sample loss*) is defined by $\hat{\mathcal{L}}_n[\alpha] \triangleq E_n l_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)$. Similarly, we define

$$\hat{V}_n(\boldsymbol{\theta}; \alpha) \triangleq E_n \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)$$
$$\hat{J}_n(\boldsymbol{\theta}; \alpha) \triangleq E_n \bigl\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha) \times \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)^{\mathrm{T}}\bigr\}$$

Throughout the paper, the vectors are arranged in column and marked in bold. Let int(S) denote the interior of a set S. Let $\|\cdot\|$ denote Euclidean norm of a vector or spectral norm of matrix. For any vector $\boldsymbol{c} \in \mathbb{R}^d$ $(d \in \mathbb{N})$ and scalar $r > 0$, let $B(\boldsymbol{c}, r) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{c}\| \le r\}$. For a positive semidefinite matrix V and a vector $\boldsymbol{x}$ of the same dimension, we shall abbreviate $\boldsymbol{x}^{\mathrm{T}} V \boldsymbol{x}$ as $\|\boldsymbol{x}\|_V^2$. For a given probability measure P_* and a measurable function m, let $\|m\|_{P_*} \triangleq (E_* m^2)^{1/2}$ denote the $L_2(P_*)$-norm. Let $\text{eig}_{\min}(V)$ (resp. $\text{eig}_{\max}(V)$) denote the smallest (resp. largest) eigenvalue of a symmetric matrix V. For a sequence of scalar random variables f_n, we write $f_n = o_p(1)$ if $\lim_{n\to\infty} f_n = 0$ in probability, and $f_n = O_p(1)$, if it is stochastically bounded.

We use $\to$ and $\to_p$ to respectively denote the deterministic and in probability convergences. Unless stated explicitly, all the limits throughout the paper are with respect to $n \to \infty$ where n is the sample size.

B. Approaching the LoL – Selection Procedure

To obtain the optimal predictive power, an appropriate model selection procedure is necessary to strike a balance between the *goodness of fit*, and *model complexity* based on the observed data. The basic idea of penalized selection is to impose an additive penalty term to the in-sample loss (i.e.

goodness of fit), so that larger models are more penalized. In this paper, we follow the aphorism that "all models are wrong", and assume that the model class under consideration is mis-specified.

Definition 3 (Efficient learning). *Our goal is to select $\hat{\alpha}_n \in \mathcal{A}_n$ that is asymptotically efficient, in the sense that*

$$\frac{\mathcal{L}_n[\hat{\alpha}_n]}{\min_{\alpha \in \mathcal{A}_n} \mathcal{L}_n[\alpha]} \rightarrow_p 1 \tag{5}$$

as $n \rightarrow \infty$.

Note that this requirement is weaker than selecting the exact optimal model $\arg\min_{\alpha \in \mathcal{A}_n} \mathcal{L}_n[\alpha]$. Similar definition has been adopted in the study of the optimality of AIC in the context of autoregressive order selection [6] and variable selection in linear regression models [7]. We propose to use the following penalized model selection procedure, which generalizes TIC from negative log-likelihood to general loss functions.

Generalized TIC (GTIC) procedure: Given data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ and a specified model class $\mathcal{A}_n$. We select a model $\hat{\alpha} \in \mathcal{A}_n$ in the following way: 1) for each $\alpha \in \mathcal{A}_n$, find the minimal loss estimator $\hat{\boldsymbol{\theta}}_n[\alpha]$ defined in (1), and record the minimum as $\hat{\mathcal{L}}_n[\alpha]$; 2) select $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}_n} \mathcal{L}_t^c[\alpha]$, where

$$\mathcal{L}_t^c[\alpha] \triangleq \hat{\mathcal{L}}_n[\alpha] + n^{-1} tr\bigl\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)^{-1} \hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\bigr\}. \tag{6}$$

Remark 1 (TIC and GTIC). *The quantity $\mathcal{L}_t^c[\alpha]$, also referred to as the corrected prediction loss, can be calculated from data, and it serves as a surrogate of the expected prediction loss $\mathcal{L}_n[\alpha]$ that is usually not analytically computable. The in-sample loss $\hat{\mathcal{L}}_n[\alpha]$ cannot be directly used as an approximation for $\mathcal{L}_n[\alpha]$, because it uses the sample approximation twice: once in the estimation of $\boldsymbol{\theta}_n^*$, and then in the approximation of $E_* l_n(\cdot, \boldsymbol{\theta}; \alpha)$ using $E_n l_n(\cdot, \boldsymbol{\theta}; \alpha)$ (the law of large numbers). For example, in a nested model class, the largest model always has the least $\hat{\mathcal{L}}_n[\alpha]$ (i.e. fits data the best). But as we discussed in the introduction, $\mathcal{L}_n[\alpha]$ is typically decreasing first and then increasing as the dimension increases.*

TIC [3] was heuristically derived as an alternative of AIC, also from an information theoretical point of view (using Kullback-Leibler divergence) [8], assuming model mis-specification. It does not appear to be widely appreciated nor used [9] compared with other information criteria such as AIC or Bayesian information criterion (BIC) [10]. In terms of provable asymptotic performance, only AIC is known to be asymptotically efficient for variable selection in linear regression models [7] and autoregressive order selection in linear time series models [6], [11], in mis-specified settings. It can be shown that TIC is close to AIC in linear models (and exactly the same for fixed-design regressions). Conceptually, TIC was proposed as a surrogate for AIC in general mis-specified settings. However, the theoretical optimality of AIC and TIC in the general context remains unknown.

Why should TIC be preferred over AIC in nonlinear models in general? Intuitively speaking, TIC has the potential of exploiting the nonlinearity while AIC does not. Recall our Example 2 in the introduction, with loss being the negative log-likelihood. It is well known from machine learning practice that neural network structures play a key role in effective prediction. However, information criteria such as AIC impose the same amount of penalty as long as the number of neurons remains the same, regardless of how neurons are configured.

In this paper, we extend the scope of allowable loss functions, and theoretically justify the use of GTIC (and thus TIC). Under some regularity conditions (elaborated in the Appendix), we shall prove that the $\hat{\alpha}_n$ selected by the GTIC procedure is asymptotically efficient (in the sense of Definition 3). This is formally stated as a theorem in Subsection II-D. Our theoretical results extend some existing statistical theories on AIC for linear models. We note that the technical analysis of high dimensional (non) linear model classes is highly nontrivial. We will develop some new technical tools in the Appendix, which may be interesting on their own rights.

C. Related Work

A wide variety of model selection techniques have been proposed in the past fifty years, motivated by different viewpoints and justified under various circumstances. State-of-art methods can be roughly categorized into two classes, the penalized selection and cross-validation. Examples are final prediction error criterion [12], AIC [1], [2], BIC [10] (and its finite sample counterpart Bayes factor [13]), minimum description length criterion [14], Hannan and Quinn criterion [15], predictive minimum description length criterion [16], [17], C_p method [18], generalized information criterion (GIC$_{\lambda_n}$) with $\lambda_n \rightarrow \infty$ [4], [19], [20], generalized cross-validation method (GCV) [21], and the bridge criterion (BC) [5], [22].

Is Cross-Validation Really The Best Choice?

The basic idea of cross-validation (CV) is to split the data into two parts, one for training and one for testing. The model with the best testing performance is selected, in the hope that it will perform well for future data as well. It is a common practice to apply 10-fold CV, 5-fold CV, 2-fold CV, or 30%-for-testing. In general, the advantages of CV method are its stability and easy implementation. However, it has been shown that only the delete-d CV method with $\lim_{n \rightarrow \infty} d/n = 1$ [23]–[26], or the delete-1 CV method [27] (or leave-one-out, LOO) can exhibit asymptotic (large sample) optimality. In fact, the former CV exhibits the same asymptotic behavior as BIC, which is typically consistent in a well-specified model class (i.e. it contains the true data generating model), but is suboptimal in a mis-specified model class. The latter CV is shown to be asymptotically equivalent to AIC and GCV if $d_n[\alpha] = o(n)$ [27], which is asymptotically efficient in a mis-specified model class, but usually overfits in a well-specified model class. We refer to [4], [5], [22], [28], [29] for more detailed discussions on the discrepancy and reconciliation of the two types of selection criteria.

In other words, common folklore that advocates the use of k-fold or 30%-for-testing CV are all asymptotically suboptimal (in the sense of Definition 3), even in linear regression models [4]. Since the only optimal CV is LOO-type (in mis-specified settings), it is more appealing to apply AIC or TIC that gives the same asymptotic performance and *significantly reduces the computational complexity* by n times. For general (mis-specified) nonlinear model class, we shall prove that GTIC procedure asymptotically approaches the LoL. While the asymptotic performance of LOO is not clear in that case, it is typically more complex to implement. To demonstrate that, we shall provide some experimental studies in the Appendix. As a result, the GTIC procedure can be a promising competitor of various types of standard CVs adopted in practice.

D. Asymptotic Analysis of the GTIC Procedure

We need the following assumptions for asymptotic analysis.

Assumption 1. *Data $\boldsymbol{Z}_i, i = 1, \dots, n$ are independent and identically distributed (i.i.d.).*

Assumption 2. *For each model $\alpha \in \mathcal{A}_n$, $\boldsymbol{\theta}_n^*[\alpha]$ (as was defined in (3)) is in the interior of the compact parameter space $\mathcal{H}_n[\alpha]$, and for all $\varepsilon > 0$ we have*

$$\liminf_{n\to\infty} \inf_{\alpha\in\mathcal{A}_n} \biggl(\inf_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]:\|\boldsymbol{\theta}-\boldsymbol{\theta}_n^*[\alpha]\|\geq\varepsilon} E_*\ell_n\bigl(\cdot,\boldsymbol{\theta};\alpha\bigr) - E_*\ell_n\bigl(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha\bigr) \biggr) \geq \eta_\varepsilon$$

for some constant $\eta_\varepsilon > 0$ that depends only on ε. Moreover, we have

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]} \Bigl| E_n\ell_n\bigl(\cdot,\boldsymbol{\theta};\alpha\bigr) - E_*\ell_n\bigl(\cdot,\boldsymbol{\theta};\alpha\bigr) \Bigr| \to_p 0,$$

as $n \to \infty$, and $\ell_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha)$ is twice differentiable in int$(\mathcal{Z})$ for all n, $\alpha \in \mathcal{A}_n$.

Assumption 3. *There exist constants $\tau \in (0, 0.5)$ and $\delta > 0$ such that*

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\delta)} n^{\tau} \Bigl\| E_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha) - E_*\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha) \Bigr\| = O_p(1).$$

Additionally, the map $\boldsymbol{\theta} \mapsto E_*\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)$ is differentiable at $\boldsymbol{\theta} \in$ int$(\mathcal{H}_n[\alpha])$ for all n and $\alpha \in \mathcal{A}_n$.

Assumption 4. *There exist constants $c_1, c_2 > 0$ such that*

$$\liminf_{n\to\infty} \min_{\alpha\in\mathcal{A}_n} eig_{\min}(V_n(\boldsymbol{\theta}_n^*;\alpha)) \geq c_1,$$
$$\limsup_{n\to\infty} \max_{\alpha\in\mathcal{A}_n} eig_{\max}(V_n(\boldsymbol{\theta}_n^*;\alpha)) \leq c_2.$$

Assumption 5. *There exist constants $r > 0$, $\gamma > 1$, and measurable functions $m[\alpha] : \mathcal{Z} \to \mathbb{R}^+ \cup \{0\}$, $\boldsymbol{z} \mapsto m[\alpha](\boldsymbol{z})$ for each $\alpha \in \mathcal{A}_n$, such that for all n and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in B(\boldsymbol{\theta}_n^*[\alpha], r)$,*

$$\|\boldsymbol{\psi}_n(\boldsymbol{z},\boldsymbol{\theta}_1;\alpha) - \boldsymbol{\psi}_n(\boldsymbol{z},\boldsymbol{\theta}_2;\alpha)\| \leq m_n[\alpha](\boldsymbol{z})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \tag{7}$$

$$E_*m_n[\alpha] < \infty. \tag{8}$$

Moreover, we have

$$\max\Bigl\{ d_n^{\gamma}\ card(\mathcal{A}_n)^{\gamma/2},\ d_n\sqrt{\log\{d_n card(\mathcal{A}_n)\}} \Bigr\} \times n^{-\tau} \Bigl\| \sup_{\alpha\in\mathcal{A}_n} m_n[\alpha] \Bigr\|_{P_*} \to 0, \tag{9}$$

and for all n,

$$E_* \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]} \|\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)\| < \infty. \tag{10}$$

Assumption 6. *There exists a constant $\delta > 0$ such that*

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\delta)} \|\hat{J}_n(\boldsymbol{\theta};\alpha) - J_n(\boldsymbol{\theta};\alpha)\| \to_p 0, \tag{11}$$

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\delta)} \|\hat{V}_n(\boldsymbol{\theta};\alpha) - V_n(\boldsymbol{\theta};\alpha)\| \to_p 0, \tag{12}$$

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\delta)} \|V_n(\boldsymbol{\theta};\alpha) - V_n(\boldsymbol{\theta}';\alpha)\| \to_p 0. \tag{13}$$

We define

$$\boldsymbol{w}_n[\alpha] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}_n(\boldsymbol{z}_i, \boldsymbol{\theta}_n^*[\alpha]; \alpha).$$

Clearly, $\boldsymbol{w}_n[\alpha]$ has zero mean and variance matrix $J_n(\boldsymbol{\theta}_n^*[\alpha]; \alpha)$, and thus

$$E_*\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}} = tr\bigl\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\bigr\}.$$

Assumption 7. *Suppose that the following regularity conditions are satisfied.*

$$\inf_{\alpha\in\mathcal{A}_n} n^{2\tau}\mathcal{R}_n[\alpha] \to \infty, \tag{14}$$

$$\sup_{\alpha\in\mathcal{A}_n} \frac{d_n[\alpha]}{n\mathcal{R}_n[\alpha]} \to 0. \tag{15}$$

Moreover, there exists a fixed constant $m_1 > 0$ such that

$$\sum_{\alpha\in\mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-2m_1} \sum_{\alpha\in\mathcal{A}_n} E_*\bigl\{l_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha) - E_*l_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)\bigr\}^{2m_1} \to 0, \tag{16}$$

there exists a fixed constant $m_2 > 0$ such that

$$\sum_{\alpha\in\mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-2m_2} \sum_{\alpha\in\mathcal{A}_n} E_*\biggl[\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}} - tr\bigl\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\bigr\}\biggr]^{2m_2} \to 0, \qquad (17)$$

and there exists a fixed constant $m_3 > 0$ such that

$$\limsup_{n\to\infty} \sum_{\alpha\in\mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-m_3}\{E_*\|\boldsymbol{w}_n[\alpha]\|^{m_3} + E_*\|\boldsymbol{w}_n[\alpha]\|^{2m_3}\} < \infty. \qquad (18)$$

Remark 2 (Intuitive explanation of each assumption). *Assumption 1 is standard for theoretical analysis and for some practical applications. In the context of regression analysis, it corresponds to the random design. In our technical proofs, it is possible to extend the assumption of i.i.d. to strong mixing which is more commonly assumed for time series data.*

Assumption 2 is the counterpart of the "separated mode" and "uniform law of large numbers" conditions that have been commonly required in proving the consistency of maximum likelihood estimator for classical statistical models [30, Theorem 5.7]. The $\boldsymbol{\theta}_n^*[\alpha]$ can be interpreted as the oracle optimum under model α, or the "projection" of true data generating model onto α.

Assumption 3 is a weaker statement to the central limit theorem (and its extension to Donsker classes) in a classical (non-high dimensional) setting. In our high dimensional setting, the assumption ensures that each projected model $\boldsymbol{\theta}_n^*[\alpha]$ behaves regularly. It implicitly builds a relation between d_n (the dimension of the largest candidate models) and n.

Assumption 4 assumes that the second derivative of the expected prediction loss has bounded eigenvalues at the optimum $\boldsymbol{\theta}_n^*[\alpha]$. It is useful because our asymptotic analysis requires "well-behaved" Taylor expansion up to the second order.

Assumption 5 is a Lipschitz-type condition. Similar but simpler forms of this have been used in classical analysis of asymptotic normality [30, Theorem 5.21]. We note that the condition (9) explicitly requires that the largest dimension d_n and the candidate size card($\mathcal{A}_n$) do not grow too fast.

Assumption 6 requires that the sample analogs of matrices $J_n(\boldsymbol{\theta};\alpha)$ and $V_n(\boldsymbol{\theta};\alpha)$ are asymptotically close to the truth (in spectral norm) in a neighborhood of $\boldsymbol{\theta}_n^*[\alpha]$. In the classical setting, it is guaranteed by the law of large numbers (applied to each matrix element). Assumption 6 also requires the continuity of $V_n(\boldsymbol{\theta};\alpha)$ in a neighborhood of $\boldsymbol{\theta}_n^*[\alpha]$.

In Assumption 7, the assumptions in (14), (15), and (18) mean that the model class is "quite mis-specified", since they require that the risks $\mathcal{R}_n[\alpha]$ (which are deterministic numbers) cannot be too small. The assumptions in equalities (16) and (17) are central moment constraints that control both the regularity of loss functions and the speed of growth of the model class.

Theorem 1. *Suppose that Assumptions 1-7 hold. Then the $\hat{\alpha}_n$ selected by GTIC procedure is asymptotically efficient (in the sense of Definition 3).*

Remark 3 (Sketch of Technical Ideas). *Classical asymptotic analysis typically relies on a type of uniform convergence of empirical process around $\boldsymbol{\theta}_n^*[\alpha]$. Because our functions are vector valued with dimension depending on data size, we cannot directly use state-of-art technical tools such as [30, Theorem 19.28]. The classical proof by White [31] (in proving asymptotic normality in mis-specified class) cannot be directly adapted, either, for parameter spaces that depend on n. We therefore need to develop some new technical tools in the proof. Due to page limits, the detailed proof will be included in a full version of this paper.*

III. Numerical Experiments

In view of Examples 1 and 2, the model classes under consideration are logistic regression and single-layer feed-forward neural networks.

We also implemented and released a python package "gtic" at *https://pypi.python.org/pypi/gtic*, in which we build a tensor graph of GTIC upon the *theano* platform. Users can simply input in their tensor variables of data and parameters, and obtain the GTIC instantly.

A. Logistic Regression Models

We consider the model class to be logistic regression. We generate data from a logistic regression model, where the coefficient vector is $\boldsymbol{\beta} = 10 \times [1^{-1.5}, \ldots, 100^{-1.5}]^{\mathrm{T}}$, and covariates $x_1, \ldots, x_{100}$ are independent standard Gaussian. Suppose that we sequentially obtain and learn the data, starting from $t = 10$, and then $t = 11, \ldots, 100$. We restrict the maximum dimension of candidate models to be $\lfloor\sqrt{t}\rfloor$ at time t (see our theoretical assumptions). Here, a model of dimension d means that the first d covariates are nonzero. The model class is nested because a small model is a special case of a large model. We summarize the results in Fig. 3 and 4.

To illustrate the efficiency of GTIC, we first simulate model selection results with batch data. We numerically compute the true prediction loss of each trained model (obtained by testing on a large dataset), and then identify the optimal model (with the least loss). Fig. 3 is a heat-map showing the true prediction loss of estimated candidate models of each dimension (y-axis) at each data size (x-axis), where the black dots indicate the model of optimal loss at each data size. The true loss is numerically computed from independently generated test data. In Fig. 4a, we compare the performance of GTIC to different types of CV. Holdout takes 70% data for training and tests on 30% data. It fluctuates throughout the experiment, and most of time it yields the

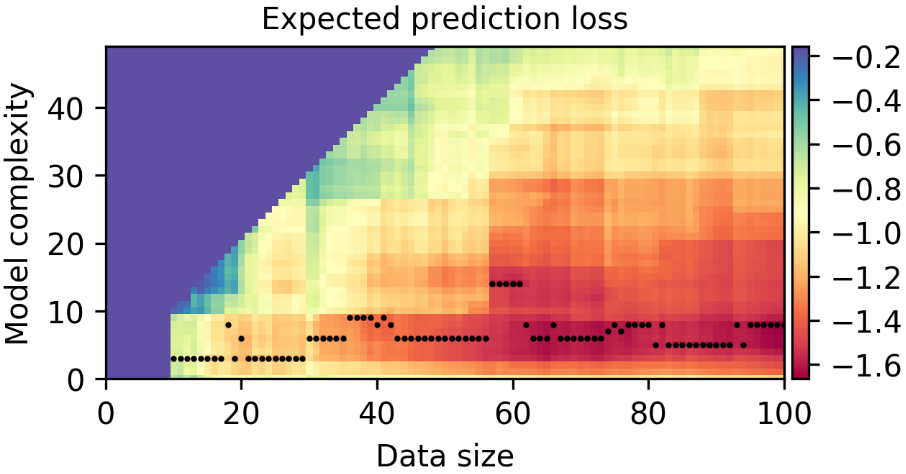
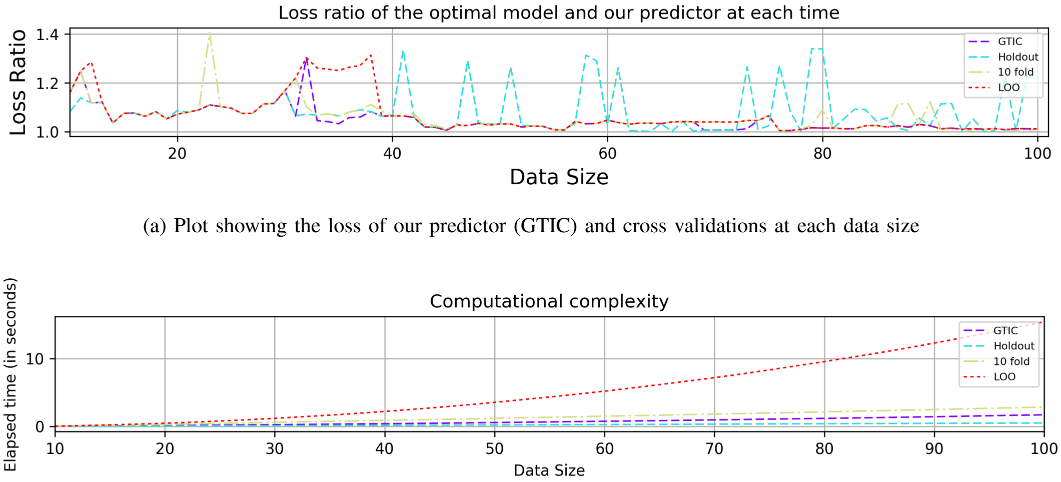


Figure 3: Experiment 1: Heat-map showing the true prediction loss of estimated candidate models of each dimension (y-axis) at each data size (x-axis), where the black dots indicate the model of optimal loss at each data size. The true loss is numerically computed from independently generated test data.



(a) Plot showing the loss of our predictor (GTIC) and cross validations at each data size

(b) Plot showing the computational costs.

Figure 4: Experiment 1: logistic regression models

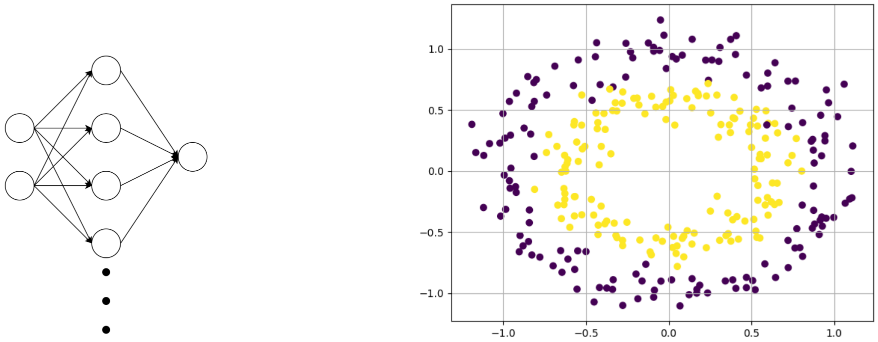
worst performance. GTIC, 10-fold CV and LOO perform well in this experiment. Although the optimal model of each data size is not always identical to our selected model, their prediction losses are very close. This result is consistent with our definition of efficient learning.

The computation cost of all approaches is provided in Fig. 4b. As shown in the figures, under logistic regression, GTIC is slightly better than 10-fold CV but worse than Holdout. Indeed, we need to compute the penalty term in GTIC. However, depending on the problem and data, we may need different number of folds for CV in order to have a satisfactory result. Since GTIC performs almost as well as LOO and 10-fold CV, we suggest using GTIC instead of guessing the optimal number of fold for CV. With GTIC, we do not need to sacrifice much on computation cost, but can still achieve theoretically justifiable result which is as good as LOO.

B. Neural Networks

We consider the model class to be single-layer feed-forward neural networks (see Fig. 5a). Neural networks are inherently miss-specified models.

Data are generated from the following way. A set of two-dimensional data are uniformly sampled from two circles (with radius ratio 0.6), corrupted by independent Gaussian noise with mean 0 and variance 0.1 (generated from python



(a) An illustration of the single-layer feed-forward neural network

(b) A set of 300 data uniformly sampled from two circles corrupted by Gaussian noise ($\mu = 0$, $\sigma^2 = 0.1$, radius ratio $= 0.6$)

Figure 5: Experiment 2: neural networks

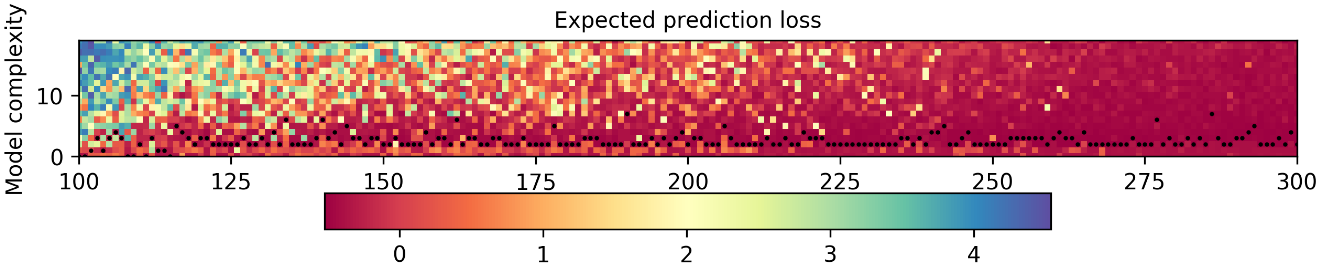


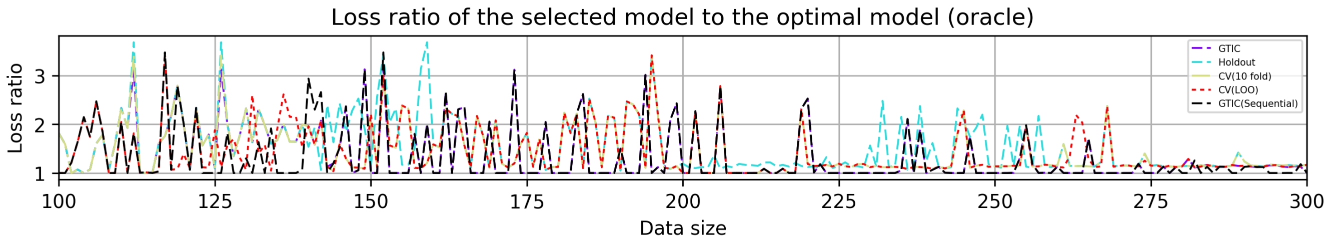
Figure 6: Experiment 2: Heat-map showing the prediction loss of estimated candidate models of each dimension (y-axis) at each data size (x-axis), where the black dots indicate the model of optimal loss at each data size.

package sklearn dataset “make_circle”). The goal is to correctly classify the data into two groups, the larger and smaller rings. Since we have two-dimensional data, our input dimension for the model is two. And because we want to classify into two groups, the output dimension is one. In this experiment, the model complexity of our model is the number of hidden nodes in the single hidden layer.

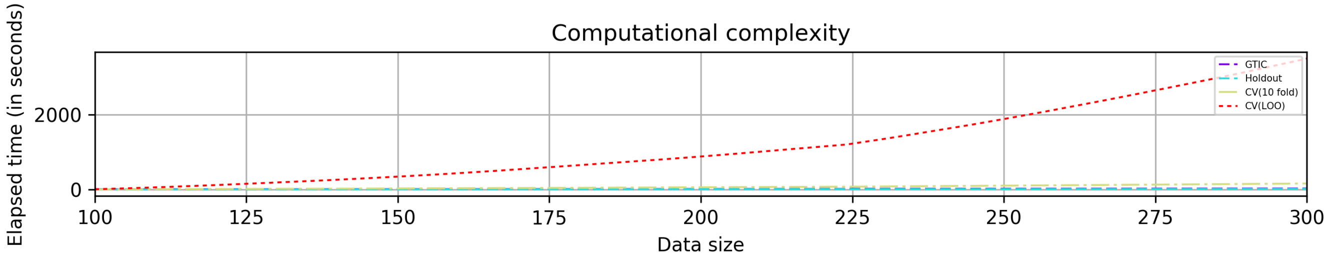
We sequentially obtain and learn the data, starting from $t = 100$, then $t = 101, \ldots, 300$. We start from 100 samples because neural networks are likely to converge to a local optimal for small sample size. The path of expansion in this case is the number of hidden nodes in the single hidden layer. Since the true model is not linearly-separable, we do need more than one hidden node to accurately classify the data. We restrict the maximum number of hidden nodes to be $\sqrt{t}$/(input dimension) due to our assumption. The path of expansion is in increasing order of the number of hidden nodes, since having a small number of hidden nodes is a special case of having more number of hidden nodes.

Similarly, the optimal model (oracle) is obtained by testing the trained model on a large dataset. The oracle loss of different models at different data size is shown in Fig. 6. With a small sample size, the cost of overfitting is considerably high. When we have enough samples for training, the cost of overfitting decreases. This effect may also depend on the dimension of input data and labels. In Fig. 7a, the loss ratio varies quite a lot when the sample size is small, but gradually converges. This is partially because the influence of overfitting on the predictive power decreases as sample size increases. In other words, even if we choose a model that is slightly overfitting, the loss ratio is still close to one.

The computational cost is shown in Fig. 7b. As expected, the computation of 10-fold CV and LOO increases significantly. However, since we can analytically compute the gradient and hessian involved in the GTIC penalty term, using symbolic expression computation software and saving them on the disk in advance, our computation cost is almost constant at each time step. Therefore, our overall computational cost is almost identical to Holdout. Furthermore, we can utilize warm-start in our implementation, which is a benefit that CV cannot enjoy in naive sequential model selection framework. Therefore, we encourage the use of GTIC in a sequential setting.



(a) Heat-map showing the prediction loss of estimated candidate models of each dimension (y-axis) at each data size (x-axis), where the black dots indicate the model of optimal loss at each data size.



(b) Heat-map showing our predictive weights over the candidate models (y-axis) at each data size (x-axis).

Figure 7: Experiment 2: neural networks

IV. Conclusions

In this work, we addressed the challenges of model selection for parametric model classes with high dimension and/or small samples. We studied a method (GTIC) to approach the limit of learning, in the sense that the predictive power of the selected model is asymptotically close to the best offered from a model class (which can depend on data size). The theoretical analysis of GTIC justifies the use of TIC for general mis-specified model classes, and extends some technical results for classical analysis on AIC in linear models. In summary, the proposed methodology is asymptotically optimal and practically useful, and may be a promising competitor of various types of cross-validations.

Acknowledgement

This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers W911NF-14-1-0508, N66001-15-C-4028, and W911NF-16-1-0561.

References

[1] H. Akaike, “Statistical predictor identification,” *Ann. Inst. Statist. Math.*, vol. 22, no. 1, pp. 203–217, 1970.

[2] ——, “Information theory and an extension of the maximum likelihood principle,” in *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.

[3] K. Takeuchi, “Distribution of informational statistics and a criterion of model fitting,” *Suri-Kagaku (Mathematical Sciences)*, no. 153, pp. 12–18, 1976.

[4] J. Shao, “An asymptotic theory for linear model selection,” *Statist. Sinica*, vol. 7, no. 2, pp. 221–242, 1997.

[5] J. Ding, V. Tarokh, and Y. Yang, “Optimal variable selection in regression models,” *http://jding.org/jie-uploads/2017/03/variable-selection.pdf*, 2016.

[6] R. Shibata, “Asymptotically efficient selection of the order of the model for estimating parameters of a linear process,” *Ann. Statist.*, vol. 8, no. 1, pp. 147–164, 1980.

[7] ——, “An optimal selection of regression variables,” *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981.

[8] S. Konishi and G. Kitagawa, “Generalised information criteria in model selection,” *Biometrika*, vol. 83, no. 4, pp. 875–890, 1996.

[9] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

[10] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[11] C.-K. Ing and C.-Z. Wei, “Order selection for same-realization predictions in autoregressive processes,” *Ann. Statist.*, vol. 33, no. 5, pp. 2423–2474, 2005.

[12] H. Akaike, “Fitting autoregressive models for prediction,” *Ann. Inst. Statist. Math.*, vol. 21, no. 1, pp. 243–247, 1969.

[13] G. Casella, F. J. Girón, M. L. Martínez, and E. Moreno, “Consistency of bayesian procedures for variable selection,” *Ann. Stat.*, pp. 1207–1228, 2009.

[14] M. H. Hansen and B. Yu, “Model selection and the principle of minimum description length,” *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[15] E. J. Hannan and B. G. Quinn, “The determination of the order of an autoregression,” *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 190–195, 1979.

[16] J. Rissanen, “Stochastic complexity and modeling,” *Ann. Statist.*, pp. 1080–1100, 1986.

[17] C.-Z. Wei, “On predictive least squares principles,” *Ann. Statist.*, pp. 1–42, 1992.

[18] C. L. Mallows, “Some comments on c p,” *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[19] R. Nishii *et al.*, “Asymptotic properties of criteria for selection of variables in multiple regression,” *Ann. Stat.*, vol. 12, no. 2, pp. 758–765, 1984.

[20] R. Rao and Y. Wu, “A strongly consistent procedure for model selection in a regression problem,” *Biometrika*, vol. 76, no. 2, pp. 369–374, 1989.

[21] P. Craven and G. Wahba, “Smoothing noisy data with spline functions,” *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1978.

[22] J. Ding, V. Tarokh, and Y. Yang, “Bridging AIC and BIC: a new criterion for autoregression,” *IEEE Trans. Inf. Theory*, 2017.

[23] S. Geisser, “The predictive sample reuse method with applications,” *J. Amer. Statist. Assoc.*, vol. 70, no. 350, pp. 320–328, 1975.

[24] P. Burman, “A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods,” *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989.

[25] J. Shao, “Linear model selection by cross-validation,” *J. Amer. Statist. Assoc.*, vol. 88, no. 422, pp. 486–494, 1993.

[26] P. Zhang, “Model selection via multifold cross validation,” *Ann. Stat.*, pp. 299–313, 1993.

[27] M. Stone, “An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion,” *J. R. Stat. Soc. Ser. B*, pp. 44–47, 1977.

[28] Y. Yang, “Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation,” *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[29] Y. Zhang and Y. Yang, “Cross-validation for selecting a model selection procedure,” *J. Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.

[30] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[31] H. White, “Maximum likelihood estimation of misspecified models,” *Econometrica*, pp. 1–25, 1982.