A PENALIZED METHOD FOR THE PREDICTIVE LIMIT OF LEARNING

Jie Ding, Enmao Diao, Jiawei Zhou, and Vahid Tarokh

ABSTRACT

Machine learning systems learn from and make predictions by building models from observed data. Because large models tend to overfit while small models tend to underfit 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. 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 (in the sense that the models' complexity can grow with data size). We released a python package with our proposed method for general usage like logistic regression and neural networks.

Index Terms— Cross-validation, Computational efficiency, Feature selection, High dimension, Limit of learning

1. 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. 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.

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 $l_n : \mathcal{Z} \times \mathcal{H}_n[\alpha] \to \mathbb{R}$, 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. 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

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

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. 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.*

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

J. Ding and V. Tarokh are with the Department of Electrical and Computer Engineering, Duke University. E. Diao and J. Zhou are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University.

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).

2. LIMIT OF LEARNING

2.1. 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 $V_n(\boldsymbol{\theta};\alpha) \triangleq E_*\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)$ and $J_n(\boldsymbol{\theta};\alpha) \triangleq E_*\big\{\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)\times\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)^{\mathrm{T}}\big\}$. Recall the definition of $\mathcal{L}_n[\alpha]$. Its sample analog (also referred to as the *in-sample loss*) is defined by $\hat{\mathcal{L}}_n[\alpha] \triangleq E_n l_n\big(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha\big)$. Similarly, we define $\hat{V}_n(\boldsymbol{\theta};\alpha) \triangleq E_n\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)$ and $\hat{J}_n(\boldsymbol{\theta};\alpha) \triangleq E_n\big\{\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)\times\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)^{\mathrm{T}}\big\}$. We let $\boldsymbol{\theta}^*_n[\alpha]$ denote the minimum loss parameter defined by

$$\boldsymbol{\theta}^*_n[\alpha] \triangleq \operatorname*{arg\,min}_{\theta\in\mathcal{H}_n[\alpha]} E_* l_n(\cdot,\boldsymbol{\theta};\alpha). \tag{2}$$

Throughout the paper, the vectors are arranged in column and marked in bold. Let $\mathrm{int}(S)$ 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 $\|\boldsymbol{x}\|^2_V$. 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 $\mathrm{eig}_{\min}(V)$ (resp. $\mathrm{eig}_{\max}(V)$) 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.

2.2. 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 misspecified.

Definition 3 (Efficient learning). *Our goal is to select $\hat{\alpha}_n\in\mathcal{A}_n$ that is asymptotically efficient, in the sense that $\mathcal{L}_n[\hat{\alpha}_n]/\min_{\alpha\in\mathcal{A}_n}\mathcal{L}_n[\alpha]\to_p 1$ as $n\to\infty$.*

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 [4] and variable selection in linear regression models [5]. 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 $\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}_n}\mathcal{L}^c_t[\alpha]$, where

$$\mathcal{L}^c_t[\alpha] \triangleq \hat{\mathcal{L}}_n[\alpha] + n^{-1}tr\big\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)^{-1}\hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\big\}.$$

2.3. 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 AIC [1, 2], BIC [6] (and its finite sample counterpart Bayes factor [7]), minimum description length criterion [8], predictive minimum description length criterion [9, 10], generalized information criterion, generalized cross-validation method (GCV) [11], and the bridge criterion (BC) [12].

Is Cross-Validation Really The Best Choice?

It is a common practice to apply 10-fold CV, 5-fold CV, 3-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 $\lim_{n\to\infty} d/n = 1$ [13–16], or the delete-1 CV method [17] (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 $d_n[\alpha] = o(n)$ [17], which is asymptotically efficient in a mis-specified model class, but usually overfits in a well-specified model class. We refer to [12, 18–20] for more detailed discussions on the discrepancy and reconciliation of the two types of selection criteria. 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 unclear in that case, typically it is more computationally cumbersome to implement LOO. As a result, the GTIC procedure can be a promising competitor of various types of standard CVs adopted in practice.

2.4. Asymptotic Analysis of the GTIC Procedure

We need the following assumptions for asymptotic analysis.

Assumption 1. *Data $\boldsymbol{Z}_i, i = 1, \ldots, n$ 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 (2)) is in the interior of the compact parameter space $\mathcal{H}_n[\alpha]$, and for all $\varepsilon > 0$ we have* $\liminf_{n\to\infty} \inf_{\alpha\in\mathcal{A}_n} \big(\inf_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]:\|\boldsymbol{\theta}-\boldsymbol{\theta}_n^*[\alpha]\|\geq\varepsilon} E_*\ell_n\big(\cdot,\boldsymbol{\theta};\alpha\big) - E_*\ell_n\big(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha\big)\big) \geq \eta_\varepsilon$ *for some constant $\eta_\varepsilon > 0$ that depends only on ε. Moreover, we have*

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

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*

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

is $O_p(1)$. 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*

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

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 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in B(\boldsymbol{\theta}_n^*[\alpha], r)$,*

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

and $E_*m_n[\alpha] < \infty$. Moreover, we have $\max\big\{d_n^\gamma\, \textit{card}(\mathcal{A}_n)^{\gamma/2}, d_n\sqrt{\log\{d_n \textit{card}(\mathcal{A}_n)\}}\big\} \times n^{-\tau}\Big\|\sup_{\alpha\in\mathcal{A}_n} m_n[\alpha]\Big\|_{P_*} \to 0$, *and for all n $E_* \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]}\|\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)\| < \infty$.*

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

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

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

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

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

$$\inf_{\alpha\in\mathcal{A}_n} n^{2\tau}\mathcal{R}_n[\alpha] \to \infty,\ \sup_{\alpha\in\mathcal{A}_n}\frac{d_n[\alpha]}{n\mathcal{R}_n[\alpha]} \to 0.$$

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

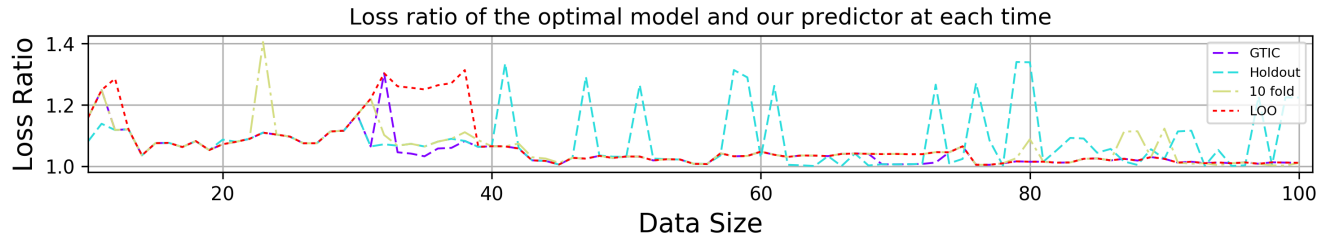
$$\sum_{\alpha\in\mathcal{A}_n}(n\mathcal{R}_n[\alpha])^{-2m_1}\sum_{\alpha\in\mathcal{A}_n}E_*\big\{l_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha) - E_*l_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\}^{2m_1} \to 0, \quad (3)$$

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

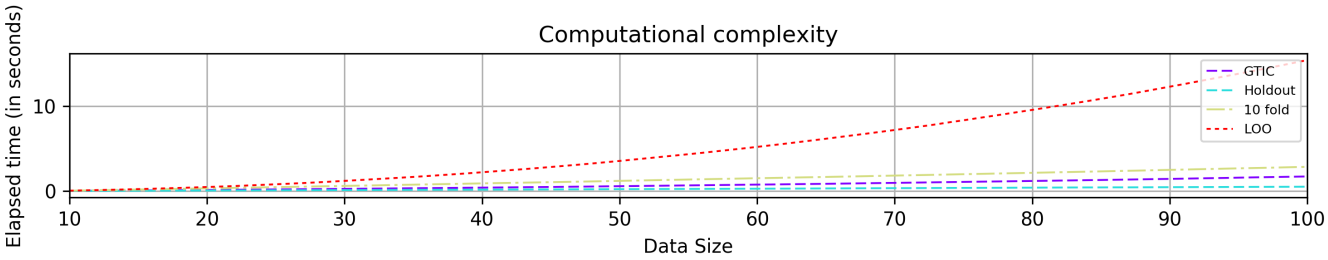
$$\sum_{\alpha\in\mathcal{A}_n}(n\mathcal{R}_n[\alpha])^{-2m_2}\sum_{\alpha\in\mathcal{A}_n}E_*\Big[\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}} - tr\big\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\}\Big]^{2m_2} \to 0, \quad (4)$$

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

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



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



(b) Plot showing the computational costs.

Fig. 1: Experiment 1: logistic regression models

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 1 (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 [21, Theorem 19.28]. The classical proof by White [22] (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.*

3. NUMERICAL EXPERIMENTS

The model classes under consideration are logistic regression. 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, applicable for both generalized linear models and single-layer feed-forward neural networks (not included due to space limitation). Users can simply provide their tensor variables of loss and parameters, and obtain the GTIC instantly.

We generate data from a logistic regression model, where the coefficient vector is $\boldsymbol{\beta} = 10 \times [1^{-1.5}, \ldots, n^{-1.5}]^{\mathrm{T}}$, and covariates $x_1, \ldots, x_n$ (with $n = 100$) are independent standard Gaussian. We restrict the maximum dimension of candidate models to be $\lfloor \sqrt{n} \rfloor$. 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. 1. 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). In Fig. 1a, 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 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. 1b. 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.

4. REFERENCES

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

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

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

[4] Ritei 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.

[5] Ritei Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981.

[6] Gideon Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[7] George Casella, F Javier Girón, M Lina Martínez, and Elias Moreno, "Consistency of bayesian procedures for variable selection," *Ann. Stat.*, pp. 1207–1228, 2009.

[8] Mark H Hansen and Bin Yu, "Model selection and the principle of minimum description length," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[9] Jorma Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, pp. 1080–1100, 1986.

[10] Ching-Zong Wei, "On predictive least squares principles," *Ann. Statist.*, pp. 1–42, 1992.

[11] Peter Craven and Grace Wahba, "Smoothing noisy data with spline functions," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1978.

[12] Jie Ding, Vahid Tarokh, and Yuhong Yang, "Bridging AIC and BIC: a new criterion for autoregression," *IEEE Trans. Inf. Theory*, 2017.

[13] Seymour Geisser, "The predictive sample reuse method with applications," *J. Amer. Statist. Assoc.*, vol. 70, no. 350, pp. 320–328, 1975.

[14] Prabir 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.

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

[16] Ping Zhang, "Model selection via multifold cross validation," *Ann. Stat.*, pp. 299–313, 1993.

[17] Mervyn 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.

[18] Jun Shao, "An asymptotic theory for linear model selection," *Statist. Sinica*, vol. 7, no. 2, pp. 221–242, 1997.

[19] Yuhong 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.

[20] Yongli Zhang and Yuhong Yang, "Cross-validation for selecting a model selection procedure," *J. Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.

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

[22] Halbert White, "Maximum likelihood estimation of misspecified models," *Econometrica*, pp. 1–25, 1982.