On Statistical Efficiency in Learning

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

***Abstract*—A central issue of many statistical learning problems is to select an appropriate model from a set of candidate models. 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. In this work, we address the critical challenge of model selection in order to strike a balance between the goodness of fit and model complexity, and thus to gain reliable predictive power. We consider the task of approaching the theoretical limit of statistical learning, meaning that the selected model has the predictive performance that is as good as the best possible model given a class of potentially mis-specified candidate models. We propose a generalized notion of Takeuchi's information criterion, and prove that the proposed method can asymptotically achieve the optimal out-sample prediction loss under reasonable assumptions. To our best knowledge, this is the first proof of the asymptotic property of Takeuchi's information criterion. Our proof applies for a wide variety of nonlinear models, loss functions, and high dimensionality (in the sense that the models' complexity can grow with sample size). The proposed method can be used as a computationally efficient surrogate for leave-one-out cross-validation. Moreover, for modeling streaming data, we propose an online algorithm that sequentially expands the model complexity in order to enhance selection stability and reduce computation cost. Experimental studies show that the proposed metthod has desirable predictive power and less computational cost compared to some existing methods. We also released a python package for applying the method to logistic regression and neural networks.**

***Index Terms*—Cross validation; Expert learning; Feature selection; Limit of learning; Model expansion.**

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 representations 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 an overly large model can lead to inflated variance in parameter estimation and thus 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.

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 $E_*(Y) = \mu = g(\beta_1 X_1 + \beta_2 X_2 + \cdots)$ where $g(\cdot)$ is a link function. 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 family 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, and the configuration of their connectivity. 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 following definition that quantifies the predictive power of each candidate model.

Definition 1 (Out-sample prediction loss): The loss function for each sample size n and $\alpha \in \mathcal{A}_n$ (model class) is a map $l_n(\cdot, \cdot; \alpha) : \mathcal{Z} \times \mathcal{H}_n[\alpha] \to \mathbb{R}$, 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. As Fig. 1 shows, for a loss function and a given dataset $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ 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, given by candidate model α, denoted by $\mathcal{L}_n(\alpha)$, the out-sample prediction loss, also referred to as the generalization error in machine learning, is defined by

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

Here, E_* denotes the expectation with respect to the distribution of a future unseen random variable $\boldsymbol{z}$ (conditional on the

J. Ding is with the School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA. E. Diao and V. Tarokh are with the Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA. J. Zhou is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02138 USA.

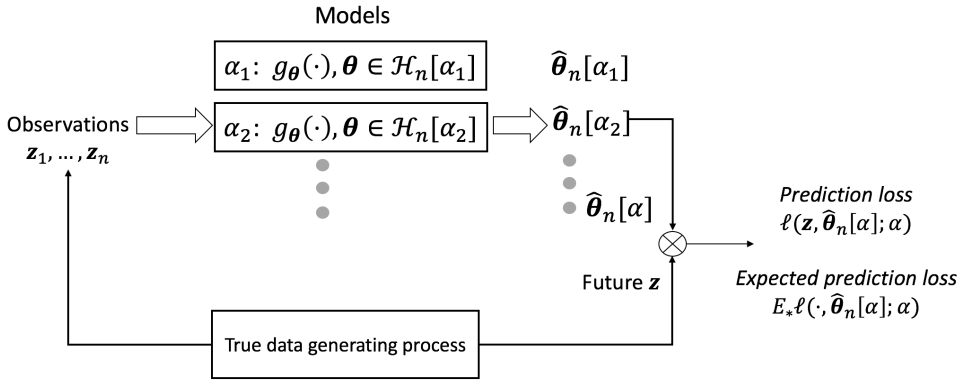


Fig. 1: Illustration of a typical learning procedure, where each candidate model α_j is trained in terms of $\hat{\boldsymbol{\theta}}_n[\alpha_j]$ in its parameter space $\mathcal{H}_n[\alpha_j]$, and then used to evaluate future data under some loss function $\ell(\cdot)$.

observed data). We also define the risk by

$$\mathcal{R}_n[\alpha] = E_{*,o}\mathcal{L}_n[\alpha],$$

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 statistical 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 $\min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n(\alpha)$, the optimal out-sample 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 sample size). It is worth noting that asymptotic analysis for a fixed number of candidate models with fixed dimensions are generally straightforward. Under some classical regularity conditions (e.g. [4, Theorem 19.28]), likelihood based principle usually selects the model that attains the smallest Kullback-Leibler divergence from the data generating model. However, our high dimensional setting considers models whose dimensions and parameter spaces may depend on sample size, and thus we cannot directly use those technical tools that have been used in classical asymptotic analysis for mis-specified modes. We will develop some new technical tools in our proof. Our theoretical results extend the classical statistical theory on AIC for linear (fixed-design) regression models to a broader range of generalized linear or nonlinear 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, in the sense that the number of candidate models or each model's complexity is allowed to grow with sample size? In the context of regression analysis, technical discussions that address the question have been elaborated in [5], [6]. 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 (Equality 44) that $\mathcal{L}_n[\alpha]$ may be expressed as

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

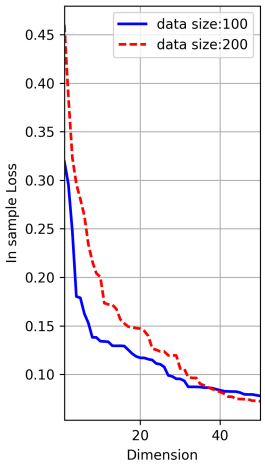
under some regularity conditions, where $V_n(\boldsymbol{\theta}; \alpha) \triangleq E_* \nabla^2_{\boldsymbol{\theta}} l_n(\cdot, \boldsymbol{\theta}; \alpha)$, and $o_p(1)$ is a sequence of random variables that converges to zero in probability. The out-sample prediction loss consists of two additive 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*

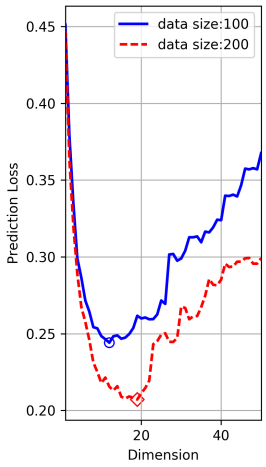
TABLE I: Some common loss functions in addition to negative log-likelihood

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	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \mathbb{R}$	$y \in \{0, 1\}$

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 the out-sample prediction loss) first improves and 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 sample sizes requires the search over a larger model class.



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



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

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

As data sequentially arrives, the selected model from our proposed method (and many other existing method such as cross validation) suffer from fluctuations (due to randomness). A conceptually appealing and computationally efficient way is to move from small model to larger models sequentially. Motivated by this, based on the proposed method, we further propose a sequential model expansion strategy that aims to facilitate interpretability of learning.

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 propose a model expansion techniques building upon a new online learning algorithm which we refer to as "graph-based" learning. The online learning algorithm may be interested on its own as it exploits graphical structure when updating the expert systems and computing the regrets. In Section IV, 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 finitely many candidate models (also called the model class), a candidate parametric 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 the sample size n, and include an α in the arguments of many variables or functions in order to emphasize their dependency on the model (and parameter 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

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

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

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

Throughout the paper, the vectors are arranged in column and marked in bold. Let $\|\cdot\|$ denote Euclidean norm of a vector or spectral norm of matrix. Let $\mathrm{int}(S)$ denote the interior of a set S. 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. Unless otherwise stated, E_* denotes the expectation with respect the true data generating process. Let $\mathrm{eig}_{\min}(V)$ (resp. $\mathrm{eig}_{\max}(V)$) denote the smallest (resp.

maximal) 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. For a fixed measurable vector-valued function $\boldsymbol{f}$, we define

$$\mathbb{G}_n \boldsymbol{f} \triangleq \sqrt{n}(E_n - E_*)\boldsymbol{f},$$

the empirical process evaluated at $\boldsymbol{f}$. For $a, b \in \mathbb{R}$, we write $a \lesssim b$ if $a \leq cb$ for a universal constant c. For a vector $\boldsymbol{a}$ or a vector-valued function $\boldsymbol{f}$, we let a_i or f_i denote the ith component.

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

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

as $n \to \infty$.

Note that this requirement is weaker than selecting the exact optimal model $\arg\min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n[\alpha]$. Also, the concept of asymptotic efficiency in model selection is reminiscent of its counterpart in parameter estimation theory. Similar definition has been adopted in the study of the optimality of AIC in the context of autoregressive order selection [7] and variable selection in linear regression models [8].

It is worth noting that the above definition is in the scope of the available data and a specified class of models. Because we are in a date-driven setting where it is unrealistic to compete with the best performance attainable with full knowledge of the underlying distribution, we chose the above rationale of efficient learning instead of using

$$\frac{\mathcal{L}_n[\hat{\alpha}_n]}{\min_{\alpha\in\mathcal{A}_n} E_* l_n\big(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha\big)} \to_p 1$$

whose denominator does not reveal the influence of finite-sample data. In other words, Definition 3 calls for a model whose predictive power can practically approach the best offered by the candidate models (i.e. the LoL in Definition 2).

A related but different school of thoughts is the so-called *structural risk minimization* in the literature of statistical learning theory. In that context, the out-sample prediction loss is usually bounded using in-sample loss plus a positive term (e.g. a function of the Vapnik-Chervonenkis (VC) dimension [9] for a classification model). A definitive treatment of this line of work can be found in, e.g., [10] and the references therein. The major difference of our setting compared with that in statistical learning is our (stronger) requirement that the selected model should exhibit prediction loss comparable to the best offered by the candidates. In other words, the positive term plus the in-sample loss should asymptotically approach the true out-sample loss (as sample size goes to infinity).

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 $\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n^c[\alpha]$, where

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

We note that the two additive terms on the right hand side of (6) represent the goodness of fit and the model complexity, respectively.

Remark 1 (TIC and GTIC): The quantity $\mathcal{L}_n^c[\alpha]$, also referred to as the corrected prediction loss, can be calculated from data, and it serves as a surrogate for the out-sample 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) [11], assuming model mis-specification. It does not appear to be widely appreciated nor used [12] compared with other information criteria such as AIC or Bayesian information criterion (BIC) [13]. In terms of provable asymptotic performance, only AIC is known to be asymptotically efficient for variable selection in linear regression models [8] and autoregressive order selection in linear time series models [7], [14], in mis-specified settings. It can be shown that TIC reduces to AIC in linear models. 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.

Penalized selection: Examples are final prediction error criterion [15], AIC [1], [2], BIC [13] (and its finite sample counterpart Bayes factor [16]), minimum description length criterion [17], Hannan and Quinn criterion [18], predictive minimum description length criterion [19], [20], C_p method [21], generalized information criterion (GIC$_{\lambda_n}$) with $\lambda_n \to \infty$ [5], [22], [23], generalized cross-validation method (GCV) [24], and the bridge criterion (BC) [6], [25].

Cross-validation (CV): The basic idea of 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, *is cross-validation really the best choice?*

In fact, it has been shown that only the delete-d CV method with $\lim_{n\to\infty} d/n = 1$ [26]–[29], or the delete-1 CV method [30] (or leave-one-out, LOO) can exhibit asymptotic (large sample) optimality. Specifically, 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 $d_n[\alpha] = o(n)$ [30], which is asymptotically efficient in a mis-specified model class, but usually overfits in a well-specified model class. We refer to [5], [6], [25], [31], [32] 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 [5]. 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 $\boldsymbol{Z}_i, i = 1, \ldots, n$ 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 (3)) is in the interior of the compact parameter space $\mathcal{H}_n[\alpha]$, and for all $\varepsilon > 0$ we have

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

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

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

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

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

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

$$\|\boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}_1; \alpha) - \boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}_2; \alpha)\| \leq m_n[\alpha](\boldsymbol{z}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \quad (7)$$
$$E_* m_n[\alpha] < \infty. \quad (8)$$

Moreover, we have

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

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

$$\sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\delta)} \|\hat{J}_n(\boldsymbol{\theta}; \alpha) - J_n(\boldsymbol{\theta}; \alpha)\| \to_p 0, \quad (10)$$

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

$$\lim_{\varepsilon\to 0} \sup_{\alpha\in\mathcal{A}_n} \sup_{\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]\cap B(\boldsymbol{\theta}_n^*[\alpha],\varepsilon)} \|V_n(\boldsymbol{\theta}; \alpha) - V_n(\boldsymbol{\theta}_n^*; \alpha)\| = 0. \quad (12)$$

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.

$$\inf_{\alpha \in \mathcal{A}_n} n^{2\tau} \mathcal{R}_n[\alpha] \to \infty, \tag{13}$$

$$\sup_{\alpha \in \mathcal{A}_n} \frac{d_n[\alpha]}{n \mathcal{R}_n[\alpha]} \to 0. \tag{14}$$

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

$$\sum_{\alpha \in \mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-2m_1} E_* \big\{ l_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) - E_* l_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) \big\}^{2m_1} \to 0, \tag{15}$$

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

$$\sum_{\alpha \in \mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-2m_2} E_* \bigg[\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}} - tr\big\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1} J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\} \bigg]^{2m_2} \to 0, \tag{16}$$

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

$$\limsup_{n \to \infty} \sum_{\alpha \in \mathcal{A}_n} (n\mathcal{R}_n[\alpha])^{-m_3} \{ E_* \|\boldsymbol{w}_n[\alpha]\|^{m_3} + E_* \|\boldsymbol{w}_n[\alpha]\|^{2m_3} \} < \infty. \tag{17}$$

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 [4, 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 out-sample 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 [4, 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 conditions (13), (14) and (17) indicate that the risks $\mathcal{R}_n[\alpha]$ are large so that the model class is virtually mis-specified. The assumptions in equalities (15) and (16) are central moment constraints that control the regularity of loss functions.

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 sample size, we cannot directly use state-of-art technical tools such as [4, Theorem 19.28]. The classical proof by White [33] (in proving asymptotic normality in mis-specified class) cannot be directly adapted, either, for parameter spaces that depend on n. Some new technical tools are needed in our proof. Our ideas of proof are sketched below.

We first prove that $\hat{\boldsymbol{\theta}}_n[\alpha]$ is n^τ-consistent (instead of the classical $\sqrt{n}$-consistency). We then prove the first key result, namely Lemma 6, that states a type of local uniform convergence. Note that its proof is nontrivial as both the empirical process and $\hat{\boldsymbol{\theta}}_n$ depend on the same observed data. Our technical tools resemble those for proving a Donsker class, but the major difference is that our model dimensions depend on n. We then prove the second key lemma (Lemma 7). It directly leads to the asymptotic normality of maximum likelihood estimators in the classical setting. It is somewhat interesting to see that the proof of Lemma 7 does not require the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_n[\alpha]$ (which usually does not hold in high dimensional settings).

E. Example

Theorem 1 is applicable to general parametric model classes, where assumptions can often be simplified. We shall use regression models as an example of applying Theorem 1. Suppose that the response variable is written as $Y = \mu(X) + \varepsilon$, where ε has mean zero and variance σ^2, and $\mu(\boldsymbol{X})$ is a (possibly nonlinear) function of d_n predictors $\boldsymbol{X} = [X_1, \ldots, X_{d_n}]^{\mathrm{T}}$. In linear models, data analysts assume that μ is a linear function of $\boldsymbol{X}$ in the form of $\mu = \beta_1 X_1 + \cdots + \beta_d X_{d_n}$, where d may or may not depend on the sample size n. We sometimes write $\mu(X)$ as μ for brevity. For simplicity, we assume that σ is known, and X is a random vector independent with ε. Also assume that $E(Y) = 0$ and $E(X_i) = 0$ $(i = 1, \ldots, d_n)$. The observed data are n independent realizations of $Z = (Y, X_1, \ldots, X_{d_n})$. The unknown parameters

are $\theta = (\beta_1, \ldots, \beta_{d_n})$. The model class, denoted by $\mathcal{A}_n$, consists of candidate models represented by $\alpha \subseteq \{1, \ldots, d_n\}$, i.e. $\mu(\boldsymbol{X}) = \sum_{i\in\alpha} \beta_i X_i$.

In regression, it is common to use the quadratic loss function

$$l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) = \left(y - \sum_{j\in\alpha} \beta_j x_j\right)^2 - \sigma^2$$

for $\boldsymbol{\theta} \in \mathcal{H}_n[\alpha]$. Note that the population loss is

$$E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) = E_*\left(\mu - \sum_{j\in\alpha} \beta_j x_j\right)^2. \tag{18}$$

Suppose that $\boldsymbol{\theta}_n^*$ is defined as in (3). We define Σ_{xx} to be the covariance matrix whose (i, j)-th element is $E_*(X_i X_j)$, $\Sigma_{x\mu}$ to be the column vector whose i-th element is $E_*(X_i\mu)$, and $\Sigma_{\mu\mu} = E_*(\mu^2)$. We similarly define $\Sigma_{xx}[\alpha]$, $\Sigma_{x\mu}[\alpha]$, $X[\alpha]$ which are the covariance matrix/vectors restricted to model $\alpha \in \mathcal{A}_n$. Simple calculations show that $\boldsymbol{\theta}_n^*[\alpha] = (\Sigma_{xx}[\alpha])^{-1}\Sigma_{x\mu}[\alpha]$ for $\mathcal{H}_n(\alpha) = \mathbb{R}^{d_n[\alpha]}$, and (18) may be rewritten as

$$\begin{aligned} E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) &= E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}_n^*[\alpha]; \alpha) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*[\alpha]\|^2_{\Sigma_{xx}[\alpha]} \\ &= \left(\Sigma_{\mu\mu} - \Sigma_{\mu x}[\alpha]\Sigma_{xx}[\alpha]^{-1}\Sigma_{x\mu}[\alpha]\right) + \\ &\quad \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*[\alpha]\|^2_{\Sigma_{xx}[\alpha]}. \end{aligned} \tag{19}$$

The decomposition in (19) has a nice interpretation in terms of bias-variance tradeoff. The first term is the $L_2(P_*)$-norm of the orthogonal complement of μ projected to the linear span of covariates, or the minimal possible loss offered by the specified model α. Clearly, it is zero if α is well-specified, and nonzero otherwise. The second term represents the variance of estimation. Using the law of large numbers and Slutsky's theorem, $\hat{\boldsymbol{\theta}}_n[\alpha] = (X[\alpha]^{\mathrm{T}} X[\alpha])^{-1} X[\alpha]^{\mathrm{T}} \boldsymbol{y}$, is a consistent estimator of $\boldsymbol{\theta}_n^*[\alpha]$, where X is the $n \times d_n[\alpha]$ design matrix whose rows are realizations of X_j ($j \in \alpha$), and $\boldsymbol{y} \triangleq [y_1, \ldots, y_n]^{\mathrm{T}}$. It is consistent because of $\hat{\boldsymbol{\theta}}_n[\alpha] = (n^{-1}X[\alpha]^{\mathrm{T}}X[\alpha])^{-1}(n^{-1}X[\alpha]^{\mathrm{T}}\boldsymbol{y})$, Theorem 1 implies the following corollary. It is possible to relax the conditions by more sophisticated verification of assumptions.

Corollary 1: Assume that $|\mu|$ and $|X_i|$ $(i = 1, \ldots, d_n)$ are bounded by a constant c that does not depend on n. Suppose the following conditions hold, then the $\hat{\alpha}_n$ selected by GTIC procedure is asymptotically efficient.

1) $X_1, \ldots, X_{d_n}$ are independent for all n;
2) $d_n = o(n^w)$, where $w < 1/6$;
3) $\inf_{\alpha\in\mathcal{A}_n} \mathcal{R}_n[\alpha] > n^{-\zeta}$, where $\zeta < 1 - 2w$;
4) $\mathrm{card}(\mathcal{A}_n) = o(n^{2(1-\zeta-w)})$.

III. Sequential Model Expansion

As explained in the introduction, in terms of predictive power, a model in a mis-specified model class could be determined to be unnecessarily large, suitable, or inadequately small, depending on specific sample size (see Fig. 2). A realistic learning procedure thus requires models of different complexity levels as more data become available.

Throughout this section, we shall use T (instead of the previously used n) to denote sample size, and subscript t as the data index, in order to emphasize the sequential setting.

A. Discussion

We have addressed the selection of an efficient model for a given number of observations. In many practical situations, data are sequentially observed. A straightforward model selection is to repeatedly apply GTIC procedure upon arrival of data. However, in a sequential setting, the following issue naturally arises:

Suppose that we successively select a model and use it to predict at each time step. The path of the historically selected models may fluctuate a lot (which will be illustrated in our numerical experiments). Instead, it is more appealing (either philosophically or computationally) to force the selected models to evolve gradually.

To address the above challenge, we first propose a concept referred to as the *graph-based* expert tracking, which extends some classical online learning techniques (Algorithm 1). Motivated by the particular path graph $1 \to 2 \to \cdots N$, where $1, 2, \ldots, N$ index the model class, we further propose a model expansion strategy (Algorithm 2), where each candidate model and its corrected prediction loss can be regarded respectively as an expert and loss at each time.

The proposed algorithm can be used for online prediction, which ensures not only statistically reliable results but also simple computation. Specifically, we propose a predictor that has cumulative out-sample prediction loss (over time) close to the following optimum benchmark:

$$\min_{\mathrm{size}(i_1,\ldots,i_T)\le k,\ i_1,\ldots,i_T\in\{1,\ldots,N\}} \sum_{t=1}^{T} \mathcal{L}_n[\alpha_{i_t}]. \tag{20}$$

where the size of a sequence $\mathrm{size}(i_1, \ldots, i_T)$ is defined as the number of t's such that $i_t \neq i_{t+1}$. In other words, the minimization is taken over all tuples $(i_1, \ldots, i_T)$ that have at most k switches and that are restricted to the chain $1 \to 2 \to \cdots$. For example, $(i_1, \ldots, i_5) = (1, 2, 2, 3, 3)$. In the above formulation, i_t and k respectively means the index of model chosen to predict at time step t, and the number of switches within T time steps.

B. Tracking the Best Expert with Graphical Constraints

In this subsection, we propose a novel graph-based expert tracking technique that motivates our algorithm in the following subsection. The discussion may be interesting on its own right, as it includes the state-of-art expert tracking framework as a special case (when the underlying graph is complete).

Suppose there are N experts. At each discrete time step $t = 1, 2, \ldots, T$, each expert gives its prediction, after which the environment reveals the truth $\boldsymbol{z}_t \in \mathcal{Y}$. In this subsection, with a slight abuse of notation, we shall also use l to denote loss functions in the context of online learning. The performance of each prediction is measured by a loss function $l : \{1, 2, \ldots, N\} \times \mathcal{Y} \to \mathbb{R}$. The smaller the loss is, the better the prediction is. In light of the model expansion we shall introduce in the next subsection, each $i = 1, \ldots, N$ represents a model, and $l(i, \boldsymbol{z}_t)$ is the prediction loss of model i which is successively re-estimated using $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t$ at time step t.

In order to aggregate all the predictions that the experts make, we maintain a weight for each expert, and update

Algorithm 1 Tracking the best expert with graphical transitional constraints

input Learning rate $\eta > 0$, sharing rate $0 < \kappa < 1/D$

output $\boldsymbol{p}_t = [p_{t,1}, \ldots, p_{t,N}]^{\mathrm{T}}$ (predictive distribution over the active models) for each $t = 1, \ldots, T$

1: Initialize $w_{1,0} = 1$ $w_{i,0} = 0$ for all $i \in \{2, \ldots, N\}$
2: **for** $t = 1 \to T$ **do**
3: Calculate the predictive distribution $p_{i,t} = w_{i,t-1} / \sum_{j=1}^{N} w_{j,t-1}$, for each $i \in \{1, \ldots, N\}$
4: Read $\boldsymbol{z}_t$, and compute $v_{i,t} = w_{i,t-1} \exp(-\eta \cdot l(i, \boldsymbol{z}_t))$, for each $i \in \{1, \ldots, N\}$
5: Let $w_{i,t} = \kappa \sum_{j=1}^{N} \beta_{ji} v_{j,t} + (1 - \kappa\beta_i) v_{i,t}$ for each $i \in \{1, \ldots, N\}$, where β_{ji}, β_i are defined in (22), (22)
6: **end for**

them upon the arrival of each new data point based on the qualities of the predictions. We denote the weight for expert $i \in \{1, \ldots, N\}$ at time t as $w_{i,t}$, and the normalized version as $W_{i,t}$. The goal is to optimally update the weights for better prediction, which is measured by the cumulative loss minus the best achievable (benchmark) loss. This measure is often called "regret" in the online learning literature.

If the benchmark in the regret is defined as the minimum cumulative loss achieved by a single expert, then the exponential re-weighting algorithm can be applied easily. In many cases the best performing expert can be different from one time segment to another, for which the fixed share algorithm [34, Chapter 5] can be a good solution with guaranteed regret bound. We consider the following problem setting that aims to reduce computational cost. The best performing expert is restricted to switch according to a *directed graph*, $G = (V, E)$ (without self-loops), with $V = \{1, \ldots, N\}$ denoting the set of nodes (representing experts) and E denoting the set of directed edges. At each time point, the best performing expert can either stay the same or jump to another node which is directly connected from the current node. Let

$$\beta_{ij} = 1_{\exists (i,j) \in E}, \tag{21}$$

which is 1 if there is a directed edge (i, j) on the graph, and 0 otherwise. Let

$$\beta_i = \sum_{j=1}^{N} \beta_{ij}, \tag{22}$$

which is the out-degree of the node i. In addition, we assume that $\max_{i \in 1, \ldots, N} \beta_i \le D$, where $0 < D < N$.

We propose Algorithm 1 to follow the best expert with the graphical transitional constraints. We use a special prior $w_{i,0}$ here to motivate content in the next subsection. It is not difficult to extend our discussion to more general priors here. The classical fixed-share algorithm can be seen as a special case when the graph is complete. The advantage of using the graph-based expert learning is to reduce the computational cost and to obtain a tighter error bound, as shown in the following Theorem 2.

Theorem 2: Suppose the loss function takes values from $[0, 1]$. For all $T \ge 1$, the output of the algorithm in Algorithm 1 satisfies

$$\sum_{t=1}^{T} \left(\sum_{i=1}^{N} l(i, \boldsymbol{z}_t) p_{i,t} - l(i_t, \boldsymbol{z}_t) \right) \le \frac{1}{\eta}(T - k - 1) \log \frac{1}{1 - \kappa D} + \frac{1}{\eta} k \log \frac{1}{\kappa} + \eta \frac{T}{8}$$

for all expert sequence $(i_1, i_2, \ldots, i_T)$ and all observation sequence $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t)$, given that $(i_1, i_2, \ldots, i_T)$ has only transitions following directed paths in graph G and size$(i_1, i_2, \ldots, i_T) \le k$. Here, the left hand side of the above inequality is referred to as *regret*.

In particular, the above regret bound has minimum $\sqrt{TS/2}$ which is achieved with the learning rate $\eta = \sqrt{8S/T}$ and sharing rate $\kappa = k/\{(T-1)D\}$, where

$$S = (T - 1) H(k/(T - 1)) + k \log D,$$

and $H(\cdot)$ is the binary entropy function defined by $H(x) \triangleq -x \log x - (1-x) \log(1-x)$ for $x \in (0, 1)$, $H(0) = H(1) = 0$.

It is interesting to see that with graphical constraint, the regret bound does not depend on N, but on the maximum out-degree D instead. Thus, the bound can be tight even when N grows exponentially in T, as long as $D \ll N$ (i.e. sparse graph).

C. Algorithm for Sequential Model Expansion

The new online learning theory proposed in the last subsection is motivated by graph-based expert tracking. Intuitively speaking, instead of using the exponentially updated weights directly, each expert borrows some weights from others, allowing poorly performing experts to quickly stand out when they start doing better. In that way, the experts are encouraged to rejuvenate their past performance and "start a new life", so that we can track the best expert in different time epochs. The classical fixed-share algorithm [34, Chapter 5] is a special case when $\beta_{ij} = 1$ for all $i \ne j$ and κ becomes $\kappa/(N-1)$, illustrated in Fig. 3(a). Our algorithm in this subsection is motivated by the particular *path graph* $1 \to 2 \to \cdots N$, where $1, 2, \ldots, N$ index the model class. In other words, we share the weights in a directional way, thus encouraging the experts to switch in a chain. The update rule is illustrated by Fig. 3(b).

Our algorithm for sequential model expansion is summarized in Algorithm 2, where each candidate model and its corrected prediction loss can be regarded respectively as an expert and loss at each time. The labeling of models $\alpha_1, \alpha_2, \ldots$ is generally in the ascending order of their dimensions. To further reduce the computational cost, we maintain only an active subset (of size K) instead of all the candidate models at each time. The active subset starts from $\{\alpha_1, \ldots, \alpha_K\}$; it switches to $\{\alpha_2, \ldots, \alpha_{K+1}\}$ when the weight of the smallest model α_1 becomes small and that of the largest model α_K becomes large; it continues to switch upon the aggregation of data.

The output of Algorithm 1 is a predictive distribution over the active models. It can be used in the following two ways in practice: 1) we randomly draw a model according to the predictive distribution and use the predictor of that model, or 2) we use the weighted average of predictors of each model

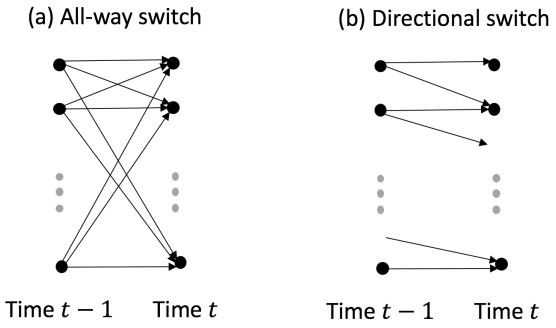


Fig. 3: Illustration of the state-of-art and our new way of redistributing the share of weights in online learning.

Algorithm 2 Sequential model expansion using GTIC-corrected loss (GTIC-sequential)

input $\{\boldsymbol{z}_t : t = 1,\dots,T\}$, $\eta > 0$, $\kappa \in [0,1]$, $w_{0,1} = 1, w_{0,2} = \cdots = w_{0,K} = 0$, candidate models $\mathcal{A}_T = \{\alpha_1, \alpha_2, \dots, \alpha_{\text{card}(\mathcal{A}_T)}\}$, $s = 0$ ($\alpha_{s+1}, \dots, \alpha_{s+K}$ are the maintained active subsets of models), $K \in \mathbb{N}$, threshold $\rho \in [0,1]$

output $\boldsymbol{p}_t = [p_{t,1}, \dots, p_{t,K}]^{\mathrm{T}}$ (predictive distribution over the active models) for each $t = 1, \dots, T$

1: **for** $t = 1 \to n$ **do**
2: Obtain $\boldsymbol{z}_t$ and compute $v_{t,k} = w_{t-1,k} \exp\{-\eta\, \mathcal{L}_t^c[\alpha_{s+k}]\}$ for each $k = 1, \dots, K$, where $\mathcal{L}_t^c[\alpha]$ is calculated from (6) and fitting the data $\boldsymbol{z}_1, \dots \boldsymbol{z}_t$ to model α.
3: Let

$$w_{t,k} = \begin{cases} (1-\kappa)\, v_{t,k} & \text{if } k = 1 \\ (1-\kappa)\, v_{t,k} + \kappa\, v_{t,k-1} & \text{if } 1 < k < K \\ v_{t,k} + \kappa\, v_{t,k-1} & \text{if } k = K \end{cases}$$

4: Let $p_{t,k} = (\sum_{k=1}^K w_{t,k})^{-1} w_{t,k}$, $k = 1, \dots, K$
5: **if** $p_{t,1} \le \rho$ and $p_{t,K} \ge 1 - \rho$ and $s + K \le \text{card}(\mathcal{A}_T)$ **then**
6: Let $s = s + 1$
7: Let $w_{t,k} = w_{t,k'}$, where $k = 1, \dots, K$ and $k' = (k + 1 \bmod K)$ (relabeling the active models)
8: **end if**
9: **end for**

according to the predictive distribution. This can be regarded as a specific ensemble learning (or model averaging) method. The following Proposition 1 shows that with appropriate learning parameters, the average predictive performance of our algorithm is asymptotically close to the average of a series of truly optimal models (i.e. optimal model expansion) allowing moderately many switches.

Proposition 1: Suppose that Assumptions 1-7 hold, and that $\sup_{1\le t\le T} \sup_{\alpha\in\mathcal{A}_n} \mathcal{L}_t^c[\alpha] < c$ almost surely for some fixed constant $c > 0$. Suppose that the lines 5-8 are removed from Algorithm 2, and that $K = \text{card}(\mathcal{A}_n)$, then its output satisfies

$$\frac{1}{T}\Bigg(\sum_{t=1}^{T}\sum_{i=1}^{\text{card}(\mathcal{A}_T)} p_{i,t}\, \mathcal{L}_t^c[\alpha_i] - \min_{\text{size}(i_1,i_2,\dots,i_T)\le k} \sum_{t=1}^{T} \mathcal{L}_t^c[\alpha_{i_t}]\Bigg) \le \frac{c}{\sqrt{2}}\sqrt{H\Big(\frac{k}{T-1}\Big)} \tag{23}$$

for all $T \ge 1$, given that

$$\kappa = \frac{k}{T-1},\ \eta = \frac{1}{c}\sqrt{8\frac{T-1}{T}H\Big(\frac{k}{T-1}\Big)}.$$

In particular, if $k = o(T)$, we have

$$\limsup_{T\to\infty} \frac{1}{T}\Bigg(\sum_{t=1}^{T}\sum_{i=1}^{\text{card}(\mathcal{A}_T)} p_{i,t}\, \mathcal{L}_t^c[\alpha_i] - \min_{\text{size}(i_1,i_2,\dots,i_T)\le k} \sum_{t=1}^{T} \mathcal{L}_n[\alpha_{i_t}]\Bigg) \le 0 \tag{24}$$

almost surely.

Remark 4 (Explanation of Algorithm 2 and Proposition 1): In addition to the (sequential) data and model class, other inputs to Algorithm 2 are two learning parameters η, κ, the number of active models K, and the threshold ρ. The parameters η and κ control the rate of learning and the rate of model expansion, respectively. The number of active models K is set to reduce the computation cost when sample size is small compared with model dimensions, and the threshold ρ is used to update our active models under consideration.

In particular, upon the arrival of a new data point or a set of data points, denoted by $\boldsymbol{z}_t$, at each time step t (line 1), we update the weight of each candidate model by a Bayes-type procedure (line 2). The loss employed in the update is the corrected prediction loss, which is directly computable from the data and which serves as an approximation of the out-sample prediction loss (as was discussed in Subsection II-B). The weights of each model are then updated following the path graph (line 3). When the weight of the smallest model becomes small and that of the largest model becomes large, it means the current active models are inadequately small. So we drop the smallest model and include the next large model into the active set, and adjust their weights accordingly (lines 5-8). In line 7, the weight of the removed model is assigned to the newly included one, so that the sum of the weights remains the same.

Proposition 1 states that the average predictive performance of our algorithm is asymptotically close to that of the optimal model expansion allowing $k = o(T)$ switches. For example, if only one point arrives at each time step, and the dimension of the optimal model is at the order of T^δ for $\delta \in (0,1)$, then the condition is trivially satisfied. The proof of Proposition 1 follows directly from Theorem 1 and Theorem 2, by using simple manipulations. For technical convenience, Proposition 1 is only proved by removing the part of maintaining an active subset (lines 5-8). We maintain an active subset only for computational purpose, and we notice that it does not deteriorate the performance from various numerical experiments. This is because the weights of the removed models are negligible.

IV. Numerical Experiments

The model classes under consideration are logistic regression and feed-forward neural networks. We also released an open source 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 provide their tensor variables of loss 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. 4 and 5.

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). In Fig. 5a, 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. However, both GTIC and 10-fold CV fluctuate a little bit. Our proposed sequential model expansion algorithm smoothly expands the model and yields the best performance compared to all the other approaches. As shown in Fig. 4a and 4b, although the optimal model of each sample size is not always identical to the selected model from our model expansion algorithm, the loss of our selected model is almost the same as the optimal model. This result is consistent with our definition of efficient learning.

The computation cost of all approaches is provided in Fig. 5b. 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.

In another experiment, we considered two underlying data generating models. One model (called $\mathcal{M}_1$) is generated using a logistic regression model with coefficients $\boldsymbol{\beta} = [1, 2^{0.1}, \ldots, p^{0.1}]$ and standard Gaussian covariates. The other model (called $\mathcal{M}_2$) is generated using a logistic regression model with coefficients $\boldsymbol{\beta} = [0.999, 0.999^2, \ldots, 0.999^p]$ and standard Gaussian covariates. We numerically compare the performance of AIC, BIC, and TIC under various choices of sample size n and number of covariates p. The performance is evaluated using out-sample prediction loss, prediction accuracy, and prediction efficiency, summarized in Tables II, III, IV, respectively. The best performing method is highlighted with bold. The results show that TIC performs the best in all cases, and BIC is always the worst for the mis-specified model class.

B. Neural Networks

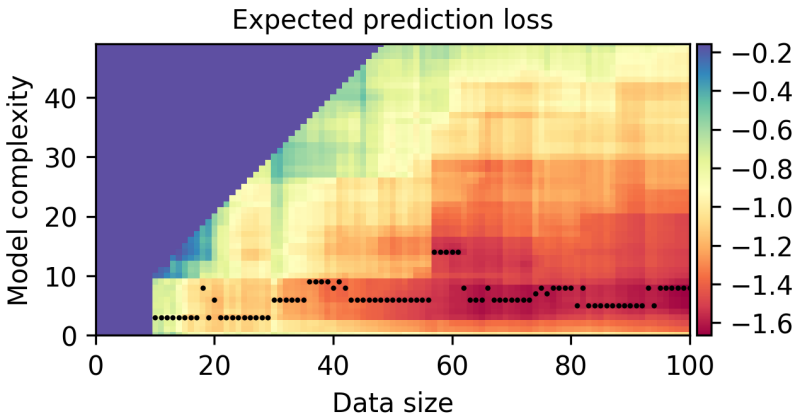
We consider the model class to be single-layer feed-forward neural networks (see Fig. 6a). Neural networks are inherently miss-specified models.

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 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 two. In this experiment, the model complexity of our model is the number of hidden nodes in the single hidden layer.

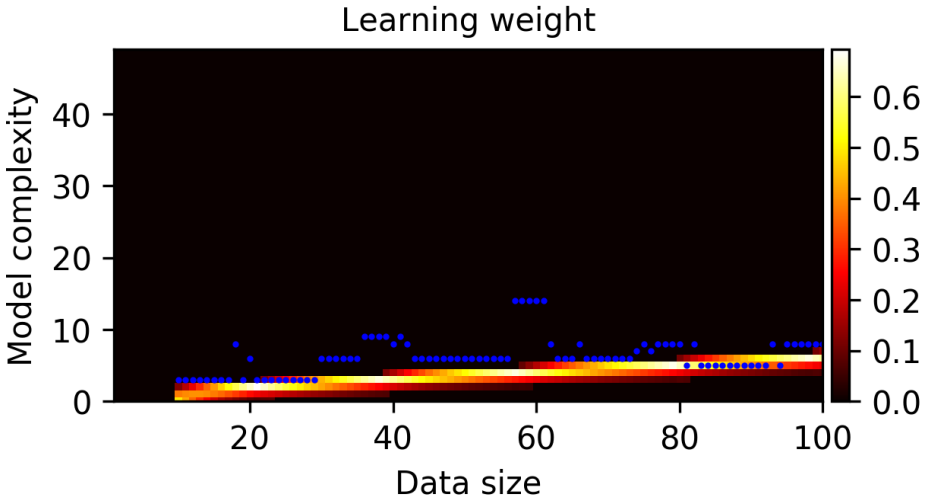
We sequentially obtain and learn the data, starting from $t = 100$, then $t = 101, \ldots, 300$. We start from 100 samples because Neural Network is likely to converge to a local optimal for small sample size. The path of expansion in this case is the number of hidden nodes in the single hidden layer. Since data are not linearly-separable, we do need at least one hidden layer 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 sample size is shown in Fig. 7a. 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. In Fig. 8a, 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. Our proposed sequential algorithm is superior to other approaches as shown in Fig. 7b, because the weight of smaller models in the active set is large enough to prevent the model to expand. As a result, we alleviate the tendency to choose the overfitting models even when their loss is relatively small.

The computational cost is shown in Fig. 8b. 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



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



(b) Heat-map showing our predictive weights over the candidate models (y-axis) at each sample size (x-axis), using sequential model expansion.

Fig. 4: Experiment 1: logistic regression models

our implementation, which is a benefit that CV cannot enjoy in naive sequential model selection framework. Therefore, we encourage the use of GTIC in sequential model expansion scheme.

V. Conclusions

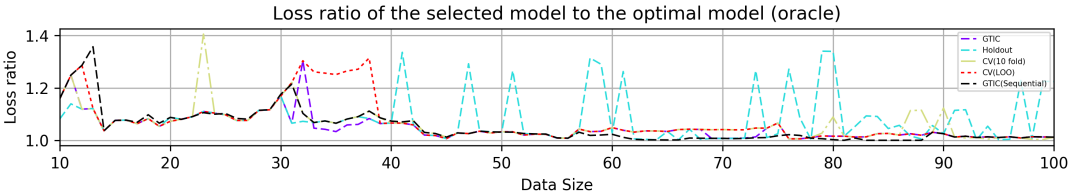
We addressed the challenges of high dimensions and/or small samples by developing a framework for model expansion. This framework aims to approach the best predictive power in sequential settings. In the first part of this paper, we explained that requiring larger model to fit more data is not only theoretically appealing but also practically useful. We then studied a method (GTIC) to approach the limit of statistical 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 sample size). The theoretical analysis of GTIC justifies the use of TIC for general mis-specified model classes, and extends some technical tools for classical analysis on AIC in linear models. In the second part of the paper, we proposed a sequential model expansion strategy for reliable online prediction with low computation cost. It is motivated by our graph-based expert tracking techniques. In summary, the proposed methodology is asymptotically optimal and practically useful, and may be a promising competitor of various types of cross-validations.

Acknowledgement

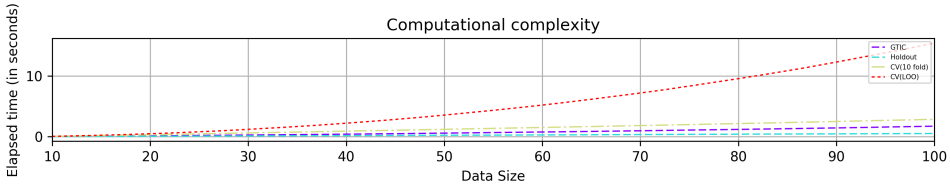
This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers W911NF1810134 and W911NF1610561. The authors thank Bingfeng Xia and Kathryn Heal for helpful suggestions.

Appendix A
Proof of Theorem 1

We start with the following technical lemmas and additional definitions.

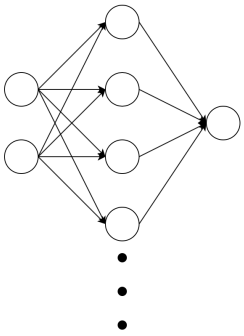


(a) Plot showing the loss of our predictor (GTIC) and cross validations at each sample size

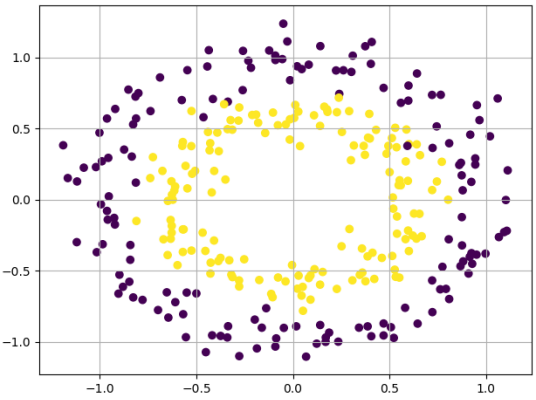


(b) Plot showing the computational costs.

Fig. 5: Experiment 1: logistic regression models



(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 $= 0.6$)

Fig. 6: Experiment 2: neural networks

Lemma 1: Suppose that Assumptions 1, 2, 3, 4, 6 hold. Then $\hat{\boldsymbol{\theta}}_n$ is n^τ-consistent uniformly over $\mathcal{A}_n$, namely $\sup_{\alpha\in\mathcal{A}_n} n^\tau \|\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\| = O_p(1)$.

Proof: Using Assumption 2 and a direct adaptation of the techniques in [4, Theorem 5.7], we can prove that $\hat{\boldsymbol{\theta}}_n[\alpha]$ is consistent in the sense that

$$\sup_{\alpha\in\mathcal{A}_n} \|\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\| = o_p(1) \tag{25}$$

as $n \to \infty$.

From the definitions of $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n^*$, we have for each $\alpha \in \mathcal{A}_n$

$$\begin{aligned} &n^\tau E_*\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) - \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= n^\tau\{0 - E_*\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= n^\tau\{E_n\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha) - E_*\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \end{aligned} \tag{26}$$

From the differentiability of the map $\boldsymbol{\theta} \mapsto E_*\psi_n(\cdot, \boldsymbol{\theta}; \alpha)$, there exists $\tilde{\boldsymbol{\theta}}[\alpha]$ such that $\|\tilde{\boldsymbol{\theta}}[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\| \le \|\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\|$, and

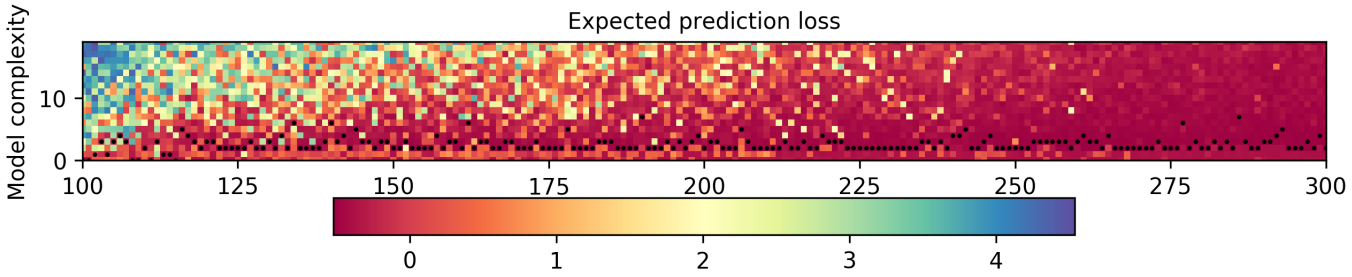
$$\begin{aligned} &E_*\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) - \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= \nabla_{\boldsymbol{\theta}} E_*\{\boldsymbol{\psi}_n(\cdot, \tilde{\boldsymbol{\theta}}[\alpha]; \alpha)\}(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]) \\ &= V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]), \end{aligned} \tag{27}$$

where the exchangeability of integral and differentiation (in the second identity) is guaranteed by (7) and (8) in Assumption 5.

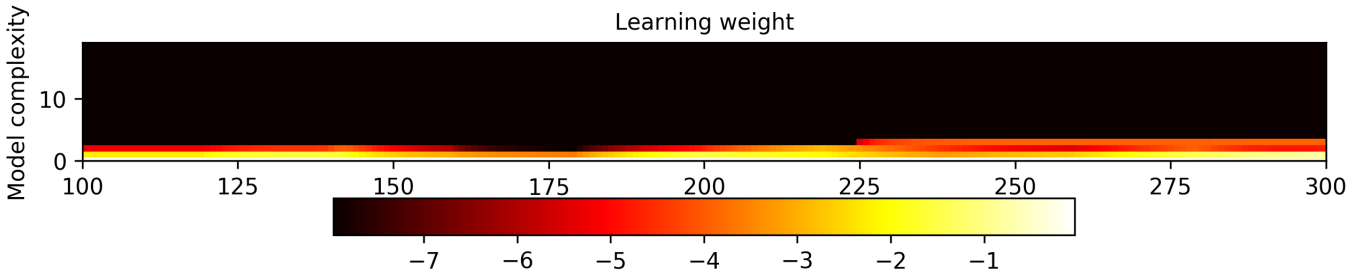
Therefore, with probability tending to one, we have

$$\begin{aligned} &\sup_{\alpha\in\mathcal{A}_n} n^\tau \|V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha])\| \\ &= \sup_{\alpha\in\mathcal{A}_n} n^\tau \|E_*\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha) - \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n; \alpha)\}\| \\ &= \sup_{\alpha\in\mathcal{A}_n} n^\tau \|E_n\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha) - E_*\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\| \\ &= O_p(1) \end{aligned}$$

where the first equality is due to (27), the second equality is due to (26), and the third equality comes from Assumption 3. By the (12) in Assumption 6 and Assumption 4, $V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)$

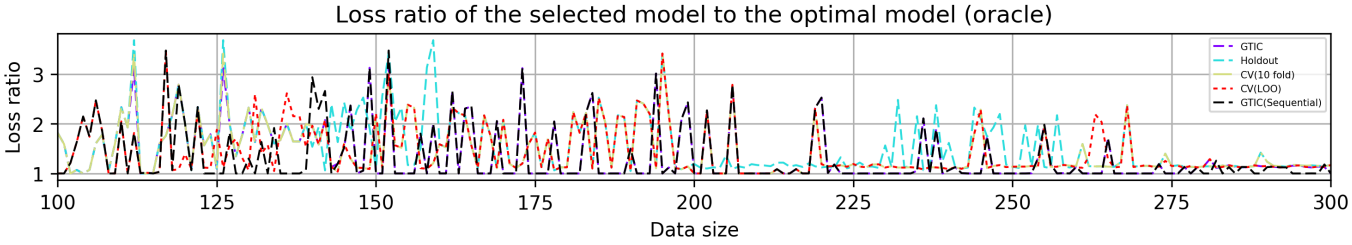


(a) Heat-map showing the prediction loss of estimated candidate models of each dimension (y-axis) at each sample size (x-axis), where the black dots indicate the model of optimal loss at each sample size.

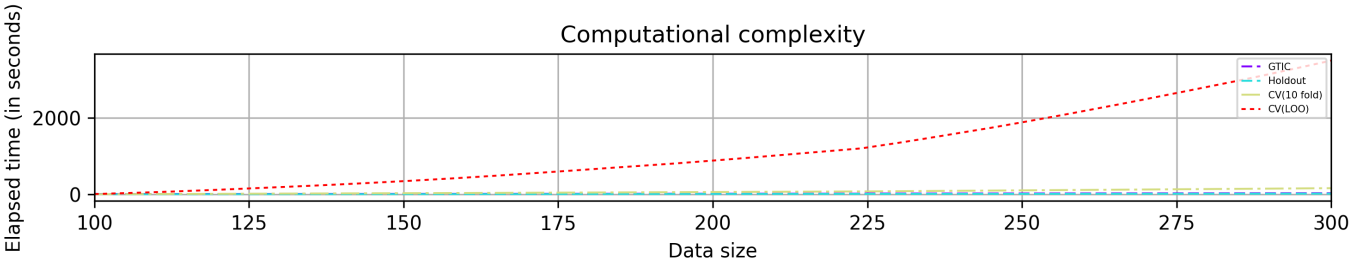


(b) Heat-map showing our predictive weights over the candidate models (y-axis) at each sample size (x-axis).

Fig. 7: Experiment 2: neural networks



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



(b) Plot showing the computational costs.

Fig. 8: Experiment 2: neural networks

is invertible for each $\alpha \in \mathcal{A}_n$, and

$$\sup_{\alpha \in \mathcal{A}_n} \|V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)^{-1}\| < 1/(2c_1)$$

with probability tending to one. It follows that

$$\begin{aligned} &\sup_{\alpha \in \mathcal{A}_n} n^{\tau} \|\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]\| \\ &\leq \sup_{\alpha \in \mathcal{A}_n} \Big\{ \|V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)^{-1}\| \times \\ &\qquad \|n^{\tau} V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha])\| \Big\} \\ &= O_p(1), \end{aligned} \tag{28}$$

which concludes the proof. ■

Before we proceed, we need the following definition.

TABLE II: Out-sample prediction loss of each method (means and standard errors)

	$\mathcal{M}_1$			$\mathcal{M}_2$		
	AIC	BIC	TIC	AIC	BIC	TIC
$n=100, p=20$	**0.42 (0.01)**	0.53 (0.03)	**0.42 (0.01)**	0.39 (0.01)	0.52 (0.04)	**0.38 (0.01)**
$n=100, p=50$	0.65 (0.02)	0.69 (0.01)	**0.50 (0.02)**	0.68 (0.03)	0.71 (0.01)	**0.47 (0.02)**
$n=300, p=60$	**0.34 (0.01)**	0.64 (0.02)	**0.34 (0.01)**	**0.30 (0.01)**	0.44 (0.05)	**0.30 (0.01)**
$n=300, p=150$	0.81 (0.05)	0.69 (0.00)	**0.53 (0.02)**	0.81 (0.03)	0.70 (0.00)	**0.50 (0.01)**
$n=500, p=100$	**0.34 (0.01)**	0.68 (0.01)	**0.34 (0.01)**	0.30 (0.01)	0.44 (0.05)	**0.29 (0.01)**
$n=500, p=250$	0.95 (0.04)	0.69 (0.00)	**0.54 (0.02)**	0.96 (0.03)	0.69 (0.00)	**0.51 (0.01)**

TABLE III: Out-sample prediction accuracy of each method (means and standard errors)

	$\mathcal{M}_1$			$\mathcal{M}_2$		
	AIC	BIC	TIC	AIC	BIC	TIC
$n=100, p=20$	**0.79 (0.01)**	0.74 (0.02)	**0.79 (0.01)**	**0.83 (0.02)**	0.75 (0.05)	**0.83 (0.02)**
$n=100, p=50$	0.69 (0.03)	0.54 (0.07)	**0.78 (0.02)**	0.73 (0.03)	0.63 (0.06)	**0.80 (0.02)**
$n=300, p=60$	**0.83 (0.01)**	0.60 (0.03)	**0.83 (0.01)**	**0.87 (0.01)**	0.80 (0.03)	**0.87 (0.01)**
$n=300, p=150$	0.72 (0.02)	0.55 (0.09)	**0.77 (0.01)**	0.76 (0.02)	0.67 (0.08)	**0.81 (0.02)**
$n=500, p=100$	**0.83 (0.01)**	0.58 (0.04)	**0.83 (0.01)**	**0.87 (0.01)**	0.77 (0.06)	**0.87 (0.01)**
$n=500, p=250$	0.74 (0.01)	0.50 (0.08)	**0.79 (0.01)**	0.73 (0.01)	0.40 (0.05)	**0.79 (0.01)**

TABLE IV: Out-sample prediction efficiency of each method (means and standard errors)

	$\mathcal{M}_1$			$\mathcal{M}_2$		
	AIC	BIC	TIC	AIC	BIC	TIC
$n=100, p=20$	0.95 (0.02)	0.55 (0.09)	**0.97 (0.01)**	0.96 (0.02)	0.56 (0.09)	**0.97 (0.02)**
$n=100, p=50$	0.64 (0.04)	0.57 (0.04)	**0.97 (0.02)**	0.59 (0.04)	0.53 (0.03)	**0.98 (0.01)**
$n=300, p=60$	**0.98 (0.01)**	0.32 (0.06)	**0.98 (0.01)**	0.96 (0.02)	0.64 (0.10)	**0.98 (0.01)**
$n=300, p=150$	0.58 (0.03)	0.70 (0.04)	**0.99 (0.01)**	0.56 (0.02)	0.66 (0.03)	**1.00 (0.00)**
$n=500, p=100$	**0.98 (0.01)**	0.32 (0.02)	**0.98 (0.01)**	0.96 (0.01)	0.62 (0.08)	**1.00 (0.00)**
$n=500, p=250$	0.50 (0.02)	0.73 (0.04)	**0.99 (0.01)**	0.48 (0.02)	0.68 (0.02)	**0.99 (0.01)**

Definition 4 (Bracketing number): Given two scalar functions f_1 and f_2, the bracket $[f_1, f_2]$ is the set of all functions f such that $f_1 \le f \le f_2$. An ε-bracket in $L_2(P_*)$ is a bracket $[f_1, f_2]$ with $E_*(f_2 - f_1)^2 < \varepsilon^2$. The bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P_*))$ is the minimum number of ε-brackets needed to cover a set $\mathcal{F}$. Moreover, the bracketing integral is defined by

$$I_{[\,]}(\delta, \mathcal{F}, L_2(P_*)) = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P_*))} d\varepsilon \quad (29)$$

for $\delta > 0$.

Remark 5: The logarithm of the above bracketing number is also referred to as bracketing entropy relative to the $L_2(P_*)$-norm. It is commonly used to describe the size of a class of functions. We will use the above definition in order to prove uniform convergence results. We refer to [35] for a different bracketing idea which was used to study the nonasymptotic estimation theory.

We have the following lemma whose proof follows directly from Definition 4 and Assumption 5.

Lemma 2: Suppose that Assumption 5 holds, and $r_n \le r$ for all n (where r has been defined in Assumption 5). Let $\mathcal{F}_n[\alpha] = \left\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha) : \boldsymbol{\theta} \in B(\boldsymbol{\theta}_n^*[\alpha], r_n) \subset \mathbb{R}^{d_n[\alpha]}\right\}$ be a collection of (vector-valued) measurable functions. Then

$$N_{[\,]}(\varepsilon, \mathcal{F}_n[\alpha], L_2(P_*)) \le \left(\varepsilon^{-1} r_n \|m_n\|_{P_*}\right)^{d_n[\alpha]}$$

for all $0 < \varepsilon < r_n$.

We prove the following technical lemmas.

Lemma 3: For any sets of functions $\mathcal{F}_j, j = 1, \ldots, k$, we have

$$\begin{aligned} &I_{[\,]}(\delta, \cup_{1\le j\le k}\mathcal{F}_j, L_2(P_*)) \\ &\le \sqrt{2\log k}\ \delta + \sqrt{k} \sup_{1\le j\le k} I_{[\,]}(\delta, \mathcal{F}_j, L_2(P_*)) \end{aligned}$$

Proof: The case $k = 1$ is straightforward. We only need to prove for $k \ge 2$. By Definition 4, we have

$$N_{[\,]}(\varepsilon, \cup_{1\le j\le k}\mathcal{F}_j, L_2(P_*)) \le \sum_{1\le j\le k} N_{[\,]}(\varepsilon, \mathcal{F}_j, L_2(P_*).$$

From (29), it suffices to prove the following result, and then let v_j's be $N_{[\,]}(\varepsilon, \mathcal{F}_j, L_2(P_*))$'s. For any numbers $v_1 \ge \cdots \ge$

$v_k \geq 1$ ($k \geq 2$), we have

$$\sqrt{\log \sum_{j=1}^{k} v_j} \leq \max\{\sqrt{2\log k},\ \sqrt{k \log v_1}\} \leq \sqrt{2\log k} + \sqrt{k \log v_1}.$$

Furthermore, it suffices to prove that

$$\log(kv_1) \leq \max\{2\log k,\ k\log v_1\}. \tag{30}$$

In fact, if $v_1 \leq k^{1/(k-1)}$, then

$$\log(kv_1) \leq \log(k \cdot k^{1/(k-1)}) = \frac{k}{k-1}\log k \leq 2\log k.$$

Otherwise, $\log(kv_1) \leq k \log v_1$, because $g : v \mapsto k\log v - \log(kv)$ is increasing on $v \geq 1$ and it equals zero when $v = k^{1/(k-1)}$. ∎

Definition 5: For any class $\mathcal{F}$ of functions $f : \mathcal{Z} \to \mathbb{R}$, a function $F : \mathcal{Z} \to \mathbb{R}$ is called an envelope function of $\mathcal{F}$, if $\sup_{f\in\mathcal{F}} |f(z)| \leq F(z) < \infty$ for every $z \in \mathcal{Z}$.

Lemma 4: ([4, Lemma 19.34]) For any class $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ such that $E_* f^2 < \delta^2$ for all f, with

$$a(\delta) = \delta / \sqrt{\max\{1, \log N_{[\,]}(\delta, \mathcal{F}, L_2(P_*))\}}$$

and F an envelope function, that

$$E_* \sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \lesssim I_{[\,]}(\delta, \mathcal{F}, L_2(P_*)) + \sqrt{n} E_*\big\{F \cdot 1_{F > \sqrt{n}a(\delta)}\big\}.$$

Here, 1_A is the indicator function of event A.

Lemma 5: Let $\mathcal{F}_n = \cup_{\alpha\in\mathcal{A}_n} \mathcal{F}_n[\alpha]$ where $\mathcal{F}_n[\alpha] = \{\boldsymbol{f}_{n,u} : u \in U[\alpha]\}$ be a class of measurable vector-valued functions. In other words, for each $\alpha \in \mathcal{A}_n$ and $u \in U[\alpha]$, $\boldsymbol{f}_{n,u} = [f_{n,u,1}, \ldots, f_{n,u,d_n[\alpha]}]^{\mathrm{T}}$ with $f_{n,u,i} : \mathcal{Z} \to \mathbb{R}$ being a scalar-valued function. The dimension $d_n[\alpha]$ can be different for $\alpha \in \mathcal{A}_n$, and we let $d_n = \max_{\alpha\in\mathcal{A}_n} d_n[\alpha]$. Assume that the following conditions hold.

(i) There is an envelope function F_n that satisfies

$$\sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]\subset\mathbb{R}^{d_n[\alpha]}, 1\leq i\leq d_n[\alpha]} |f_{n,u,i}(\boldsymbol{z})| \leq F_n(\boldsymbol{z}) < \infty$$

for every $\boldsymbol{z} \in \mathcal{Z}$;

(ii) There exists a deterministic sequence $\{\delta_n\}$ such that

$$d_n \sqrt{\log\{d_n \mathrm{card}(\mathcal{A}_n)\}}\delta_n \to 0, \tag{31}$$

and

(iii) The bounded moment condition:

$$\delta_n^{-2} E_* F_n^2 \to 0;$$

(iv) The bounded class condition:

$$\sqrt{d_n^{3/2}\ \mathrm{card}(\mathcal{A}_n)} \times \sup_{\alpha\in\mathcal{A}_n, 1\leq i\leq d_n[\alpha]} I_{[\,]}(\delta_n, \mathcal{F}_{n,i}[\alpha], L_2(P_*)) \to 0,$$

where we let $\mathcal{F}_{n,i}[\alpha] = \{f_{n,u,i} : u \in U[\alpha]\}$.

Then we have $\sup_{\boldsymbol{f}\in\mathcal{F}_n} \|\mathbb{G}_n \boldsymbol{f}\| \to_p 0$ as $n \to \infty$.

Proof: By Markov's inequality, it suffices to prove that $E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]} \|\mathbb{G}_n \boldsymbol{f}_{n,u}\| \to 0$ as $n \to 0$.

Condition (iii) implies that for all sufficiently large n,

$$\sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], i=1,\ldots,d_n} E_* f_{n,u,i}^2 \leq E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], i=1,\ldots,d_n} f_{n,u,i}^2 < \delta_n^2. \tag{32}$$

Let $\delta_n, a_n(\delta_n)$ be the constants given in Lemma 4 corresponding to $\delta = \delta_n$ and

$$\tilde{\mathcal{F}}_n = \bigcup_{\alpha\in\mathcal{A}_n, 1\leq i\leq d_n[\alpha]} \mathcal{F}_{n,i}[\alpha].$$

From inequality (32) and Lemma 4, we have

$$\begin{aligned} &E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], 1\leq i\leq d_n[\alpha]} |\mathbb{G}_n f_{n,u,i}| \\ &\lesssim I_{[\,]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*)) + \sqrt{n} E_*\big\{F_n \cdot 1_{F_n > \sqrt{n}a_n(\delta_n)}\big\} \\ &\leq I_{[\,]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*)) + \frac{1}{a_n(\delta_n)} E_* F_n^2, \end{aligned} \tag{33}$$

where the second inequality comes from the fact that

$$1_{F_n > \sqrt{n}a_n(\delta_n)} \leq \frac{F_n}{\sqrt{n}a_n(\delta_n)} 1_{F_n > \sqrt{n}a_n(\delta_n)} \leq \frac{F_n}{\sqrt{n}a_n(\delta_n)}.$$

By the definition of $a_n(\cdot)$, $I_{[\,]}(\delta, \tilde{\mathcal{F}}_n, L_2(P_*))$, and the fact that $N_{[\,]}(\delta, \tilde{\mathcal{F}}_n, L_2(P_*))$ is non-increasing in δ, we have

$$\begin{aligned} \frac{1}{a_n(\delta_n)} &= \frac{1}{\delta_n}\sqrt{\max\{1, \log N_{[\,]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*))\}} \\ &\leq \frac{1}{\delta_n^2} I_{[\,]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*)). \end{aligned}$$

It follows that the right hand side of (33) is upper bounded by

$$I_{[\,]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*))\big(1 + \delta_n^{-2} E_* F_n^2\big).$$

Therefore, by Lemma 3 and simple manipulations, we have

$$\begin{aligned} &E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]} \|\mathbb{G}_n \boldsymbol{f}_{n,u}\| \\ &\leq E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]} \sum_{i=1}^{d_n} |\mathbb{G}_n f_{n,u,i}| \\ &\leq d_n E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], 1\leq i\leq d_n[\alpha]} |\mathbb{G}_n f_{n,u,i}| \\ &\leq (A_1 + A_2)\big(1 + \delta_n^{-2}\ E_* F_n^2\big), \end{aligned} \tag{34}$$

where

$$A_1 = d_n\sqrt{2\log\{d_n \mathrm{card}(\mathcal{A}_n)\}}\delta_n,$$
$$A_2 = d_n^{3/2}\sqrt{\mathrm{card}(\mathcal{A}_n)} \sup_{\alpha\in\mathcal{A}_n, 1\leq i\leq d_n[\alpha]} I_{[\,]}(\delta_n, \mathcal{F}_{n,i}[\alpha], L_2(P_*)),$$

Assumptions (ii), (iii), and (iv) guarantee that the right hand side of (34) goes to zero as $n \to \infty$, which concludes the proof. ∎

Using the above results, we can prove the following key technical lemma.

Lemma 6: Suppose that Assumptions 1-6 hold. Then

$$\sup_{\alpha\in\mathcal{A}_n} \|\mathbb{G}_n \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha) - \mathbb{G}_n \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha)\| = o_p(1). \tag{35}$$

Proof: For a constant c, consider the class $\mathcal{F}_n = \cup_{\alpha\in\mathcal{A}_n}\mathcal{F}_n[\alpha]$, with $\mathcal{F}_n[\alpha] \triangleq \{\boldsymbol{f}_{n,\boldsymbol{u}} : \boldsymbol{u} \in U[\alpha]\}$, $U[\alpha] = \{[u_1,\ldots,u_{d_n[\alpha]}]^T : \sum_{i=1}^{d_n[\alpha]} u_i^2 = c\}$, and

$$\boldsymbol{f}_{n,\boldsymbol{u}}(\cdot) = \boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha] + n^{-\tau}\boldsymbol{u};\alpha) - \boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha).$$

Suppose that $\varepsilon,\delta > 0$ are fixed constants. It suffices to prove that the left hand side of (35) is less than δ with probability at least $1-\varepsilon$ for all sufficiently large n. By Lemma 1, there exists a constant $c > 0$ such that $\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha) - \mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)$ falls into the class $\mathcal{F}_n$ with probability at least $1-\varepsilon/2$ for all sufficiently large n. Therefore, we only need to prove that for any given constant $c>0$, $\sup_{\boldsymbol{f}\in\mathcal{F}_n}\|\mathbb{G}_n\boldsymbol{f}\| \to_p 0$. It remains to prove that there are δ_n's that satisfy Conditions (i)-(iv) of Lemma 5.

We define $\mathcal{F}_{n,i}[\alpha]$ as was in Lemma 5, and define $m_n(\cdot) = \sup_{\alpha\in\mathcal{A}_n} m_n[\alpha](\cdot)$. By Assumption 5, we can use $F_n(\cdot) \triangleq cn^{-\tau}\sup_{\alpha\in\mathcal{A}_n} m_n[\alpha](\cdot)$ as the envelop function for each $f_{n,u,i}(\cdot)$, and we have

$$E_*F_n^2 \le C_1 \triangleq c^2n^{-2\tau}E_*m_n^2$$

Let

$$C_2 = d_n\sqrt{\log\{d_n\text{card}(\mathcal{A}_n)\}}.$$

Because of (9) in Assumption 5, we have

$$C_2^2C_1 = c^2n^{-2\tau}d_n^2\log\{d_n\text{card}(\mathcal{A}_n)\}\ E_*m_n^2 \to 0 \qquad (36)$$

This implies the existence of a sequence δ_n (e.g. $\delta_n = C_1^{1/4}C_2^{-1/2}$) such that

$$\delta_nC_2 \to 0, \quad \delta_n^{-2}C_1 \to 0,$$

which further implies Conditions (ii) and (iii) in Lemma 5.

To conclude the proof, we prove that Condition (iv) in Lemma 5 holds for any $\delta_n \to 0$. From Lemma 2, we have for each $\alpha\in\mathcal{A}_n, 1\le i\le d_n[\alpha]$ that

$$\begin{aligned}
&I_{[\,]}(\delta_n,\mathcal{F}_{n,i}[\alpha],L_2(P_*))\\
&\le \int_0^{\delta_n}\left[\max\left\{0, d_n\log\left(\varepsilon^{-1}cn^{-\tau}\|m_n\|_{P_*}\right)\right\}\right]^{1/2}d\varepsilon\\
&= \int_0^{\min\{\delta_n,cn^{-\tau}\|m_n\|_{P_*}\}}\left[d_n\log\left(\varepsilon^{-1}cn^{-\tau}\|m_n\|_{P_*}\right)\right]^{1/2}d\varepsilon
\end{aligned} \qquad (37)$$

Because condition (9) implies that $n^{-\tau}\|m_n\|_{P_*}\to 0$, the value of ε in the integral is close to one. This implies that for all sufficiently large n, the integrand in (37) is upper bounded by $d_n^{1/2}\varepsilon^{-\rho}$, where $1/(1-\rho)=\gamma$ and γ is given in Assumption 5. Therefore, for all sufficiently large n, the right hand side of (37) is upper bounded by

$$\int_0^{cn^{-\tau}\|m_n\|_{P_*}} d_n^{1/2}\varepsilon^{\rho}d\varepsilon = (1-\rho)^{-1}d_n^{1/2}(cn^{-\tau}\|m_n\|_{P_*})^{1-\rho},$$

which does not depend on α, i. This further implies

$$\begin{aligned}
&\sqrt{d_n^{3/2}\ \text{card}(\mathcal{A}_n)} \times \sup_{\alpha\in\mathcal{A}_n,1\le i\le d_n[\alpha]} I_{[\,]}(\delta_n,\mathcal{F}_{n,i}[\alpha],L_2(P_*))\\
&\le (1-\rho)^{-1}c^{1-\rho}d_n\sqrt{\text{card}(\mathcal{A}_n)}(n^{-\tau}\|m_n\|_{P_*})^{1-\rho}\\
&= (1-\rho)^{-1}c^{1-\rho}\left(d_n^{\gamma}\ \text{card}(\mathcal{A}_n)^{\gamma/2}\ n^{-\tau}\|m_n\|_{P_*}\right)^{1-\rho}\\
&\to 0
\end{aligned} \qquad (38)$$

where the last limit is due to (9) in Assumption 5. ∎

We next prove the second key technical lemma.

Lemma 7: Suppose that Assumptions 1-6 hold. Assume that the map $\boldsymbol{\theta}\mapsto E_*\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)$ is differentiable at a $\boldsymbol{\theta}_n^*$ for all n. Then we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]) = &-\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}+\nu_{1,n}[\alpha]\}\cdot\\
&\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{\psi}_n(\boldsymbol{z}_i,\boldsymbol{\theta}_{\boldsymbol{n}}^*[\alpha];\alpha)+\nu_{2,n}[\alpha]
\end{aligned}$$

where $\nu_{1,n}[\alpha]$ is a positive semidefinite matrix and $\boldsymbol{\nu}_{2,n}[\alpha]$ is a vector such that $\sup_{\alpha\in\mathcal{A}_n}\|\nu_{1,n}[\alpha]\| \to_p 0$ and $\sup_{\alpha\in\mathcal{A}_n}\|\boldsymbol{\nu}_{2,n}[\alpha]\|\to_p 0$.

Proof: By the definitions of $\boldsymbol{\theta}_n^*$ and $\hat{\boldsymbol{\theta}}_n$, we have

$$\begin{aligned}
&\sqrt{n}E_*\{\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)\}\\
&=\sqrt{n}\{E_*\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-0\}\\
&=\sqrt{n}\{E_*\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-E_n\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\}\\
&=-\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\
&=-\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\boldsymbol{\nu}_n
\end{aligned} \qquad (39)$$

where the last equality is due to Lemma 6, and $\|\boldsymbol{\nu}_n\| = o_p(1)$,

From the differentiability of the map $\boldsymbol{\theta}\mapsto E_*\psi_n(\cdot,\boldsymbol{\theta};\alpha)$, there exists $\tilde{\boldsymbol{\theta}}[\alpha]$ such that $\|\tilde{\boldsymbol{\theta}}[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\| \le \|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|$, and

$$\begin{aligned}
&E_*\{\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)-\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\}\\
&=\nabla_{\boldsymbol{\theta}}E_*\{\boldsymbol{\psi}_n(\cdot,\tilde{\boldsymbol{\theta}}[\alpha];\alpha)\}(\boldsymbol{\theta}_n^*[\alpha]-\hat{\boldsymbol{\theta}}_n[\alpha])\\
&=V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)(\boldsymbol{\theta}_n^*[\alpha]-\hat{\boldsymbol{\theta}}_n[\alpha]),
\end{aligned} \qquad (40)$$

where the exchangeability of integral and differentiation (in the second identity) is guaranteed by (7) and (8) in Assumption 5. Multiplying the matrix $\sqrt{n}V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}$ to both sides of (40) and using equality (39), we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]) = &-V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\\
&V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}\boldsymbol{\nu}_n.
\end{aligned} \qquad (41)$$

We conclude the proof by applying Assumption 4 (with the constant c_2) and (11) in Assumption 6 to equality (41). ∎

Proof of Theorem 1

In order to prove that the minimum of $\mathcal{L}_n^c[\alpha]$ asymptotically approaches the minimum of $\mathcal{L}_n[\alpha]$ (in the sense of Definition 3), we only need to prove that $\mathcal{L}_n^c[\alpha]/\mathcal{L}_n[\alpha] = 1+o_p(1)$ where $o_p(1)$ is uniform in $\alpha\in\mathcal{A}_n$. In other words,

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{\mathcal{L}_n^c[\alpha]-\mathcal{L}_n[\alpha]}{\mathcal{L}_n[\alpha]}\right| \to_p 0.$$

Recall the definition of $\mathcal{R}_n[\alpha]$. It further suffices to prove that

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{\mathcal{L}_n^c[\alpha]-\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]}\right|\to_p 0, \tag{42}$$

and

$$\sup_{\alpha\in\mathcal{A}_n}\frac{\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]}\to_p 1. \tag{43}$$

By the definition of loss $\mathcal{L}_n[\alpha]$ and Taylor expansion, we have for each $\alpha\in\mathcal{A}_n$

$$\begin{aligned}\mathcal{L}_n[\alpha] &= E_* l_n(\boldsymbol{z},\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\ &= E_* l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])^{\mathrm{T}}\cdot\\ &\quad\frac{\partial}{\partial\boldsymbol{\theta}}E_* l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+\\ &\quad\frac{1}{2}\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|^2_{\nabla^2_{\boldsymbol{\theta}}E_* l_n(\boldsymbol{z},\tilde{\boldsymbol{\theta}}[\alpha];\alpha)}\\ &= E_* l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+\frac{1}{2}\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|^2_{V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)}\end{aligned} \tag{44}$$

where $\tilde{\boldsymbol{\theta}}[\alpha]$ in the second equality is a vector satisfying $\|\tilde{\boldsymbol{\theta}}[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|\le\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|$, and the exchangeability of expectation and differentiation in the third equality is guaranteed by (8) in Assumption 5, and the consistency of $\hat{\boldsymbol{\theta}}_n[\alpha]$. We note that by Assumption 4, the equality (44) further implies (4) presented in our introduction.

Similarly, we have

$$\begin{aligned}\hat{\mathcal{L}}_n[\alpha] &= \frac{1}{n}\sum_{i=1}^n l_n(\boldsymbol{z}_i,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\ &= \frac{1}{n}\sum_{i=1}^n l_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\\ &\quad(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^n\boldsymbol{\psi}_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\\ &\quad\frac{1}{2}\left\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\right\|^2_{\hat{V}_n(\tilde{\tilde{\boldsymbol{\theta}}}[\alpha])}.\end{aligned} \tag{45}$$

From identities (44) and (45), we may write

$$\begin{aligned}&\mathcal{L}_n[\alpha]-\hat{\mathcal{L}}_n[\alpha]-\frac{1}{n}tr\left\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)^{-1}\hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\right\}\\ &= A_3[\alpha]+A_4[\alpha]+A_5[\alpha]+A_6[\alpha]\end{aligned}$$

where we define

$$\begin{aligned}A_3[\alpha] &= \frac{1}{2}\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|^2_{V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)-\hat{V}_n(\tilde{\tilde{\boldsymbol{\theta}}}[\alpha])}\\ A_4[\alpha] &= -\frac{1}{n}\sum_{i=1}^n\{l_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*;\alpha)-E_* l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)\}\\ A_5[\alpha] &= \frac{1}{n}\Big\{tr\big\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\}\\ &\qquad -tr\big\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)^{-1}\hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\big\}\Big\}\\ A_6[\alpha] &= -(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^n\boldsymbol{\psi}_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*[\alpha];\alpha)\\ &\qquad -\frac{1}{n}tr\big\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\}.\end{aligned}$$

In view of (42), it suffices to prove that

$$\sup_{\alpha\in\mathcal{A}_n}\frac{|A_k[\alpha]|}{\mathcal{R}_n[\alpha]}\to_p 0 \tag{46}$$

as $n\to\infty$ for $k=3,4,5,6$, and the limit (43).

By the n^τ-consistency of $\hat{\boldsymbol{\theta}}_n[\alpha]$ uniformly over $\mathcal{A}_n$ (Lemma 1) and Assumption 6,

$$\sup_{\alpha\in\mathcal{A}_n}\frac{|A_3[\alpha]|}{\mathcal{R}_n[\alpha]}=\sup_{\alpha\in\mathcal{A}_n}\frac{1}{2}\frac{n^{-2\tau}}{\mathcal{R}_n[\alpha]}\|\boldsymbol{\nu}_n\|^2_{V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)-\hat{V}_n(\tilde{\tilde{\boldsymbol{\theta}}}[\alpha])}$$

where $\sup_{\alpha\in\mathcal{A}_n}\|\boldsymbol{\nu}_n\|=O_p(1)$. Thus, given assumption (13), (46) with $k=3$ can be proved.

By Chebyshev's inequality, for any positive constant $\delta>0$, we have

$$\begin{aligned}&P_*\left(\sup_{\alpha\in\mathcal{A}_n}\frac{|A_4[\alpha]|}{\mathcal{R}_n[\alpha]}>\delta\right)\\ &\le\sum_{\alpha\in\mathcal{A}_n}P_*\left(\frac{|A_4[\alpha]|}{\mathcal{R}_n[\alpha]}>\delta\right)\\ &\le\sum_{\alpha\in\mathcal{A}_n}\frac{E_*\{l_n(\boldsymbol{z}_1,\boldsymbol{\theta}_n^*;\alpha)-E_* l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)\}^{2m_1}}{\delta^{2m_1}n^{2m_1}\mathcal{R}_n[\alpha]^{2m_1}}.\end{aligned} \tag{47}$$

Thus, given assumption (15), (46) with $k=4$ can be proved.

For brevity, we temporarily denote

$$V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha),\ \hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha),\ J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha),\ \text{and}\ \hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)$$

respectively by

$$V[\alpha],\ \hat{V}[\alpha],\ J[\alpha],\ \text{and}\ \hat{J}[\alpha].$$

Then

$$\begin{aligned}&tr\{V[\alpha]^{-1}J[\alpha]\}-tr\{\hat{V}[\alpha]^{-1}\hat{J}[\alpha]\}\\ &=tr\{V[\alpha]^{-1}(J[\alpha]-\hat{J}[\alpha])\}+tr\{(V[\alpha]^{-1}-\hat{V}[\alpha]^{-1})\hat{J}[\alpha]\}.\end{aligned}$$

To prove (46) with $k=5$, we only need to show that

$$\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n\mathcal{R}_n[\alpha]}tr\{V[\alpha]^{-1}(J[\alpha]-\hat{J}[\alpha])\}\to_p 0, \tag{48}$$

$$\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n\mathcal{R}_n[\alpha]}tr\{(V[\alpha]^{-1}-\hat{V}[\alpha]^{-1})\hat{J}[\alpha]\}\to_p 0. \tag{49}$$

We only prove (48), and then (49) follows similar arguments. Suppose that $\boldsymbol{z}$ is a $\mathcal{N}(0,I)$ random variable of dimension $d_n[\alpha]$, and $V[\alpha]^{-1/2}$ is a positive semidefinite matrix whose square equals $V[\alpha]^{-1}$. Because of Assumption 4 and 5, (48) could be rewritten as

$$\begin{aligned}&\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n\mathcal{R}_n[\alpha]}E\left\{\boldsymbol{z}^{\mathrm{T}}V[\alpha]^{-1/2}(J[\alpha]-\hat{J}[\alpha])V[\alpha]^{-1/2}\boldsymbol{z}\right\}\\ &=o_p(1)\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n\mathcal{R}_n[\alpha]}E\|V[\alpha]^{-1/2}\boldsymbol{z}\|^2\\ &=o_p(1)\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n\mathcal{R}_n[\alpha]}E\|\boldsymbol{z}\|^2\\ &=o_p(1)\sup_{\alpha\in\mathcal{A}_n}\frac{d_n[\alpha]}{n\mathcal{R}_n[\alpha]}\to_p 0\end{aligned}$$

where the first equality is due to (10) in Assumption 6, the second equality is due to Assumption 4, and the last equality is guaranteed by assumption (14).

Next, we prove (46) with $k = 6$. Applying Lemma 7, we could rewrite

$$\frac{|A_6[\alpha]|}{\mathcal{R}_n[\alpha]} = A_7[\alpha] + A_8[\alpha] + A_9[\alpha],$$

where we define

$$A_7[\alpha] = \frac{\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}^*_n[\alpha];\alpha)^{-1}} - tr\big\{V_n(\boldsymbol{\theta}^*_n[\alpha];\alpha)^{-1}J_n(\boldsymbol{\theta}^*_n[\alpha];\alpha)\big\}}{n\mathcal{R}_n[\alpha]}$$

$$A_8[\alpha] = \frac{\|\boldsymbol{w}_n[\alpha]\|^2_{\nu_{1,n}[\alpha]}}{n\mathcal{R}_n[\alpha]}, \quad A_9[\alpha] = \frac{\boldsymbol{\nu}_{2,n}[\alpha]^{\mathrm{T}}\boldsymbol{w}_n[\alpha]}{n\mathcal{R}_n[\alpha]}.$$

Using assumption (16) and similar arguments as in (47), we can prove $\limsup_{\alpha\in\mathcal{A}_n}|A_7[\alpha]| \to_p 0$. Similarly, because

$$|A_8[\alpha]| = o_p(1)\frac{\|\boldsymbol{w}_n[\alpha]\|^2}{n\mathcal{R}_n[\alpha]}$$

where $o_p(1)$ is uniform in $\mathcal{A}_n$, assumption (17) guarantees that $\sup_{\alpha\in\mathcal{A}_n}A_8[\alpha] \to_p 0$. Cauchy inequality and assumption (17) also imply that

$$\sup_{\alpha\in\mathcal{A}_n}|A_9[\alpha]| \le \sup_{\alpha\in\mathcal{A}_n}\frac{\|\boldsymbol{\nu}_{2,n}[\alpha]\| \times \|\boldsymbol{w}_n[\alpha]\|}{n\mathcal{R}_n[\alpha]} \to_p 0. \quad (50)$$

Finally, we prove (43). From (44) and τ-consistency of $\hat{\boldsymbol{\theta}}_n[\alpha]$, we have

$$\begin{aligned}\mathcal{L}_n[\alpha] &= E_* l_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\ &= E_* l_n(\cdot, \boldsymbol{\theta}^*_n[\alpha];\alpha) + n^{-2\tau}O_p(1)\end{aligned}$$

where $O_p(1)$ is uniformly in $\mathcal{A}_n$. Therefore

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}_n}\frac{\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]} &= 1 + \sup_{\alpha\in\mathcal{A}_n}\frac{\mathcal{L}_n[\alpha] - E_*\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]}\\ &= 1 + O_p(1)\sup_{\alpha\in\mathcal{A}_n}\frac{1}{n^{2\tau}\mathcal{R}_n[\alpha]} \to_p 1.\end{aligned}$$

APPENDIX B
PROOF OF COROLLARY 1

A sketch of the proof is outlined below. We only need to verify Assumptions 2 to 7. Assumption 4 is implied by the assumption that X are independent and $V_n(\boldsymbol{\theta}^*_n;\alpha) = 2\Sigma_{xx} = 2I$. Due to the boundedness condition $\|\boldsymbol{\theta}^*_n[\alpha]\| = \|(\Sigma_{xx}[\alpha])^{-1}\Sigma_{x\mu}[\alpha]\| < c\sqrt{d_n}$ for some constant c. We choose $\mathcal{H}_n[\alpha]$ to be $\{\boldsymbol{\theta}\in\mathbb{R}^{d_n[\alpha]} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_n[\alpha]\| < c\sqrt{d_n}\}$. We choose any fixed τ satisfying

$$\max\left\{2w, \frac{\zeta}{2}\right\} < \tau \le \frac{1}{2} - w. \quad (51)$$

For Assumption 2,

$$E_*\ell_n\big(\cdot,\boldsymbol{\theta};\alpha\big) - E_*\ell_n\big(\cdot,\boldsymbol{\theta}^*_n[\alpha];\alpha\big) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_n[\alpha]\|^2_{\Sigma_{xx}[\alpha]} \ge \varepsilon^2$$

for all $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_n[\alpha]\| \ge \varepsilon$. Moreover, $E_n\ell_n\big(\cdot,\boldsymbol{\theta};\alpha\big) - E_*\ell_n\big(\cdot,\boldsymbol{\theta};\alpha\big)$ has mean 0 and variance $n^{-1}\mathrm{Var}\{(Y - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}[\alpha])^2\} = O(d_n^2/n) = o(1)$ uniformly in $\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]$ and $\alpha\in\mathcal{A}_n$.

For Assumption 3, $E_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha) - E_*\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta};\alpha)$ has mean zero and covariance $n^{-1}\mathrm{Var}\{(Y - \boldsymbol{\theta}^{\mathrm{T}}X[\alpha])X[\alpha]^{\mathrm{T}}\}$. Let $A = E_*\{(Y - \boldsymbol{\theta}^{\mathrm{T}}X[\alpha])^2X[\alpha]X[\alpha]^{\mathrm{T}}\}$. Let $\|\cdot\|_F$ denote the Frobenius norm. Since

$$\begin{aligned}&\left\|n^{-1}\mathrm{Var}\{(Y - \boldsymbol{\theta}^{\mathrm{T}}X[\alpha])X[\alpha]^{\mathrm{T}}\}\right\| \le n^{-1}\|A\|\\ &= n^{-1}O(d_n)\|E_*\{X[\alpha]X[\alpha]^{\mathrm{T}}\}\|\\ &\le n^{-1}O(d_n)\|E_*\{X[\alpha]X[\alpha]^{\mathrm{T}}\}\|_F \le n^{-1}O(d_n^2)\end{aligned}$$

uniformly in $\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]$ and $\alpha\in\mathcal{A}_n$. Thus any τ satisfying (51) suffices.

For Assumption 5,

$$\begin{aligned}&\|\boldsymbol{\psi}_n(\boldsymbol{z},\boldsymbol{\theta}_1;\alpha) - \boldsymbol{\psi}_n(\boldsymbol{z},\boldsymbol{\theta}_2;\alpha)\| = \|X[\alpha]X[\alpha]^{\mathrm{T}}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|\\ &\le \|X[\alpha]X[\alpha]^{\mathrm{T}}\|_F\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \le cd_n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.\end{aligned}$$

So $m_n[\alpha] = cd_n$ suffices. This together with the condition $2w < \tau$ implies (9).

For Assumption 6, similar as before, it can be shown that $\|\hat{J}_n(\boldsymbol{\theta};\alpha) - J_n(\boldsymbol{\theta};\alpha)\| = O(n^{-1/2}d^2) = o(1)$, and $\|\hat{V}_n(\boldsymbol{\theta};\alpha) - V_n(\boldsymbol{\theta};\alpha)\| = o(1)$ uniformly in $\boldsymbol{\theta}\in\mathcal{H}_n[\alpha]$ and $\alpha\in\mathcal{A}_n$. Also, $V_n(\boldsymbol{\theta}^*_n;\alpha) - V_n(\boldsymbol{\theta};\alpha) = 0$.

For Assumption 7, (13) is implied by $\zeta/2 < \tau$, and (14) is implied by $\zeta < 1 - w$. Since

$$\begin{aligned}&E_*\big\{l_n(\cdot,\boldsymbol{\theta}^*_n[\alpha];\alpha) - E_*l_n(\cdot,\boldsymbol{\theta}^*_n[\alpha];\alpha)\big\}\\ &\le E_*\big\{l_n(\cdot,\boldsymbol{\theta}^*_n[\alpha];\alpha)\big\}^2 = O(d_n^2),\end{aligned}$$

(15) holds under $m_1 = 1$ and $\zeta < 1 - w$. Similar calculations as before show that (16) and (17) are implied by $\zeta < 1 - 2w$ and $m_2 = m_3 = 1$.

APPENDIX C
PROOF OF THEOREM 2

First, we introduce the concept of "compound experts". A compound expert is defined as an expert sequence $(i_1, i_2, \ldots, i_T)$ whose size $\leq k$ with some prescribed $k > 0$. Then in order to tackle the problem of "tracking the best expert", we could simply apply the exponentially re-weighting algorithm over all the possible compound experts, which can yield provable tight regret bounds. The reason why this simple strategy is not used in practice is that the number of compound experts is usually too large to manage, while the fixed share algorithm greatly reduces the computational complexity and has similar regret bounds.

For our extension of "tracking the best expert" with graphical transitional constraints, following a similar proving strategy used in [34, Chapter 5], we first prove an equivalence between the results of the exponentially re-weighting algorithm over compound experts and the algorithm that we propose, and then apply the regret bound for the former algorithm directly.

The exponentially re-weighting algorithm that we are considering here is as follows. At each time $t = 0, 1, \ldots, T$, the distribution over the compound experts is maintained by $w'_t(i_1, i_2, \ldots, i_T)$ (not necessarily normalized) for all the sequences $(i_1, i_2, \ldots, i_T)$. The initial distribution is

$$\begin{aligned}
&w'_0(i_1, i_2, \ldots, i_T)\\
&= w'_0(i_1) w'_0(i_2|i_1) w'_0(i_3|i_1, i_2) \cdots w'_0(i_T|i_1, \ldots, i_{T-1})\\
&= w'_0(i_1) w'_0(i_2|i_1) w'_0(i_3|i_2) \cdots w'_0(i_T|i_{T-1})\\
&= w'_0(i_1) \prod_{t=1}^{T-1} w'_0(i_{t+1}|i_t)\\
&= 1_{i_1=1} \prod_{t=1}^{T-1} \Big[(1-\kappa\beta_{i_t}) 1_{i_{t+1}=i_t} + \kappa\beta_{i_t,i_{t+1}} 1_{i_{t+1}\neq i_t}\Big],
\end{aligned}$$

where the second equality is due to Markovian property. This initial distribution over compound experts ensures that only the "valid" expert sequences (those follow graphical transitions) have positive probabilities. Based on the exponentially re-weighting updating rule, the distribution at each time instant $t = 1, 2, \ldots, T$ becomes $w'_t(i_1, i_2, \ldots, i_T) = w'_0(i_1, i_2, \ldots, i_T) \exp(-\eta \sum_{s=1}^{t} l(i_s, \boldsymbol{z}_s))$.

Marginally, at time t,

$$w'_{i,t} = \sum_{i_1,\ldots,i_t,i,i_{t+2},\ldots,i_T} w'_t(i_1, \ldots, i_t, i_{t+2}, \ldots, i_T).$$

Then we have $p'_{i,t} = w'_{i,t}/W'_t$ with $W'_t = \sum_{j=1}^{N} w'_{j,t}$, and $p'_{i,0} = w'_{i,0} = 1_{i=1}$. The exponentially forecaster draws action according to expert i at time $t+1$ with probability $p'_{i,t}$.

Lemma 8: For all $\kappa \in (0, 1/D)$, for any sequence of T outcomes, and for all $t = 0, 1, \ldots, T$, the predictive distribution $p_{i,t}$ for $i = 1, \ldots, N$ generated by our proposed Algorithm 1 is the same as the predictive distribution $p'_{i,t}$ for $i = 1, \ldots, N$ that is maintained by the special exponentially re-weighting algorithm described above.

Proof: It is enough to show that for all i and t, $w_{i,t} = w'_{i,t}$. We proceed by induction on t. For $t = 0$, $w_{i,0} = w'_{i,0} = 1_{i=1}$ for all i. For the induction step, assume that $w_{i,s} = w'_{i,s}$ for all i and all $s < t$. We then have

$$\begin{aligned}
w'_{i,t} &= \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_T} w'_t(i_1, \ldots, i_t, i, i_{t+2}, \ldots, i_T)\\
&= \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_T} e^{-\eta \sum_{s=1}^{t} l(i_s,\boldsymbol{z}_s)} \times\\
&\qquad w'_0(i_1, \ldots, i_t, i, i_{t+2}, \ldots, i_T)\\
&= \sum_{i_1,\ldots,i_t} e^{-\eta \sum_{s=1}^{t} l(i_s,\boldsymbol{z}_s)} w'_0(i_1, \ldots, i_t, i)\\
&= \sum_{i_1,\ldots,i_t} e^{-\eta \sum_{s=1}^{t} l(i_s,\boldsymbol{z}_s)} w'_0(i_1, \ldots, i_t) \frac{w'_0(i_1, \ldots, i_t, i)}{w'_0(i_1, \ldots, i_t)}\\
&= \sum_{i_1,\ldots,i_t} e^{-\eta \sum_{s=1}^{t} l(i_s,\boldsymbol{z}_s)} w'_0(i_1, \ldots, i_t) \times\\
&\qquad \Big[(1-\kappa\beta_{i_t}) 1_{i=i_t} + \kappa\beta_{i_t,i} 1_{i\neq i_t}\Big]\\
&= \sum_{i_1,\ldots,i_t} e^{-\eta l(i_t,\boldsymbol{z}_t)} \exp\Big(-\eta \sum_{s=1}^{t-1} l(i_s, \boldsymbol{z}_s)\Big) \times\\
&\qquad w'_0(i_1, \ldots, i_t) \Big[(1-\kappa\beta_{i_t}) 1_{i=i_t} + \kappa\beta_{i_t,i} 1_{i\neq i_t}\Big]\\
&= \sum_{i_t} e^{-\eta l(i_t,\boldsymbol{z}_t)} w'_{i_t,t-1} \Big[(1-\kappa\beta_{i_t}) 1_{i=i_t} + \kappa\beta_{i_t,i} 1_{i\neq i_t}\Big].
\end{aligned}$$

By induction hypothesis, $w'_{i,t}$ further equals

$$\begin{aligned}
&\sum_{i_t} e^{-\eta l(i_t,\boldsymbol{z}_t)} w_{i_t,t-1} \Big[(1-\kappa\beta_{i_t}) 1_{i=i_t} + \kappa\beta_{i_t,i} 1_{i\neq i_t}\Big]\\
&= \sum_{i_t} v_{i_t,t-1} \Big[(1-\kappa\beta_{i_t}) 1_{i=i_t} + \kappa\beta_{i_t,i} 1_{i\neq i_t}\Big]\\
&= (1-\kappa\beta_i) v_{i,t} + \kappa \sum_{j=1}^{N} \beta_{ji} v_{j,t} = w_{i,t}
\end{aligned}$$

where the last equality is by $\beta_{ii} = 0$. ∎

Lemma 9: For all $T \geq 1$, if $l \in [0, 1]$ and we run the exponentially weighted forecaster over compound experts as described before, we will have

$$\sum_{t=1}^{T} \sum_{i=1}^{N} p'_{i,t} l(i, \boldsymbol{z}_t) \leq \frac{1}{\eta} \ln \frac{1}{W'_T} + \frac{\eta}{8} T$$

Proof: First, notice that

$$\begin{aligned}
W'_t &= \sum_{i=1}^{N} w'_{i,t}\\
&= \sum_{i=1}^{N} \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_T} w'_t(i_1, \ldots, i_t, i, i_{t+2}, \ldots, i_T)\\
&= \sum_{i_1,\ldots,i_T} w'_t(i_1, \ldots, i_T).
\end{aligned}$$

Then, we also have

$$\begin{aligned}
&\sum_{i=1}^{N} p'_{i,t} l(i, \boldsymbol{z}_t) = \sum_{i_t} l(i_t, \boldsymbol{z}_t) \frac{w'_{i_t,t}}{W'_{t-1}} \\
&= \sum_{i_t} l(i_t, \boldsymbol{z}_t) \frac{\sum_{i_1,\ldots,i_{t-1},i_{t+1},\ldots,i_T} w'_{t-1}(i_1, \ldots, i_T)}{W'_{t-1}} \\
&= \sum_{i_1,\ldots,i_T} \frac{w'_t(i_1, \ldots, i_T)}{W'_{t-1}} l(i_t, \boldsymbol{z}_t).
\end{aligned}$$

Then we can directly apply Lemma 5.1 in [34, Chapter 5] by noticing that $W'_0 = 1$. ∎

Proof of Theorem 2

Proof: According to Lemma 8, it is equivalent to prove the bound for the equivalent exponentially weighted forecaster. There we have

$$\begin{aligned}
&w'_0(i_1, \ldots, i_T) \\
&= 1_{i_1=1} \prod_{t=1}^{T-1} \left[(1 - \kappa\beta_{i_t}) 1_{i_{t+1}=i_t} + \kappa\beta_{i_t,i_{t+1}} 1_{i_{t+1}\neq i_t} \right] \\
&\geq (1 - \kappa D)^{T-k-1} \kappa^k
\end{aligned}$$

for all the sequence $(i_1, \ldots, i_T)$ with size $\leq k$ and transitions restricted on the graph.

Also, we have

$$\ln w'_T(i_1, \ldots, i_T) = \ln w'_0(i_1, \ldots, i_T) - \eta \sum_{t=1}^{T} l(i_t, \boldsymbol{z}_t).$$

And $W'_T \geq w'_T(i_1, \ldots, i_T)$. Then by Lemma 9 and some simple manipulations, we will get

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\sum_{i=1}^{N} l(i, \boldsymbol{z}_t) p_{i,t} - l(i_t, \boldsymbol{z}_t) \right) \\
&\leq \frac{1}{\eta} (T - k - 1) \log \frac{1}{1 - \kappa D} + \frac{1}{\eta} k \log \frac{1}{\kappa} + \eta \frac{T}{8}.
\end{aligned}$$

In order to minimize the above bound with respect to (κ, η), we first take derivative w.r.t κ and setting it to zero gives $\kappa = k/\big((T-1)D\big)$. Then the bound becomes $S/\eta + \eta T/8$. Minimizing w.r.t. η gives the minimal bound $\sqrt{TS/2}$ with $\eta = \sqrt{8S/T}$. This concludes the proof. ∎

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

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

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

[7] 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.

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

[9] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," 1971.

[10] V. Koltchinskii, "Oracle inequalities in empirical risk minimization and sparse recovery problems," vol. 2033, 2011.

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

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

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

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] 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.

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

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

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

[27] 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.

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

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

[30] 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.

[31] 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.

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

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

[34] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[35] V. Spokoiny *et al.*, "Parametric estimation. finite sample theory," *Ann. Stat.*, vol. 40, no. 6, pp. 2877–2909, 2012.