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 model fitting 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 method has desirable predictive power and less computational cost compared to some popular 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 fitting performance 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 suitable model selection procedure, which will be elaborated 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 $x_1, x_2, \ldots$ 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 $\boldsymbol{z} = [y, x_1, x_2, \ldots]^{\mathrm{T}}$, unknown parameters are $\boldsymbol{\theta} = [\beta_1, \beta_2, \ldots]^{\mathrm{T}}$, and models are subsets of $\{\beta_1, \beta_2, \ldots\}$. 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.

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.

This manuscript was presented in part at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing.

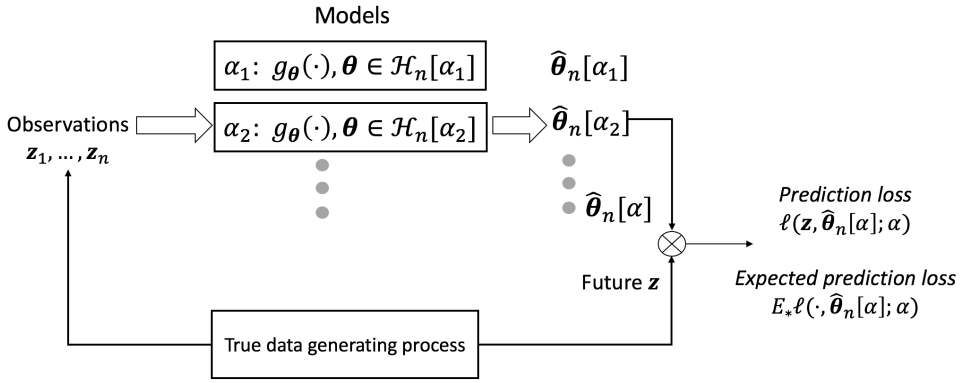


Fig. 1: Illustration of a typical learning procedure, where each candidate model α_j is trained in terms of $\hat{\boldsymbol{\theta}}_n[\alpha_j]$ in its parameter space $\mathcal{H}_n[\alpha_j]$, and then used to evaluate future data under some loss function $\ell(\cdot)$.

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 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ which are independent and identically distributed (i.i.d.), each candidate model α produces an estimator $\hat{\boldsymbol{\theta}}_n[\alpha]$ (referred to as the minimum loss estimator) defined by

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

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 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.

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

Definition 2 (Limit of learning (LoL)): 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

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

Using Taylor expansion under some regularity conditions, $\mathcal{L}_n[\alpha]$ may be expressed as

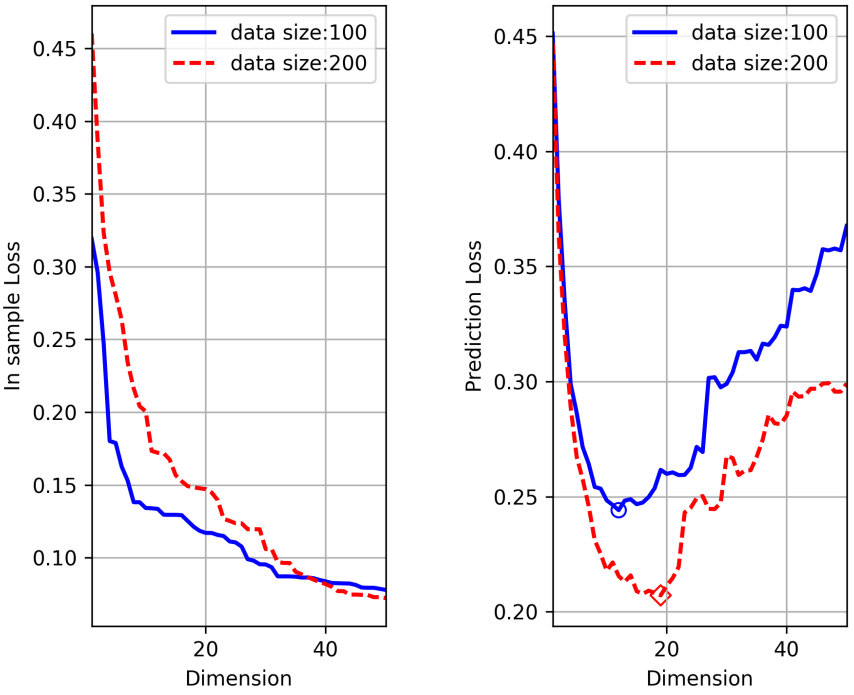
$$\mathcal{L}_n[\alpha] = E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}_n^*[\alpha]; \alpha) + \frac{1}{2}\big(\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\big)^{\mathrm{T}} V_n(\boldsymbol{\theta}_n^*[\alpha]; \alpha)\big(\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\big)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

where $V_n(\boldsymbol{\theta}; \alpha) \stackrel{\Delta}{=} E_* \nabla^2_{\boldsymbol{\theta}} l_n(\cdot, \boldsymbol{\theta}; \alpha)$, and $o_p(1)$ is a sequence of random variables that converges to zero in probability. The main idea of (4) is to expand $\mathcal{L}_n[\alpha]$ at a projection point $\boldsymbol{\theta}_n^*[\alpha]$ under some uniform convergence condition in its vicinity. Theoretical justifications of (4) or its variants for a model whose dimension depends on n have been studied in several earlier work, e.g. in [7]–[9]. 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 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 model fitting as measured by in-sample loss improves (Fig. 2a), while the predictive power as 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.

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 III, we propose a computationally efficient method that determines the most appropriate learning model as more data become available. We prove that the LoL can



(a) The fitting performance 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.

be asymptotically approached under some regularity assumptions. In Section IV, 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 V, 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. 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. We refer to [10]–[14] for more surveys. In this section, we briefly review the some closely related work in information criterion and cross validation, and a derivation of TIC which was

A. Information Criteria

Examples of penalized selection include final prediction error criterion [15], AIC [1], [2], TIC [3], BIC [16] and its Bayesian counterpart Bayes factor [17], minimum description length criterion [18], Hannan and Quinn criterion [19], predictive minimum description length criterion [20], [21], C_p method [22], generalized information criterion (GIC_{λ_n}) with $\lambda_n \to \infty$ [5], [23], [24], generalized cross-validation method (GCV) [25], the Goldenshluger-Lepski method [26]–[28], and the bridge criterion (BC) [29]. Recently, a regularization approach named as information criterion estimation (ICE) [30] is proposed that extends TIC to handle non-MLE estimates in over-parameterized models. An extension of AIC and Mallows' C_p method is the 'slope heuristics' approach proposed in [31], [32] for Gaussian model selection and later developed to more general settings [33]–[35]. The main idea of slope

heuristics is to recognize the existence of a minimal penalty such that the out-sample prediction loss of the selected model with lighter penalties explode, and to show that a penalty equal to twice the minimal penalty often enables model selection that meets the inequality: $\mathcal{L}_n[\hat{\alpha}_n] \leq c_n \min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n[\alpha] + \eta_n$, also called oracle inequality, for c_n close to 1 and η_n negligible with respect to the value of $\min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n[\alpha]$. In theory, the asymptotic efficiency is a limiting requirement of the oracle inequality with $c_n = 1 + o_p(1)$ and $\eta_n = o_p(1) \times \min_{\alpha\in\mathcal{A}_n} \mathcal{L}_n[\alpha]$ as $n \to \infty$. There have been fruitful results in non-asymptotic quantifications of c_n and η_n using concentration inequalities (see e.g. [34]–[39]). Non-asymptotic analysis is often based on concentration inequalities or Stein's method [11], [40]. In this work, we are not looking for oracle inequalities with non-asymptotic analysis. On the other hand, the recent development of slope heuristics has motivated data-driven construction of penalty terms instead of using pre-determined penalty functions. An example in this direction is the dimension jump method [32], [41] which, for a given penalty shape, identifies the suitable multiplicative constant by searching for a significant jump of the selected dimension against different constants.

B. Cross-validation (CV)

The basic idea of cross-validation [42], [43] 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 $\lim_{n\to\infty} d/n = 1$ [44]–[47], or the delete-1 CV method (or leave-one-out, LOO) [48] can exhibit asymptotic 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/TIC and GCV if $d_n[\alpha] = o(n)$ [5], [48], which is asymptotically efficient in a mis-specified model class, but usually overfits in a well-specified model class. An appropriate choice of the splitting ratio often depends on specific learning tasks, such as the prediction of unobserved data, selection of model, selection of other criteria [49], goodness-of-fit test [50]. We refer to [5], [13], [14], [29], [51] for more detailed discussions on the discrepancy and reconciliation of different CVs.

In particular, for the prediction purpose, common folklore that advocates the use of k-fold or 30%-for-testing CV are 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.

C. Background of TIC

TIC [3] was heuristically derived as an alternative of AIC, also from an information-theoretic view rooted in Kullback-Leibler (KL) divergence. In the seminar work of [48], TIC is shown to be asymptotically equivalent to cross-validation when the purpose is to minimize the KL divergence, and AIC is a special case of TIC when the models under consideration are well-specified. It does not appear to be widely appreciated nor used [52] compared with other information criteria such as AIC or Bayesian information criterion (BIC) [16]. In terms of provable asymptotic performance, only AIC is known to be asymptotically efficient for variable selection in regression models [53] and autoregressive order selection in time series models [54], [55] when models are mis-specified. Conceptually, TIC was proposed as a surrogate for AIC in general mis-specified settings, but the optimality of AIC and TIC in the general context remains unknown. As the original paper of TIC [3] was not written in English, we review it for the completeness of the paper. Similar derivations can be found in, e.g. [30].

Suppose that our goal is to select the model α that minimizes logarithmic loss $E_*\{-\log p_{\hat{\boldsymbol{\theta}}_n[\alpha]}(Y)\}$ (or equivalently, minimizes the KL divergence from the true data-generating distribution), where $\hat{\boldsymbol{\theta}}_n[\alpha]$ is the MLE under model α. For notational convenience, we drop the model index α and focus on one model. The motivation of

TIC was to approximate $E_*\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z})\}$ by $n^{-1}\sum_{i=1}^n\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z}_i)\}+\lambda_n$, where the first term is computable from data and the second term is to be asymptotically approximated. Under some regularity conditions, the classical sandwich formula of MLE [56, Theorem 3.2] gives $\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\to_d\mathcal{N}(0,V^{-1}JV^{-1})$ for some $\boldsymbol{\theta}^*$ in the parameter space, with

$$V=-E_*\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log p_{\boldsymbol{\theta}}(Y)\right\},\ J=E_*\left\{\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}^*}(Y)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}^*}(Y)\right)^{\mathrm{T}}\right\}.$$

Applying Taylor expansion at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, we have

$$E_*\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z})\}\approx E_*\{-\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z})\}+\frac{1}{2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^{\mathrm{T}}V(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)$$

$$n^{-1}\sum_{i=1}^n\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z}_i)\}\approx n^{-1}\sum_{i=1}^n\{-\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)\}-(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^{\mathrm{T}}n^{-1}\sum_{i=1}^n\frac{\partial\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}}+\frac{1}{2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^{\mathrm{T}}V(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)$$

and thus

$$\lambda_n=E_*\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z})\}-n^{-1}\sum_{i=1}^n\{-\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z}_i)\}\approx(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^{\mathrm{T}}n^{-1}\sum_{i=1}^n\frac{\partial\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}}\tag{5}$$

for large n. Using

$$n^{-1}\sum_{i=1}^n\frac{\partial\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}}=n^{-1}\sum_{i=1}^n\frac{\partial\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}}-n^{-1}\sum_{i=1}^n\frac{\partial\log p_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}}\approx n^{-1}\sum_{i=1}^n\frac{\partial^2\log p_{\boldsymbol{\theta}^*}(\boldsymbol{z}_i)}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_n)$$

and the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)$, we may further approximate λ_n by $(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^{\mathrm{T}}V(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\approx n^{-1}E(\boldsymbol{\zeta}^{\mathrm{T}}V\boldsymbol{\zeta})=n^{-1}tr(V^{-1}J)$, where $\boldsymbol{\zeta}\sim\mathcal{N}(0,V^{-1}JV^{-1})$. For a well-specified model, we have $V=J$ and $\lambda_n\approx n^{-1}d$ with d denoting the model dimension, and thus TIC becomes AIC.

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 III-C. 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.

III. 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 $E_nf(\cdot)=n^{-1}\sum_{i=1}^nf(\boldsymbol{z}_i)$. 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 $\hat{\mathcal{L}}_n[\alpha] \triangleq E_n l_n\big(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big)$. Similarly, we define

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

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 $B(\boldsymbol{c}, r) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{c}\| \leq r\}$. 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 model fitting 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, 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

$$\frac{\mathcal{L}_n[\hat{\alpha}_n]}{\min_{\alpha \in \mathcal{A}_n} \mathcal{L}_n[\alpha]} \to_p 1 \tag{6}$$

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 [54] and variable selection in linear regression models [53].

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 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 [57] for a classification model). A definitive treatment of this line of work can be found in, e.g., [39], [58]–[60] and the references therein. The major difference of our setting compared with that in learning theory is our requirement that the positive term plus the in-sample loss should asymptotically approach the true out-sample loss (as sample size goes to infinity).

Another related notion often used to describe model selection performance is minimax-rate optimality [10], [51]. In nonparametric estimation of the regression function $f \in \mathcal{F}$ under the squared loss, tight minimax risk bounds for $\inf_{\hat{f}} \sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^n E_*(\hat{f}(x_i) - f(x_i))^2$ have been obtained since the pioneering work of [61], [62] (see [63], [64] for more discussions). A model selection method ν is said to be minimax-rate optimal over $\mathcal{F}$, if $\sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^n E_*\{\hat{f}_\nu(x_i) - f(x_i)\}^2$ converges at the same rate as the aforementioned minimax risk, where $\hat{f}_\nu$ is the least squares estimate of f under the variables selected by ν. In contrast to the notion of asymptotic efficiency which we focus on here, minimax-rate optimality allows the true data-generating model to vary and thus it is a stronger requirement. The asymptotic efficiency is in a pointwise sense, meaning that the data are already generated by some fixed but unknown data-generating process. It has been proved that AIC is minimax-rate optimal for a range of variable selection tasks, and there exists no model selection method that achieves such optimality as well as selection consistency [51]. Meanwhile, it is possible to simultaneously combine asymptotic efficiency and selection consistency, and that motivated recent research in reconciling AIC-type and BIC-type model selection methods [29], [49], [65], [66].

We propose to use the following penalized model selection procedure, which extends 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

$$\mathcal{L}_n^c[\alpha] \triangleq \hat{\mathcal{L}}_n[\alpha] + n^{-1} tr\big\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)^{-1}\hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\big\}. \tag{7}$$

We note that the two additive terms on the right hand side of (7) represent the fitting performance and the model complexity, respectively.

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.

C. 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 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 [67] which is more commonly assumed for time series data.

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

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

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

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

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

Assumption 2 is the counterpart of the separated mode and uniform law of large number conditions that have been commonly required in proving the consistency of maximum likelihood estimator for classical statistical models (see, e.g. [4, Theorem 5.7]). The $\boldsymbol{\theta}_n^*[\alpha]$ can be interpreted as the oracle optimum under model α, or a "projection" point of the true data generating distribution onto the model α.

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 $\boldsymbol{\theta} \in \text{int}(\mathcal{H}_n[\alpha])$ for all n and $\alpha \in \mathcal{A}_n$.

Assumption 3 is a weaker statement compared with 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 sample size n. As was pointed out by an anonymous reviewer, it is technically possible to replace n^{τ} with a weaker requirement, say $n^{0.5}/(\log n)^{\gamma}$ for any constant $\gamma > 0$.

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

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

Assumption 4 assumes that the second derivative of the out-sample prediction loss has bounded eigenvalues at the optimum $\boldsymbol{\theta}_n^*[\alpha]$. The lower bound c_1 indicates that the loss function is strongly convex for all models, and the upper bound c_2 requires the loss functions to be reasonably smooth. This assumption is used in our asymptotic analysis to ensure reasonable Taylor expansions up to the second order.

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

$$\| \boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}_1; \alpha) - \boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}_2; \alpha) \| \leq m_n[\alpha](\boldsymbol{z}) \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|, \tag{8}$$
$$E_* m_n[\alpha] < \infty. \tag{9}$$

Moreover, we have

$$\max\left\{ d_n^{\gamma} \ \text{card}(\mathcal{A}_n)^{\gamma/2}, \ d_n \sqrt{\log\{d_n \text{card}(\mathcal{A}_n)\}} \right\} \ \times n^{-\tau} \left\| \sup_{\alpha \in \mathcal{A}_n} m_n[\alpha] \right\|_{P_*} \to 0. \tag{10}$$

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 (10) explicitly requires that the largest dimension d_n and the candidate size $\text{card}(\mathcal{A}_n)$ do not grow too fast as n goes to infinity. The condition (10) is used to bound the rate of convergence of the empirical process $\mathbb{G}_n \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)$ in the vicinity of $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n[\alpha]$. Similar conditions were often used to establish asymptotic results such as the Cramér-Rao bound [68, Theorem 18].

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

$$\sup_{\alpha \in \mathcal{A}_n} \sup_{\boldsymbol{\theta} \in \mathcal{H}_n[\alpha] \cap B(\boldsymbol{\theta}_n^*[\alpha], \delta)} \| \hat{J}_n(\boldsymbol{\theta}; \alpha) - J_n(\boldsymbol{\theta}; \alpha) \| \to_p 0, \tag{11}$$
$$\sup_{\alpha \in \mathcal{A}_n} \sup_{\boldsymbol{\theta} \in \mathcal{H}_n[\alpha] \cap B(\boldsymbol{\theta}_n^*[\alpha], \delta)} \| \hat{V}_n(\boldsymbol{\theta}; \alpha) - V_n(\boldsymbol{\theta}; \alpha) \| \to_p 0, \tag{12}$$
$$\lim_{\varepsilon \to 0} \sup_{\alpha \in \mathcal{A}_n} \sup_{\boldsymbol{\theta} \in \mathcal{H}_n[\alpha] \cap B(\boldsymbol{\theta}_n^*[\alpha], \varepsilon)} \| V_n(\boldsymbol{\theta}; \alpha) - V_n(\boldsymbol{\theta}_n^*; \alpha) \| = 0. \tag{13}$$

Assumption 6 requires that the sample analogs of the 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 $\boldsymbol{\theta}_n^*[\alpha]$.

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.

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

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

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

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

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

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

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

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

In Assumption 7, the conditions (14), (15) and (18) indicate that the risks $\mathcal{R}_n[\alpha]$ for all α are not small so that the model class is virtually mis-specified. The assumptions in equalities (16) and (17) are central moment constraints that control the regularity of loss functions. Similar conditions were often used to establish the asymptotic performance of model selection, for example [5, Condition (2.6)] and [69, Condition (A.3)].

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).

Classical asymptotic analysis for general parametric models with i.i.d. observations typically relies on a type of uniform convergence of empirical process around $\boldsymbol{\theta}_n^*[\alpha]$ within a fixed parameter space. Because our functions are vector valued with dimension depending on the sample size n, we cannot directly use state-of-art technical tools such as those in [4, Theorem 19.28]. The classical proof by White [56] (in proving asymptotic normality in mis-specified class) cannot be directly adapted, either, for parameter spaces that depend on n. On the other hand, though asymptotic analysis for criteria such as AIC, C_p, CV, GIC often consider models that depend on n (see e.g. [5], [51], [53], [69], it is often studied in the context of fixed-design regression models, so the technical tools there cannot be directly applied for our purpose.

Some new technical tools are needed in our proof. Here we sketch some technical ideas in the proof. 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.

D. 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 $Y=\mu(\boldsymbol{X})+\varepsilon$, where ε is a random noise with 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 $\mu=\beta_1X_1+\cdots+\beta_{d_n}X_{d_n}$, where d_n may or may not depend on the sample size n. We sometimes write $\mu(\boldsymbol{X})$ 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 $\boldsymbol{\theta}=(\beta_1,\ldots,\beta_{d_n})$. The model class, denoted by $\mathcal{A}_n$, consists of candidate models represented by $\alpha\subseteq\{1,\ldots,d_n\}$, i.e. $\mu(\boldsymbol{X})=\sum_{i\in\alpha}\beta_iX_i$.

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

$$E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) = E_* \left(\mu - \sum_{j \in \alpha} \beta_j x_j \right)^2. \tag{19}$$

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 $\Sigma_{xx}[\alpha]$, $\Sigma_{x\mu}[\alpha]$, $\boldsymbol{X}[\alpha]$ 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 (19) may be rewritten as

$$\begin{aligned} E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) &= E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}_n^*[\alpha]; \alpha) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*[\alpha]\|^2_{\Sigma_{xx}[\alpha]} \\ &= \left(\Sigma_{\mu\mu} - \Sigma_{\mu x}[\alpha] \Sigma_{xx}[\alpha]^{-1} \Sigma_{x\mu}[\alpha] \right) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*[\alpha]\|^2_{\Sigma_{xx}[\alpha]}. \end{aligned} \tag{20}$$

The decomposition in (20) 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. Evaluating $\hat{V}_n(\boldsymbol{\theta}; \alpha)$ and $\hat{J}_n(\boldsymbol{\theta}; \alpha)$ in this specific case, we obtain

$$\hat{V}_n(\boldsymbol{\theta}; \alpha) = 2E_n(\boldsymbol{X}[\alpha] \boldsymbol{X}[\alpha]^{\mathrm{T}}), \quad \hat{J}_n(\boldsymbol{\theta}; \alpha) = 4E_n(\boldsymbol{X}[\alpha] \boldsymbol{X}[\alpha]^{\mathrm{T}} R[\alpha]^2), \quad R[\alpha] \stackrel{\Delta}{=} Y - \hat{\boldsymbol{\theta}}_n[\alpha]^{\mathrm{T}} \boldsymbol{X}[\alpha].$$

Note that when $R[\alpha]$ is close to the independent noise term, then $E_*(R[\alpha]^2) \approx \sigma^2$ and the GTIC penalty in (7) is around $n^{-1}(2d\sigma^2)$ which approximates the AIC and Mallows' C_p method. Theorem 1 implies the following corollary. In verifying the previous assumptions such as Assumption 2 for this corollary, we used the fact that $\|\boldsymbol{\theta}_n^*[\alpha]\| = \|(\Sigma_{xx}[\alpha])^{-1} \Sigma_{x\mu}[\alpha]\| = \|\Sigma_{x\mu}[\alpha]\| \leq c^2 d_n$, and the least squares estimates fall into $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq 2c\sqrt{d_n}\}$ with high probability (due to the concentration inequalities for bounded X_i and μ). It is possible to relax the conditions by a 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 with zero mean and unit variance 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) $\mathrm{card}(\mathcal{A}_n) = o(n^{2(1-\zeta-w)})$.

IV. 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 candidate models, 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:

$$\min_{\text{size}(i_1,\ldots,i_T)\le k,\ i_1,\ldots,i_T\in\{1,\ldots,N\}} \sum_{t=1}^{T} \mathcal{L}_n[\alpha_{i_t}]. \tag{21}$$

where the size of a sequence $\text{size}(i_1, \ldots, i_T)$ 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 fully-connected/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 $l : \{1, 2, \ldots, N\} \times \mathcal{Y} \to \mathbb{R}$. A smaller loss indicates a better prediction. 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 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 a 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 [70]–[72]. The regret is a relevant criterion of evaluating the predictive performance in sequential settings, since the model and model parameters have to be adjusted on a rolling basis as new data arrives, and a selected model at a time step t_1 may not be suitable at another time step t_2. If the benchmark in the regret is defined as the minimum cumulative loss achieved by a single expert in hindsight, namely $\min_{1\le i^*\le N} \sum_{t=1}^{T} l(i^*, \boldsymbol{z}_t)$, then it is standard to apply the exponential re-weighting procedure which produces some desirable regret bound [71, Chapter 2]. In many cases the best performing expert can be different from one time segment to another, motivating the benchmark

$$\min_{\text{size}(i_1,\ldots,i_T)\le k,\, i_1,\ldots,i_T\in\{1,\ldots,N\}} \sum_{t=1}^{T} l(i_t, \boldsymbol{z}_t)$$

where k denotes the maximum number of switches of the best experts in hindsight. In this scenario, the fixed share algorithm [71, Chapter 5] can be a good solution with guaranteed regret bound. We consider the following problem setting that aims to significantly reduce computational costs.

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

$$\beta_{ij} = 1_{\exists (i,j)\in E}, \tag{22}$$

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), (23)
6: **end for**

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

$$\beta_i = \sum_{j=1}^{N} \beta_{ij}, \tag{23}$$

which is the out-degree of the node i. In addition, we assume that $\max_{i \in 1, \ldots, N} \beta_i \leq D$, 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. A regret bound will be derived that only depends on the graph degree D instead of the number of experts N. To the best of the authors' knowledge, the framework concerning dynamic regret with graph constraints stated here has not been studied before.

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

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

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$(i_1, i_2, \ldots, i_T) \leq k$.

The left hand side of the above inequality is referred to as *regret*. In order to minimize the above regret bound with respect to (κ, η), we first take derivative with respect to the sharing rate κ and solve the first order equation to obtain $\kappa = k/\big((T-1)D\big)$. Then the bound becomes $S/\eta + \eta T/8$. We further minimize it over the learning rate η to obtain $\eta = \sqrt{8S/T}$. The corresponding minimal bound is calculated to be $\sqrt{TS/2}$. Here

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

and $H(\cdot)$ is the binary entropy function defined by $H(x) \stackrel{\Delta}{=} -x \log x - (1 - x) \log(1 - x)$ for $x \in (0, 1)$, $H(0) = H(1) = 0$. The T is interpreted as a pre-determined stopping time, when the performance of the data-driven algorithm is to be contrasted with that of the optimal graph search (with k node switches). In particular, for small k/T, the average of the regret is at the order of

$$\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} l(i, \boldsymbol{z}_t) p_{i,t} - l(i_t, \boldsymbol{z}_t) \right) \leq \frac{\sqrt{TS/2}}{T} \sim T^{-1/2} \cdot \left\{ T \cdot \frac{k}{T} \cdot \log T + k \log D \right\}^{1/2} \sim (k/T)^{1/2} \log(DT)$$

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).

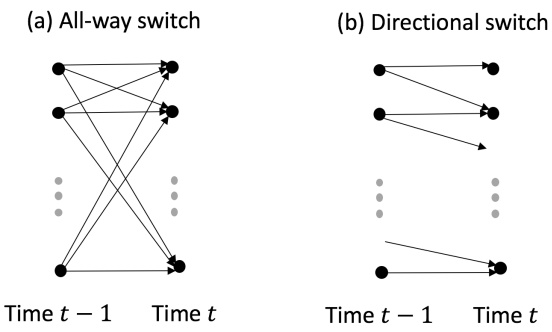


Fig. 3: Illustration of the state-of-art and our new way of redistributing the share of weights in online learning.

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 [71, Chapter 5] is a special case when $\beta_{ij} = 1$ for all $i \neq j$ 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 models and the corresponding D is 1. 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 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 $\sup_{1\leq t\leq T}\sup_{\alpha\in\mathcal{A}_n}\mathcal{L}_t^c[\alpha] < c$ 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

$$\frac{1}{T}\left(\sum_{t=1}^{T}\sum_{i=1}^{\text{card}(\mathcal{A}_T)} p_{i,t}\,\mathcal{L}_t^c[\alpha_i] - \min_{\text{size}(i_1,i_2,\ldots,i_T)\leq k}\sum_{t=1}^{T}\mathcal{L}_t^c[\alpha_{i_t}]\right) \leq \frac{c}{\sqrt{2}}\sqrt{H\left(\frac{k}{T-1}\right)} \tag{24}$$

for all $T \geq 1$, given that

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

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

$$\limsup_{T\to\infty}\frac{1}{T}\left(\sum_{t=1}^{T}\sum_{i=1}^{\text{card}(\mathcal{A}_T)} p_{i,t}\,\mathcal{L}_t^c[\alpha_i] - \min_{\text{size}(i_1,i_2,\ldots,i_T)\leq k}\sum_{t=1}^{T}\mathcal{L}_n[\alpha_{i_t}]\right) \leq 0 \tag{25}$$

almost surely.

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

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

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

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 $k = 1, \ldots, K$, where $\mathcal{L}_t^c[\alpha]$ is calculated from (7) and fitting the data $\boldsymbol{z}_1, \ldots \boldsymbol{z}_t$ 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 $p_{t,k} = (\sum_{k=1}^{K} w_{t,k})^{-1} w_{t,k}$, $k = 1, \ldots, K$

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 $k = 1, \ldots, K$ and $k' = (k + 1 \bmod K)$ (relabeling the active models)

8: **end if**

9: **end for**

Next, we explain some details regarding 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 III-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 (with $D = 1$), 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 mainly for computational purpose, and we experimentally observed that it does not deteriorate the predictive performance. Theoretically, this is because the excluded models are often overly large or small, and their weights are thus negligible. Next, we make some specific assumptions to illustrate the above idea. Suppose that the corrected prediction loss of model α with dimension d_α is approximately $\mathcal{L}_t^c[\alpha] = c_1 d_\alpha^{-\gamma} + c_2 d_\alpha / t$ for some positive constants c_1, c_2, γ, consisting of a bias and a variance term at each time step t. Let $w_t[\alpha], v_t[\alpha]$ denote the counterpart of $w_{t,k}, v_{t,k}$ for each model $\alpha \in \mathcal{A}_n$ if it were calculated, and suppose that $v_t[\alpha_1] \le \cdots \le v_t[\alpha_m]$ for a certain m. For any $1 \le i < j \le m$, we have

$$\begin{aligned} w_t[\alpha_j] &\ge (1-\kappa) v_t[\alpha_j] = (1-\kappa) \exp\{-\eta \mathcal{L}_t^c[\alpha_j]\} w_{t-1}[\alpha_j] \\ &\ge (1-\kappa)^T \exp\{-\eta \sum_{t=1}^{T} Lc[\alpha_j]\} \sim (1-\kappa)^T \exp\{-\eta c_1 T d_{\alpha_j}^{-\gamma} - \eta c_2 d_{\alpha_j} \log T\} \end{aligned}$$

while on the other hand

$$w_t[\alpha_i] \le v_t[\alpha_i] = w_{t-1}[\alpha_i]\exp\{-\eta\mathcal{L}_t^c[\alpha_i]\}$$
$$\le \exp\{-\eta\sum_{t=1}^{T}\mathcal{L}_t^c[\alpha_i]\} \sim \exp\{-\eta c_1 T d_{\alpha_i}^{-\gamma} - \eta c_2 d_{\alpha_i}\log T\}$$

For small k, it can be verified that $(1-\kappa)^T \sim \exp(-k)$ and

$$w_t[\alpha_i]/w_t[\alpha_j] = O(1)\times\exp\{-\eta c_1 T(d_{\alpha_i}^{-\gamma} - d_{\alpha_j}^{-\gamma})\} = O(1)\times\exp\{-c_3(\log T)(d_{\alpha_j} - d_{\alpha_i})\}.$$

for $d_{\alpha_i} < d_{\alpha_j} \le c_4(T/\log T)^{1/(1+\gamma)}$ for some constants c_3, c_4. This indicates that relative weights of underfitting models will exponentially decay as the dimension deviates more from the optimum. We leave a more sophisticated analysis for future work.

An anonymous reviewer pointed out that the above sequential model selection would be most useful when combined with a sequential update of the model parameters (for a given model). Suppose that $\hat{\theta}_n$ is the MLE of a parameter θ under t data observations. Here we summarize two common methods that could be potentially used in online implementations: 1) calculate asymptotic expression of $\hat{\theta}_t - \hat{\theta}_{t-1}$, also called the asymptotic influence function, to update from $\hat{\theta}_{t-1}$ to $\hat{\theta}_t$ (see e.g. [4, Theorem 5.23]), and 2) use $\hat{\theta}_{t-1}$ as a warm start for estimating $\hat{\theta}_t$ when using iterative algorithms such as the (stochastic) gradient descent and Newton-Raphson method. Additionally, as another anonymous reviewer pointed out, the sequential algorithm is compatible with any model selection criterion that enjoys the efficiency property, and Proposition 1 is not only specific to GTIC.

V. Numerical Experiments

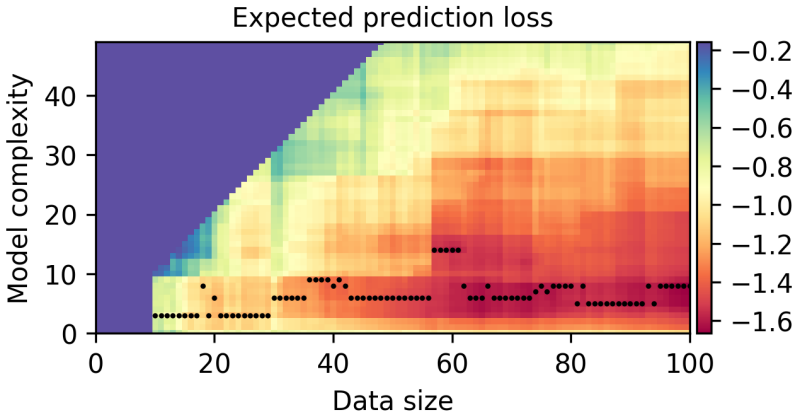
The model classes under consideration are logistic regression and feed-forward neural networks. We also released an open source python package 'gtic' at https://pypi.python.org/pypi/gtic, in which we build a tensor graph of GTIC upon the 'theano' platform. Users can simply provide their tensor variables of loss and parameters, and obtain the GTIC instantly.

A. Logistic Regression Models

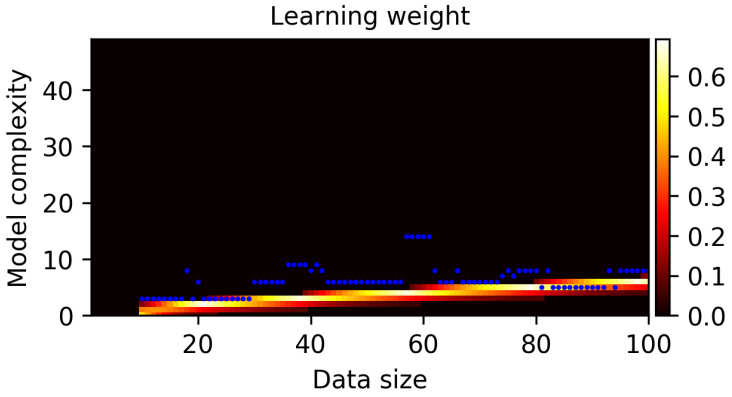
We generate data from a logistic regression model, where the coefficient vector is $\boldsymbol{\beta} = 10\times[1^{-1.5},\ldots,100^{-1.5}]^{\mathrm{T}}$, and covariates $x_1,\ldots,x_{100}$ are independent standard Gaussian. Suppose that we sequentially obtain and learn the data, starting from $t=10$, and then $t=11,\ldots,100$. We consider the model class that consists of logistic regression models of dimensions $1,\ldots,\lfloor\sqrt{t}\rfloor$ at each time step t. A model of dimension d means that the coefficients of the first d variables $x_1,\ldots,x_d$ are unknown parameters and the coefficients of the remaining variables are restricted to be zero. The model class is nested because a small model is a special case of a large model. Since the maximum dimension of candidate models is restricted to be no larger than $\sqrt{100}=10$, the model class is considered as mis-specified. 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 (Fig. 5a). 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. GTIC is no more than 10 times faster than 10-fold CV, this is because we need to compute the penalty term in GTIC plus the evaluation of n data, while each of the 10 folds uses less than n data. 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



(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

CV, we suggest using GTIC instead of guessing or searching 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, slope heuristics (denoted by SH), and TIC under various choices of sample size n and number of covariates p. The SH method is based on the AIC shape (linear in dimension) and the 'dimension jump' approach to determine the multiplicative constant [34]. 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. The results are summarized with three evaluation metrics, mainly because they are all used in practice and their interpretations are different. The out-sample prediction loss of Tables II, also referred to as the "testing loss", shows the KL divergence from the true data generating model to the learned model (up to an additive constant). The prediction accuracy of Table III, also referred to as the "classification accuracy", shows the accuracy of threshold decisions and it is not a proper loss for training. The prediction efficiency of Table IV shows the relative deviation from the underlying truth compared with the best candidate model in hindsight.

TABLE II: Performance comparison in terms of the out-sample prediction loss, for data generated from two different logistic regression models. The values are averaged from 50 independent replications, with standard errors in the parentheses.

	$\mathcal{M}_1$				$\mathcal{M}_2$			
	AIC	BIC	SH	TIC	AIC	BIC	SH	TIC
$n = 100, p = 20$	**0.41 (0.01)**	0.53 (0.03)	**0.41 (0.01)**	**0.41 (0.01)**	0.38 (0.02)	0.54 (0.03)	0.38 (0.02)	**0.37 (0.01)**
$n = 100, p = 50$	0.66 (0.02)	0.68 (0.01)	0.61 (0.01)	**0.51 (0.02)**	0.68 (0.03)	0.71 (0.01)	0.69 (0.02)	**0.47 (0.02)**
$n = 300, p = 60$	**0.34 (0.01)**	0.64 (0.02)	**0.34 (0.02)**	**0.34 (0.01)**	**0.30 (0.01)**	0.44 (0.05)	0.31 (0.01)	**0.30 (0.01)**
$n = 300, p = 150$	0.81 (0.05)	0.69 (0.00)	0.67 (0.02)	**0.53 (0.02)**	0.81 (0.03)	0.70 (0.00)	0.65 (0.01)	**0.50 (0.01)**
$n = 500, p = 100$	**0.34 (0.01)**	0.68 (0.01)	0.36 (0.01)	**0.34 (0.01)**	0.30 (0.01)	0.44 (0.05)	**0.29 (0.01)**	**0.29 (0.01)**
$n = 500, p = 250$	0.95 (0.04)	0.69 (0.00)	0.65 (0.02)	**0.54 (0.02)**	0.96 (0.03)	0.69 (0.00)	0.55 (0.01)	**0.51 (0.01)**

TABLE III: Performance comparison in terms of the out-sample prediction accuracy, for data generated from two different logistic regression models. The values are averaged from 50 independent replications, with standard errors in the parentheses.

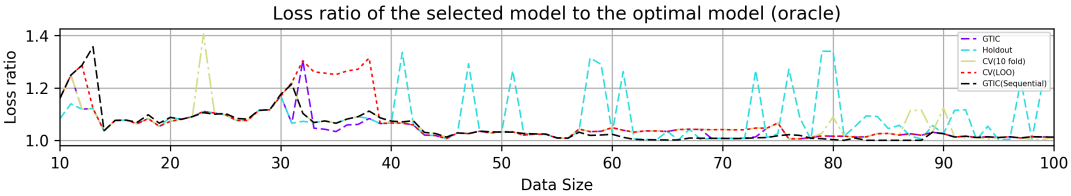
	$\mathcal{M}_1$				$\mathcal{M}_2$			
	AIC	BIC	SH	TIC	AIC	BIC	SH	TIC
$n = 100, p = 20$	**0.80 (0.01)**	0.74 (0.02)	**0.80 (0.01)**	**0.80 (0.01)**	0.82 (0.02)	0.75 (0.05)	0.82 (0.02)	**0.84 (0.01)**
$n = 100, p = 50$	0.68 (0.03)	0.54 (0.07)	0.73 (0.02)	**0.77 (0.02)**	0.73 (0.03)	0.63 (0.06)	0.73 (0.03)	**0.80 (0.02)**
$n = 300, p = 60$	**0.82 (0.01)**	0.60 (0.03)	**0.82 (0.01)**	**0.82 (0.01)**	**0.86 (0.01)**	0.80 (0.03)	**0.86 (0.02)**	**0.86 (0.01)**
$n = 300, p = 150$	0.73 (0.02)	0.57 (0.09)	0.74 (0.02)	**0.78 (0.01)**	0.76 (0.02)	0.67 (0.08)	0.79 (0.03)	**0.81 (0.02)**
$n = 500, p = 100$	**0.82 (0.01)**	0.57 (0.04)	0.81 (0.01)	**0.82 (0.01)**	**0.87 (0.01)**	0.77 (0.06)	**0.87 (0.01)**	**0.87 (0.01)**
$n = 500, p = 250$	0.72 (0.01)	0.50 (0.08)	0.76 (0.01)	**0.80 (0.01)**	0.72 (0.01)	0.40 (0.05)	0.75(0.02)	**0.79 (0.01)**

TABLE IV: Performance comparison in terms of the out-sample prediction efficiency, for data generated from two different logistic regression models. The values are averaged from 50 independent replications, with standard errors in the parentheses.

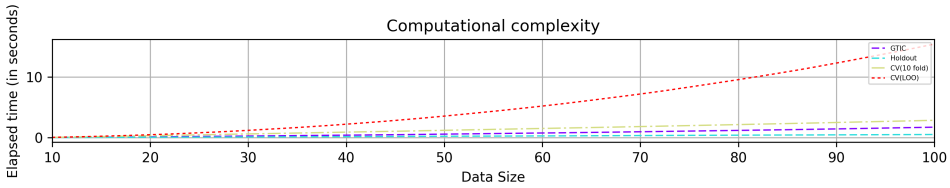
	$\mathcal{M}_1$				$\mathcal{M}_2$			
	AIC	BIC	SH	TIC	AIC	BIC	SH	TIC
$n = 100, p = 20$	**0.97 (0.01)**	0.58 (0.08)	**0.97 (0.02)**	**0.97 (0.01)**	0.97 (0.02)	0.58 (0.09)	0.97 (0.02)	**0.99 (0.02)**
$n = 100, p = 50$	0.67 (0.04)	0.60 (0.04)	0.80 (0.03)	**0.98 (0.02)**	0.60 (0.04)	0.54 (0.03)	0.61 (0.03)	**0.98 (0.01)**
$n = 300, p = 60$	**0.98 (0.01)**	0.32 (0.06)	**0.98 (0.01)**	**0.98 (0.01)**	0.96 (0.02)	0.64 (0.10)	0.96 (0.02)	**0.98 (0.01)**
$n = 300, p = 150$	0.58 (0.03)	0.70 (0.04)	0.76 (0.02)	**0.99 (0.01)**	0.56 (0.03)	0.66 (0.04)	0.79 (0.03)	**1.00 (0.00)**
$n = 500, p = 100$	**0.98 (0.01)**	0.32 (0.02)	0.95 (0.01)	**0.98 (0.01)**	0.96 (0.02)	0.62 (0.08)	0.98 (0.02)	**1.00 (0.00)**
$n = 500, p = 250$	0.50 (0.02)	0.73 (0.04)	0.81 (0.02)	**0.99 (0.01)**	0.48 (0.02)	0.68 (0.02)	0.92 (0.02)	**0.99 (0.01)**

TABLE V: Performance comparison in terms of the average efficiency in a sequential setting, for data and models described in Subsection V-B. The values are averaged from 20 independent replications (standard errors within 0.3).

Holdout	CV (10 fold)	CV (LOO)	AIC	BIC	SH	GTIC	GTIC (Sequential)
0.62	0.67	0.87	0.71	0.58	0.69	0.75	0.92

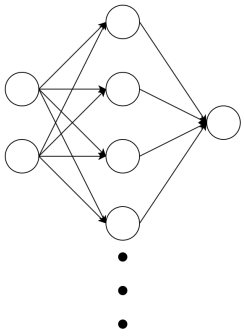


(a) Plot showing the loss of our predictor (GTIC) and cross validations at each sample size

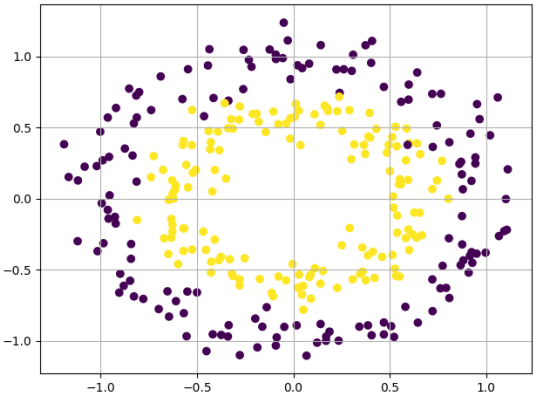


(b) Plot showing the computational costs.

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

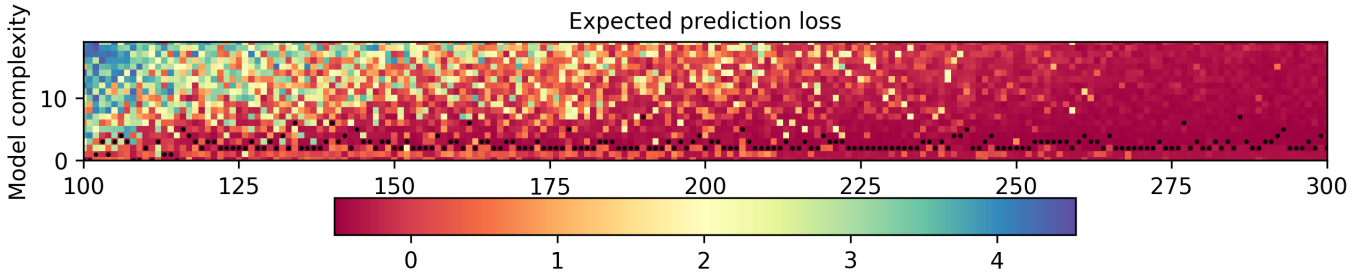
B. Neural Networks

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

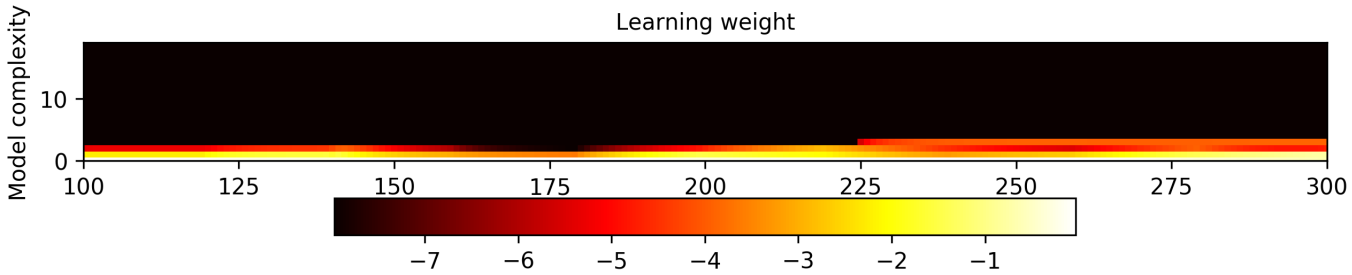
Data are generated from the following way. A set of two-dimensional data are uniformly sampled from two circles (with radius ratio 0.6), corrupted by independent Gaussian noise with mean 0 and variance 0.1 (generated from python package 'sklearn' dataset "make_circle"). The goal is to correctly classify 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 $\sqrt{t}/(\text{input dimension})$ 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

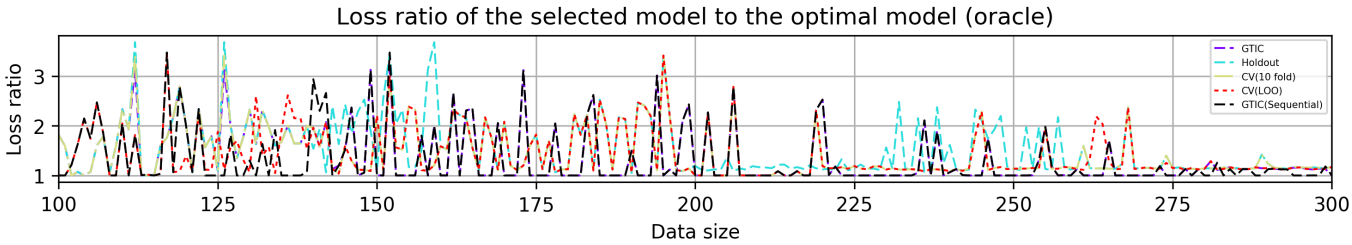


(a) Heat-map showing the prediction loss of estimated candidate models of each dimension (y-axis) at each sample size (x-axis), where the black dots indicate the model of optimal loss at each sample size.

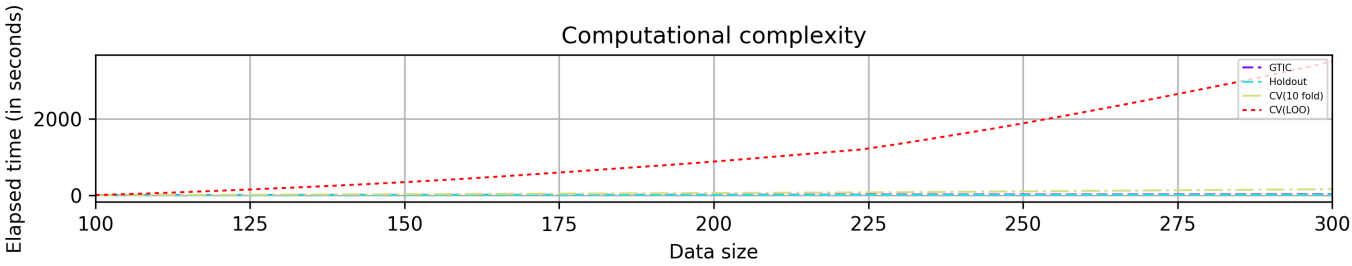


(b) Heat-map showing our predictive weights over the candidate models (y-axis) at each sample size (x-axis).

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

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. The results in Fig. 7b show that the weights of smaller models in the active set are large enough to prevent the model from expanding. As a result, we alleviate the tendency to choose the overfitting models even when their loss is relatively small. The average efficiency over all the time steps is reported in Table V, where we considered the use of various selection criteria in sequential settings (using Algo. 2).

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 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.

VI. Conclusion

In the framework of parametric models with possibly expanding model dimensions and model space, we studied a method 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. The proposed method, GTIC, is an extension of an information criterion by Takeuchi to more general loss functions. Our theoretical analysis of GTIC justifies the use of TIC for general mis-specified model classes, and extends some technical tools for classical analysis of AIC in linear models. Moreover, the proposed approach serves as an alternative of leave-one-out cross validation that is in general not accessible due to its computational burden. In the second part of the paper, we also proposed a sequential model expansion algorithm for reliable online prediction with low computation cost, based on our new graph-based expert tracking techniques. In summary, the proposed methodology is asymptotically optimal and practically useful, and it can be a promising competitor of cross-validation in both batch and online settings.

Appendix A
Proof of Theorem 1

We start with the following technical lemmas and additional definitions.

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

$$\sup_{\alpha\in\mathcal{A}_n} \|\hat{\boldsymbol{\theta}}_n[\alpha] - \boldsymbol{\theta}_n^*[\alpha]\| = o_p(1) \tag{26}$$

as $n \to \infty$.

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

$$\begin{aligned} &n^\tau E_*\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) - \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= n^\tau\{0 - E_*\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= n^\tau\{E_n\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha) - E_*\boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \end{aligned} \tag{27}$$

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

$$\begin{aligned} &E_*\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha) - \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\} \\ &= \nabla_{\boldsymbol{\theta}} E_*\{\boldsymbol{\psi}_n(\cdot, \tilde{\boldsymbol{\theta}}[\alpha]; \alpha)\}(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]) \\ &= V_n(\tilde{\boldsymbol{\theta}}[\alpha]; \alpha)(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]), \end{aligned} \tag{28}$$

where the exchangeability of integral and differentiation (in the second identity) is guaranteed by (8) and (9) 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 (28), the second equality is due to (27), and the third equality comes from Assumption 3. By the (13) in Assumption 6 and Assumption 4, $V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)$ 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

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{A}_n} n^\tau \|\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha]\| \\
&\leq \sup_{\alpha\in\mathcal{A}_n} \left\{ \|V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}\| \cdot \|n^\tau V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)(\boldsymbol{\theta}_n^*[\alpha] - \hat{\boldsymbol{\theta}}_n[\alpha])\| \right\} \\
&= O_p(1), \qquad (29)
\end{aligned}$$

which concludes the proof. ■

Before we proceed, we need the following definition.

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 $f_1 \leq f \leq f_2$. 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

$$I_{[\,]}(\delta, \mathcal{F}, L_2(P_*)) = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P_*))}d\varepsilon \qquad (30)$$

for $\delta > 0$.

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 [73] 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 $r_n \leq r$ 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

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

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

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

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

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

From (30), 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 $v_1 \geq \cdots \geq v_k \geq 1$ ($k \geq 2$), we have

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

Furthermore, it suffices to prove that

$$\log(kv_1) \leq \max\{2\log k,\ k\log v_1\}. \qquad (31)$$

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

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

Otherwise, $\log(kv_1) \le k \log v_1$, because $g : v \mapsto k\log v - \log(kv)$ is increasing on $v \ge 1$ 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 $\sup_{f\in\mathcal{F}} |f(z)| \le F(z) < \infty$ 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

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

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

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

$$d_n \sqrt{\log\{d_n \mathrm{card}(\mathcal{A}_n)\}}\delta_n \to 0, \tag{32}$$

and

(iii) The bounded moment condition:

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

(iv) The bounded class condition:

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

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,

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], i=1,\ldots,d_n} E_* f_{n,u,i}^2 &\le E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], i=1,\ldots,d_n} f_{n,u,i}^2 \\ &< \delta_n^2.\end{aligned} \tag{33}$$

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

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

From inequality (33) and Lemma 4, we have

$$\begin{aligned} & E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], 1\le i\le d_n[\alpha]} |\mathbb{G}_n f_{n,u,i}| \\ & \lesssim I_{[\ ]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*)) + \sqrt{n} E_*\big\{F_n \cdot 1_{F_n > \sqrt{n} a_n(\delta_n)}\big\} \\ & \le I_{[\ ]}(\delta_n, \tilde{\mathcal{F}}_n, L_2(P_*)) + \frac{1}{a_n(\delta_n)} E_* F_n^2, \end{aligned} \tag{34}$$

where the second inequality comes from the fact that

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

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

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

It follows that the right hand side of (34) 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

$$\begin{aligned} & E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]} \|\mathbb{G}_n \boldsymbol{f}_{n,u}\| \\ & \le E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha]} \sum_{i=1}^{d_n} |\mathbb{G}_n f_{n,u,i}| \\ & \le d_n E_* \sup_{\alpha\in\mathcal{A}_n, u\in U[\alpha], 1\le i\le d_n[\alpha]} |\mathbb{G}_n f_{n,u,i}| \\ & \le (A_1 + A_2)\big(1 + \delta_n^{-2}\ E_* F_n^2\big), \end{aligned} \tag{35}$$

where

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

Assumptions (ii), (iii), and (iv) guarantee that the right hand side of (35) 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

$$\sup_{\alpha\in\mathcal{A}_n} \|\mathbb{G}_n \boldsymbol{\psi}_n(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha) - \mathbb{G}_n \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}_n^*[\alpha]; \alpha)\| = o_p(1). \tag{36}$$

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 (36) 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 $F_n(\cdot) \stackrel{\Delta}{=} cn^{-\tau}\sup_{\alpha\in\mathcal{A}_n} m_n[\alpha](\cdot)$ as the envelop function for each $f_{n,u,i}(\cdot)$, and we have

$$E_* F_n^2 \leq C_1 \stackrel{\Delta}{=} c^2 n^{-2\tau} E_* m_n^2$$

Let

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

Because of (10) in Assumption 5, we have

$$C_2^2 C_1 = c^2 n^{-2\tau} d_n^2 \log\{d_n \text{card}(\mathcal{A}_n)\}\ E_* m_n^2 \to 0 \tag{37}$$

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

$$\delta_n C_2 \to 0, \quad \delta_n^{-2} C_1 \to 0,$$

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 $\alpha \in \mathcal{A}_n, 1 \leq i \leq d_n[\alpha]$ that

$$\begin{aligned}
&I_{[\ ]}(\delta_n, \mathcal{F}_{n,i}[\alpha], L_2(P_*)) \\
&\leq \int_0^{\delta_n} \left[\max\left\{0, d_n \log\left(\varepsilon^{-1} cn^{-\tau}\|m_n\|_{P_*}\right)\right\}\right]^{1/2} d\varepsilon \\
&= \int_0^{\min\{\delta_n, cn^{-\tau}\|m_n\|_{P_*}\}} \left[d_n \log\left(\varepsilon^{-1} cn^{-\tau}\|m_n\|_{P_*}\right)\right]^{1/2} d\varepsilon
\end{aligned} \tag{38}$$

Because condition (10) 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 (38) 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 (38) 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

$$\begin{aligned}
&\sqrt{d_n^{3/2}\ \text{card}(\mathcal{A}_n)} \times \sup_{\alpha\in\mathcal{A}_n, 1\leq i\leq d_n[\alpha]} I_{[\ ]}(\delta_n, \mathcal{F}_{n,i}[\alpha], L_2(P_*)) \\
&\leq (1-\rho)^{-1} c^{1-\rho} d_n \sqrt{\text{card}(\mathcal{A}_n)} (n^{-\tau}\|m_n\|_{P_*})^{1-\rho} \\
&= (1-\rho)^{-1} c^{1-\rho} \left(d_n^{\gamma}\ \text{card}(\mathcal{A}_n)^{\gamma/2}\ n^{-\tau}\|m_n\|_{P_*}\right)^{1-\rho} \\
&\to 0
\end{aligned} \tag{39}$$

where the last limit is due to (10) 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

$$\begin{aligned}
&\sqrt{n}E_*\{\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)\}\\
&=\sqrt{n}\{E_*\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-0\}\\
&=\sqrt{n}\{E_*\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)-E_n\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\}\\
&=-\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\
&=-\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\boldsymbol{\nu}_n
\end{aligned}\tag{40}$$

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

$$\begin{aligned}
&E_*\{\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)-\boldsymbol{\psi}_n(\cdot,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\}\\
&=\nabla_{\boldsymbol{\theta}}E_*\{\boldsymbol{\psi}_n(\cdot,\tilde{\boldsymbol{\theta}}[\alpha];\alpha)\}(\boldsymbol{\theta}_n^*[\alpha]-\hat{\boldsymbol{\theta}}_n[\alpha])\\
&=V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)(\boldsymbol{\theta}_n^*[\alpha]-\hat{\boldsymbol{\theta}}_n[\alpha]),
\end{aligned}\tag{41}$$

where the exchangeability of integral and differentiation (in the second identity) is guaranteed by (8) and (9) in Assumption 5. Multiplying the matrix $\sqrt{n}V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}$ to both sides of (41) and using equality (40), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])=-V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}\mathbb{G}_n\boldsymbol{\psi}_n(\cdot,\boldsymbol{\theta}_n^*[\alpha];\alpha)+V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)^{-1}\boldsymbol{\nu}_n.\tag{42}$$

We conclude the proof by applying Assumption 4 (with the constant c_2) and (12) in Assumption 6 to equality (42). ■

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

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{\mathcal{L}_n^c[\alpha]-\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]}\right|\to_p 0,\tag{43}$$

and

$$\sup_{\alpha\in\mathcal{A}_n}\frac{\mathcal{L}_n[\alpha]}{\mathcal{R}_n[\alpha]}\to_p 1.\tag{44}$$

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

$$\begin{aligned}
\mathcal{L}_n[\alpha]&=E_*l_n(\boldsymbol{z},\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\
&=E_*l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\theta}}E_*l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+\frac{1}{2}\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|^2_{\nabla^2_{\boldsymbol{\theta}}E_*l_n(\boldsymbol{z},\tilde{\boldsymbol{\theta}}[\alpha];\alpha)}\\
&=E_*l_n(\boldsymbol{z},\boldsymbol{\theta}_n^*[\alpha];\alpha)+\frac{1}{2}\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\|^2_{V_n(\tilde{\boldsymbol{\theta}}[\alpha];\alpha)}
\end{aligned}\tag{45}$$

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 (9) in Assumption 5, and the consistency of $\hat{\boldsymbol{\theta}}_n[\alpha]$. We note that by Assumption 4, the equality (45) further implies (4) presented in our introduction.

Similarly, we have

$$\begin{aligned}
\hat{\mathcal{L}}_n[\alpha]&=\frac{1}{n}\sum_{i=1}^n l_n(\boldsymbol{z}_i,\hat{\boldsymbol{\theta}}_n[\alpha];\alpha)\\
&=\frac{1}{n}\sum_{i=1}^n l_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*[\alpha];\alpha)+(\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha])^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^n\boldsymbol{\psi}_n(\boldsymbol{z}_i,\boldsymbol{\theta}_n^*[\alpha];\alpha)+\frac{1}{2}\left\|\hat{\boldsymbol{\theta}}_n[\alpha]-\boldsymbol{\theta}_n^*[\alpha]\right\|^2_{\hat{V}_n(\tilde{\tilde{\boldsymbol{\theta}}}[\alpha])}.
\end{aligned}\tag{46}$$

From identities (45) and (46), we may write

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

where we define

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

In view of (43), it suffices to prove that

$$\sup_{\alpha \in \mathcal{A}_n} \frac{|A_k[\alpha]|}{\mathcal{R}_n[\alpha]} \to_p 0 \tag{47}$$

as $n \to \infty$ for $k = 3, 4, 5, 6$, and the limit (44).

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 (14), (47) with $k = 3$ can be proved.

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

$$\begin{aligned}&P_*\bigg(\sup_{\alpha \in \mathcal{A}_n} \frac{|A_4[\alpha]|}{\mathcal{R}_n[\alpha]} > \delta \bigg) \\ &\le \sum_{\alpha \in \mathcal{A}_n} P_*\bigg(\frac{|A_4[\alpha]|}{\mathcal{R}_n[\alpha]} > \delta \bigg) \\ &\le \sum_{\alpha \in \mathcal{A}_n} \frac{E_* \{ l_n(\boldsymbol{z}_1, \boldsymbol{\theta}_n^*; \alpha) - E_* l_n(\boldsymbol{z}, \boldsymbol{\theta}_n^*[\alpha]; \alpha) \}^{2m_1}}{\delta^{2m_1} n^{2m_1} \mathcal{R}_n[\alpha]^{2m_1}}.\end{aligned} \tag{48}$$

Thus, given assumption (16), (47) 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 (47) with $k = 5$, we only need to show that

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

$$\sup_{\alpha \in \mathcal{A}_n} \frac{1}{n \mathcal{R}_n[\alpha]} tr\{(V[\alpha]^{-1} - \hat{V}[\alpha]^{-1}) \hat{J}[\alpha]\} \to_p 0. \tag{50}$$

We only prove (49), and then (50) 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, (49) could be rewritten as

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

where the first equality is due to (11) in Assumption 6, the second equality is due to Assumption 4, and the last equality is guaranteed by assumption (15).

Next, we prove (47) 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

$$\begin{aligned}
A_7[\alpha] &= \frac{\|\boldsymbol{w}_n[\alpha]\|^2_{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1}} - tr\big\{V_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)^{-1} J_n(\boldsymbol{\theta}_n^*[\alpha];\alpha)\big\}}{n\mathcal{R}_n[\alpha]}, \\
A_8[\alpha] &= \frac{\|\boldsymbol{w}_n[\alpha]\|^2_{\nu_{1,n}[\alpha]}}{n\mathcal{R}_n[\alpha]}, \quad A_9[\alpha] = \frac{\boldsymbol{\nu}_{2,n}[\alpha]^{\mathrm{T}}\boldsymbol{w}_n[\alpha]}{n\mathcal{R}_n[\alpha]}.
\end{aligned}$$

Using assumption (17) and similar arguments as in (48), 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 (18) guarantees that $\sup_{\alpha\in\mathcal{A}_n} A_8[\alpha] \to_p 0$. Cauchy inequality and assumption (18) also imply that

$$\sup_{\alpha\in\mathcal{A}_n} |A_9[\alpha]| \le \sup_{\alpha\in\mathcal{A}_n} \frac{\|\boldsymbol{\nu}_{2,n}[\alpha]\| \times \|\boldsymbol{w}_n[\alpha]\|}{n\mathcal{R}_n[\alpha]} \to_p 0. \tag{51}$$

Finally, we prove (44). From (45) 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

$$\max\left\{2w, \frac{\zeta}{2}\right\} < \tau \leq \frac{1}{2} - w. \tag{52}$$

For Assumption 2,

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

for all $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*[\alpha]\| \geq \varepsilon$. 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 $n^{-1}\text{Var}\{(Y - \boldsymbol{\theta}^{\text{T}}\boldsymbol{X}[\alpha])^2\} = O(d_n^2/n) = o(1)$ 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 $n^{-1}\text{Var}\{(Y - \boldsymbol{\theta}^{\text{T}}X[\alpha])X[\alpha]^{\text{T}}\}$. Let $A = E_*\{(Y - \boldsymbol{\theta}^{\text{T}}X[\alpha])^2X[\alpha]X[\alpha]^{\text{T}}\}$. Let $\|\cdot\|_F$ denote the Frobenius norm. Since

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

uniformly in $\boldsymbol{\theta} \in \mathcal{H}_n[\alpha]$ and $\alpha \in \mathcal{A}_n$. Thus any τ satisfying (52) suffices.

For Assumption 5,

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

So $m_n[\alpha] = cd_n$ suffices. This together with the condition $2w < \tau$ implies (10).

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, (14) is implied by $\zeta/2 < \tau$, and (15) is implied by $\zeta < 1 - w$. Since

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

(16) holds under $m_1 = 1$ and $\zeta < 1 - w$. Similar calculations as before show that (17) and (18) 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 [71, 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

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

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

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

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

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

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 [71, 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 obtain

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

■

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] S. Portnoy *et al.*, "Asymptotic behavior of m estimators of p regression parameters when p2/n is large. i. consistency," *Ann. Stat.*, vol. 12, no. 4, pp. 1298–1309, 1984.

[8] S. Portnoy, "On the central limit theorem in r p when p goes to infinity," *Probab. Theory Relat. Fields*, vol. 73, no. 4, pp. 571–583, 1986.

[9] ——, "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity," *Ann. Stat.*, pp. 356–366, 1988.

[10] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probability theory and related fields*, vol. 113, no. 3, pp. 301–413, 1999.

[11] P. Massart, *Concentration inequalities and model selection*. Springer, 2003, vol. 6.

[12] G. Claeskens, N. L. Hjort *et al.*, "Model selection and model averaging," *Cambridge Books*, 2008.

[13] S. Arlot, A. Celisse *et al.*, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[14] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, 2018.

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

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

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

[18] 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.

[19] 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.

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

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

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

[23] 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.

[24] 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.

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

[26] O. Lepskii, "On a problem of adaptive estimation in gaussian white noise," *Theory Probab. Its Appl.*, vol. 35, no. 3, pp. 454–466, 1991.

[27] A. Goldenshluger, O. Lepski *et al.*, "Universal pointwise selection rule in multivariate function estimation," *Bernoulli*, vol. 14, no. 4, pp. 1150–1190, 2008.

[28] A. Goldenshluger and O. Lepski, "General selection rule from a family of linear estimators," *Theory Probab. Its Appl.*, vol. 57, no. 2, pp. 209–226, 2013.

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

[30] M. Dixon and T. Ward, "Takeuchi's information criteria as a form of regularization," *arXiv preprint arXiv:1803.04947*, 2018.

[31] L. Birgé and P. Massart, "Gaussian model selection," *J. Eur. Math. Soc.*, vol. 3, no. 3, pp. 203–268, 2001.

[32] ——, "Minimal penalties for gaussian model selection," *Probab. Theory Relat. Fields*, vol. 138, no. 1-2, pp. 33–73, 2007.

[33] J.-P. Baudry, C. Maugis, and B. Michel, "Slope heuristics: overview and implementation," *Stat. Comput.*, vol. 22, no. 2, pp. 455–470, 2012.

[34] S. Arlot, "Minimal penalties and the slope heuristics: a survey," *arXiv preprint arXiv:1901.07277*, 2019.

[35] ——, "Rejoinder on: Minimal penalties and the slope heuristics: a survey," *arXiv preprint arXiv:1909.13499*, 2019.

[36] S. Boucheron, G. Lugosi, and P. Massart, "Concentration inequalities using the entropy method," *Ann. Probab.*, vol. 31, no. 3, pp. 1583–1614, 2003.

[37] S. Boucheron, G. Lugosi, and O. Bousquet, "Concentration inequalities," in *Summer School on Machine Learning*. Springer, 2003, pp. 208–240.

[38] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

[40] R. E. Gaunt, A. M. Pickett, G. Reinert *et al.*, "Chi-square approximation by stein's method with application to pearson's statistic," *Ann. Appl. Probab*, vol. 27, no. 2, pp. 720–756, 2017.

[41] S. Arlot and P. Massart, "Data-driven calibration of penalties for least-squares regression," *J. Mach. Learn. Res.*, vol. 10, no. Feb, pp. 245–279, 2009.

[42] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Stat. Soc. B*, vol. 36, no. 2, pp. 111–133, 1974.

[43] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

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

[45] 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.

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

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

[48] 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.

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

[50] J. Zhang, J. Ding, and Y. Yang, "A binary regression adaptive goodness-of-fit test," *arXiv preprint arxiv:1911.03063*, 2019.

[51] 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.

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

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

[54] ——, "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.

[55] 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.

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

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

[58] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1926–1940, 1998.

[59] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[60] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Summer School on Machine Learning*. Springer, 2003, pp. 169–207.

[61] M. S. Pinsker, "Optimal filtering of square-integrable signals in gaussian noise," *Probl. Peredachi Inf.*, vol. 16, no. 2, pp. 52–68, 1980.

[62] M. Pinsker and S. Efroimovich, "Learning algorithm for nonparametric filtering," *Autom. Remote Control*, vol. 45, no. 11, pp. 1434–1440, 1984.

[63] M. Nussbaum, "Minimax risk: Pinsker bound," *Encyclopedia of Statistical Sciences*, vol. 3, pp. 451–460, 1999.

[64] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[65] C.-K. Ing, "Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series," *Ann. Statist.*, vol. 35, no. 3, pp. 1238–1277, 2007.

[66] T. v. Erven, P. Grünwald, and S. De Rooij, "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma," *J. R. Stat. Soc. Ser. B.*, vol. 74, no. 3, pp. 361–417, 2012.

[67] R. C. Bradley, "Basic properties of strong mixing conditions," in *Dependence in probability and statistics*. Springer, 1986, pp. 165–192.

[68] T. S. Ferguson, *A course in large sample theory*. Routledge, 2017.

[69] K.-C. Li, "Asymptotic optimality for Cp, Cl, cross-validation and generalized cross-validation: discrete index set," *Ann. Stat.*, pp. 958–975, 1987.

[70] G. Stoltz and G. Lugosi, "Internal regret in on-line portfolio selection," *Mach. Learn.*, vol. 59, no. 1-2, pp. 125–159, 2005.

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

[72] G. Stoltz and G. Lugosi, "Learning correlated equilibria in games with compact sets of strategies," *Econ. Behav.*, vol. 59, no. 1, pp. 187–208, 2007.

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