Asymptotically Optimal Prediction for Time-Varying Data Generating Processes

Jie Ding, Jiawei Zhou, and Vahid Tarokh

***Abstract*—We develop a methodology (referred to as kinetic prediction) for predicting time series undergoing unknown abrupt changes in their data generating distributions. Based on Kolmogorov-Tikhomirov's ε-entropy, we propose a concept called ε-predictability that quantifies the size of a model class (which can be parametric or nonparametric) and the maximal number of structural changes that guarantee the achievability of asymptotically optimal prediction. Moreover, for parametric distribution families, we extend the aforementioned kinetic prediction with discretized function spaces to its counterpart with continuous function spaces, and propose a sequential Monte Carlo based implementation. We also extend our methodology for predicting smoothly varying data generating distributions. Under reasonable assumptions, we prove that the average predictive performance converges almost surely to the oracle bound, which corresponds to the case that the data generating distributions are known in advance. The results also shed some light on the so called "prediction-inference dilemma". Various examples and numerical results are provided to demonstrate the wide applicability of our methodology.**

***Index Terms*—Change points; Kinetic prediction; Kolmogorov-Tikhomirov ε-entropy; Optimal prediction; Sequential Monte-Carlo; Smooth variations; Time series; Online tracking.**

I. Introduction

In data analysis, sequentially observed data (either in time or space) usually exhibit occasional but abrupt changes in mean, variance, or correlation. In the presence of time-varying data generating processes, using a fixed (parametric or nonparametric) model may not be adequate for prediction. This motivates the use of time-varying models, which are becoming increasingly important for online data analysis. Roughly speaking, we could categorize the changes in time-varying data generating processes into two kinds: "abrupt changes" and "smooth variations", where the former refers to a serious sudden change in data generating distribution, and the latter refers to successive changes of small magnitude per time step. Both kinds of changes are commonly observed in practice. In order to put our ideas in focus, consider a scalar time series $\{Y_t\}_{t=1,2,...}$, each Y_t being an independent Gaussian random variable with mean θ_t and unit variance. Suppose that we are to predict (the distribution of) Y_{T+1} at each time T. Without recognizing potential change points, one may use the maximum likelihood estimator (MLE) $\hat{\theta} = \sum_{t=1}^{T} Y_t/T$ for prediction. The predicted distribution of Y_{T+1} based on such an estimator is asymptotically close to the population distribution given that θ_t does not change. However, suppose for instance that $\theta_1 = \cdots = \theta_{T/2} \neq \theta_{T/2+1} = \cdots = \theta_T$ with T being an even number. Without recognizing an abrupt change in the data-generating model at time $T/2+1$, the MLE is close to $(\theta_{T/2} + \theta_{T/2+1})/2$ for large T, which can produce a large bias in prediction. Thus, it is important to first determine the change point (by applying a change detection procedure), and then apply inference to each segment where the underlying data-generating process does not vary. This has motivated a large body of the work on (both online and offline) change point detection.

Classical change point analysis can be roughly categorized into two scenarios, namely online detection and offline detection. Deeply rooted in sequential analysis [1], [2], online change detection is often studied in terms of trade-offs between the probability of false alarm and detection delay, from different perspectives such as the cumulative sum (CUSUM) [3], [4], Shiryaev-Roberts procedure [5], [6], minimax detection [7], [8], high dimensional detection [9] (see, e.g., [10] for more references). Various tests have been developed for tracking changes in time series statistics including the mean [11], the variance [12], the autocovariance function [11], [13], and the spectrum [14]. Offline change detection aims to discover multiple change points from a given sequence of data, typically by minimizing the within-segment sum of loss plus penalties (from frequentist perspective), or the negative model evidence (from Bayesian perspective). Large sample analysis focuses on the consistent selection of the total number of changes (as sequence length goes to infinity) [15]–[18]. A more comprehensive literature review on offline analysis can be found in [18]. In both online or offline cases, state-of-the-art methods and

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

This work is supported by the Defense Advanced Research Projects Agency (DARPA) grant numbers N66001-15-C-4028 and W911NF-16-1-0561. Part of this work was presented at the GlobalSIP 2017 conference.(Corresponding author: Jie Ding, e-mail: djrlthu@gmail.com)

analysis may not be directly applicable to the situations where the number of change points grows as a function of sample size, and where consecutive change points occur closely (also referred to as the transient changes). Also, for handling multiple changes with unknown after-change distributions, a practical algorithm with asymptotic optimality remains an important open challenge. These motivated our present work.

In this paper, we develop a new framework for online change point analysis that can achieve asymptotically optimal prediction under various types of changes (including abrupt changes and smooth variations) and a wide class of evaluation metrics. We address change point analysis from a perspective that is fundamentally different from the existing literature. In particular, we assume a deterministic stopping time (similar to the classical offline detection), while we perform analysis online (similar to the classical online detection). Moreover, our evaluation metric for online analysis is to perform asymptotically optimal prediction, instead of detection delay or accuracy of change detection. More discussions are given in the rest of the introduction. For example, an investment analyst may be interested in online optimizing the predictive performance of his/her modeling of asset prices, which are likely to undergo market changes, while stopping at a prescribed time (e.g. at expiration dates of futures contracts).

We refer the set of potential density functions that may generate the data in various epochs of time as model class. We call a model *nonparametric* (resp. *parametric*) whenever the elements of the model class are not (resp. are) parameterized by a finite number of unknown parameters. We next make the following simple observation. Consider a sequence of independent Gaussian random variables with changes in mean, occurring at time $t = 2, 4, \ldots, 2^k$ (where $2^{k-1} < T \leq 2^{k+1}$). For offline change point analysis, the last few changes (with sufficient samples in between) may be easily detected, but it is unclear whether the change points at the beginning can be accurately discovered. Nevertheless, in view of short-period changes, it is not clear that this lack of detection accuracy results in a significant loss in predictive performance. Similarly, in online change point analysis, it is not clear if spurious changes or undetected changes have significant impacts on the average predictive performance. We therefore ask the following critical question:

(Q1) Is it possible to achieve optimal prediction without inferring the number and locations of abrupt change points first (for either parametric or nonparametric model classes)?

In other words, is *prediction without inference* possible? In this paper, we provide a positive answer to the above question by proposing a new framework for statistical prediction of time series. Our results are appealing since a small change that tends to be overlooked may indicate its insignificance or non-persistency. Moreover, we present a methodology that achieves asymptotic optimality (to be rigorously defined) for a wide range of nonparametric and parametric model classes, and for a wide range of abrupt change models. Interestingly, we also show that there exists no statistical procedure that is consistent in selecting the number of change points, under the same set of assumptions that guarantee optimal prediction. It is worth noting that the ideas on the inference-prediction dilemma are in parallel to those discovered in the literature of high dimensional regression analysis [19], and in information criteria for model selection [20]–[23].

Additionally, we are interested in prediction under smoothly varying data-generating processes. These processes often occur in practice, such as in environmental science where temperature and humidity vary smoothly [24], in finance where regression models with time-varying coefficients are more favored [25], or in control where the trajectory of a target can be cast as smoothly-varying parameters [26]. There exists a large literature on modeling smooth changes by spectral arguments, including evolutionary spectra [27], [28], locally stationary time series [29], or other time/frequency analysis such as wavelet approach [30].

In time-domain formulations, a systematic way to characterize time-varying models is through the parametric state space models [31], where hidden states may be interpreted as parameters. For linear state space models with known parameters, state-of-the-art inference employs the Kalman filtering algorithm [26] (resp. the Bellman-Ford/Viterbi algorithm [32]) for continuous (resp. discrete) state space. For nonlinear state space models with known or unknown parameters, Bayesian inference implemented with Sequential Monte Carlo (SMC) and Markov chain Monte Carlo (MCMC) may be applied to numerically estimate the full posterior of the states [33]. Other methodology to model smooth variations is to apply local least-squares method or its variants to estimate varying coefficient regression models [34], autoregressive models [35], Cox models [36], nonlinear additive models [37], etc. For example, to model small drifts of nonlinear time series models over time, a computationally efficient method is to employ an online adaptive filtering on the coefficients of spline basis functions [37]. Nevertheless, the above methods are not directly applicable to some practical situations where parameters follow a mixture of abrupt changes and slow variations. For example, data are generated from $Y_t = \theta_t^* + e_t$, where e_t are independent standard Gaussian random variables, the squared loss (or score in

our notation) is used to evaluate predictive performance, and θ_t can be well approximated by an epoch of linear trend, followed by a sinusoid trend, etc. Suppose we constructed a predictor $\hat{\theta}_t$ for Y_t at each time t, then our average score is given by $S(Y_1,\ldots,Y_T) = T^{-1}\sum_{t=1}^{T}(Y_t-\hat{\theta}_t)^2 = T^{-1}\sum_{t=1}^{T}(\theta_t^*-\hat{\theta}_t+e_t)^2$. The *oracle score* (when the true θ_t^*'s are known in the first place) is given by $S^*(Y_1,\ldots,Y_T) = T^{-1}\sum_{t=1}^{T}e_t^2$. Is it possible that $S(Y_1,\ldots,Y_T) - S^*(Y_1,\ldots,Y_T) \to 0$ (either in probability or almost surely) as $T\to\infty$? We then raise the following question. To address smooth variations, we focus on parametric model classes for simplicity of presentation.

(Q2) Is it possible to achieve optimal prediction (oracle score) when the underlying parameters (for a parametric model class) undergo slow variations as well as abrupt changes?

Clearly, there is no chance to achieve the above convergence for θ_t^*'s that vary arbitrarily at each time instance. Nevertheless, we will show that if the smooth variations of θ_t^*'s can be approximated by locally deterministic trends (which can be nonlinear), it is possible to approach the oracle performance. This is intuitively reasonable. For instance if θ_t^*'s follow a linear trend in a time epoch of length ΔT, then the dimension of unknown parameters reduces to only two (i.e. slope and intercept). This intuition can be extended to polynomial or other nonlinear trends.

Contributions and outline: In light of above discussions and challenges, we are motivated to come up with a general solution that can 1) predict the outcomes of both abrupt and smooth time-varying data generating processes, and 2) achieve optimal prediction score by requiring only mild assumptions on the laws governing the changes. To this end, we propose a methodology referred to as "kinetic prediction" to predict time series data, for both nonparametric and parametric model classes. And under mild assumptions, we prove that the average predictive performance asymptotically approaches the theoretical limit (oracle bound). The outline of this paper is given next. In Section II, we establish the notation and review the mathematical background. We also provide mathematical formulation of the theoretical limits of prediction (oracle bound), and the scoring rules that we will use for the quantification of predictive performance. In Section III, we consider optimal prediction under unknown abrupt changes in either nonparametric or parametric models. We propose a concept called "ε-predictability", to describe the size of a model class that guarantees the achievability of optimal predictive performance. To this end, we apply an ε-net to the function (resp. parameter) space of the nonparametric (resp. parametric) models, and sequentially adjust predictive weights of the elements of the net. This offers a predictive prior for the arrival of the next observation. We will also discuss the aforementioned "inference-prediction dilemma", and the related literature in non-statistical online learning. In Section IV, we address the optimal prediction from another perspective. Instead of discretizing the function space, we propose kinetic prediction that uses a distribution over a continuous function space and sequentially update it. The method in its form resembles classical Bayes' rule. However, they are not the same, and we shall mention some of their (dis)similarities. We then propose a Monte-Carlo technique for efficient computation of predictors in a sequential manner. The new algorithm–which is alternative to ε-net based methods–can more efficiently search and sample from the parameter space.

In Section V, we extend the solutions in Section III to handle smooth variations in addition to abrupt changes. We only require that the smooth variations (of parameters) between each pair of abrupt changes approximately follow a locally deterministic evolution path. For simplicity, we will focus on parametric models in Section V. The generalization to non-parametric models is straightforward.

II. Background and Problem Setup

We first introduce some notation, followed by a mathematical description of the problems to be addressed.

A. Notation

Throughout the paper, we assume that data is sequentially generated from some unknown probabilistic distributions. Let $Y_t \in \mathcal{Y}$ be a time series, where Y_t only depends on $Y_{1:t-1}$. The data generating distribution densities of Y_t is denoted by $g_t(\cdot \mid Y_{1:t-1})$. We may also write this as g_t when there are no ambiguities. For $t=1$, $g_1(\cdot \mid Y_{1:0})$ is simply interpreted as a density $g_1(\cdot)$ of Y_1.

We use $g_t^*, \mathbb{E}_*(\cdot)$ to respectively denote the true generating density of Y_t at time t, and the expectation operator with respect to the true data generating process. We will assume that the density g_t at each time belongs to a subset $\mathcal{G}$ of a metric space (of functions) $(\mathscr{G}, d_{\mathscr{G}})$, where $\mathscr{G}$ is a set of density functions. The diameter of a set $U\subseteq\mathscr{G}$ is defined by $\sup_{f,\tilde{f}\in U} d_{\mathscr{G}}(f,\tilde{f})$. Let

$$H: q \mapsto -q\log q - (1-q)\log(1-q)$$

for $q\in(0,1)$ and $H(0)=H(1)=0$ denote the binary entropy function. Let $\mathrm{card}(A)$ denote the cardinality of a finite set A.

Throughout the paper, random variables and their realizations are often represented by upper-case and lower-case letters, respectively. In this work, unless otherwise specified, all the limits are taken by allowing $T\to\infty$, where T denotes the data size. We let $a.s.$ denote "almost

surely", and $\mathbb{N}$ denote the set of all positive integers. Any vector $[Y_a, \ldots, Y_b]^{\mathrm{T}}$ may be compactly re-written as $Y_{a:b}$. Let $\lfloor a \rfloor$ (resp. $\lceil a \rceil$) denote the largest (resp. smallest) integer that is no larger (resp. smaller) than a. We may denote a general (resp. Gaussian) distribution with mean μ and variance σ^2 by $[\mu, \sigma^2]$ (resp. $\mathcal{N}(\mu, \sigma^2)$).

Suppose that $\hat{g}_t$ is our predicted density at each time t using only the observations of $Y_1, \ldots, Y_{t-1}$ (we consider a point estimator for now). To evaluate the predictive performance, we need to specify a scoring function (also called a statistical scoring rule [38]–[40]) $s : \mathscr{G} \times \mathcal{Y} \to \mathbb{R}$. A common practice is to use the logarithmic score (also referred to as the negative log-likelihood), $s(g, y) = -\log g(y)$. There are many other "reasonable" scores discussed in the next subsection. With some abuse of notation, we shall sometimes refer to the parametric density g_θ by parameter θ, and to $s(g_\theta, y)$ as $s(\theta, y)$ whenever there are no ambiguities.

B. Background

The optimal sequential predictive performance for time steps $t = 1, 2, \cdots, T$ is defined to be the infimum of the composite score

$$S(Y_1, \ldots, Y_T) = \sum_{t=1}^{T} s(\hat{g}_t(\cdot \mid Y_{1:t-1}), Y_t) \tag{1}$$

over all the choices $\hat{g}_t(\cdot \mid Y_{1:t-1})$. We point out that the notation of $\hat{g}_t(\cdot \mid Y_{1:t-1})$ in (1) is to emphasize the potential dependency of $\hat{g}_t$ on the past observations. We have used upper-case S to denote the composite (or cumulative) score.

Suppose that a genie presents us with the true g_t^*, then the composite score with genie's aid is

$$S^*(Y_1, \ldots, Y_T) = \sum_{t=1}^{T} s(g_t^*(\cdot \mid Y_{1:t-1}), Y_t).$$

We may write $g_t(\cdot \mid Y_{1:t-1})$ as g_t (for either $g_t = \hat{g}_t$ or $g_t = g_t^*$) for brevity when there are no ambiguities.

For a wide range of proper scoring rules (to be defined later), we shall prove under mild conditions that

$$\liminf_{T\to\infty} \frac{1}{T}\big\{S(Y_1, \ldots, Y_T) - S^*(Y_1, \ldots, Y_T)\big\} \geq 0, \quad a.s. \tag{2}$$

In other words, the total score from any inference procedure cannot be better than the genie-aided score in the above sense. A proof of this observation will be given in Subsection II-C.

It is common to assume that all the data is generated by the same data generating distribution g^* (i.e. $g_t^* \equiv g^*$), and perform the estimation of g^* in a statistical inference task. However, in many practical applications, both abrupt changes and smooth changes in the true density g_t^* exist. Moreover, the number of abrupt changes and their types (e.g. the slope of a linear drift) are unknown (in advance). It is clear that if g_t^* is time-varying in an arbitrary way at each time step, it is impossible to achieve any asymptotic optimality guarantees. But if the changes are restricted to some extent, as we briefly discussed in the introduction, it is possible to design a more flexible method. In this light, we propose a methodology that achieves optimal asymptotic prediction performance i.e.

$$\frac{1}{T}\big\{S(Y_1, \ldots, Y_T) - S^*(Y_1, \ldots, Y_T)\big\} \to_{a.s.} 0, \tag{3}$$

or equivalently (in view of (2))

$$\limsup_{T\to\infty} \frac{1}{T}\big\{S(Y_1, \ldots, Y_T) - S^*(Y_1, \ldots, Y_T)\big\} \leq 0, \quad a.s. \tag{4}$$

Before we proceed to the next subsection, we provide some examples for illustration of the main idea. These examples will be revisited later in remarks and numerical experiments.

Example 1 (Gaussian with one unknown abrupt change in mean). *Recall the parametric example in the introduction, where $Y_t \sim \mathcal{N}(\theta_t^*, 1)$, $\theta_t^* \in \Theta$ (a compact space), and $s(g_\theta, y) = \frac{1}{2}(y-\theta)^2 + c$ is the logarithmic score (with c being a constant). In this case, our goal (Equation (3)) discussed above can be cast as*

$$\frac{1}{T}\left\{\sum_{t=1}^{T}(\theta_t^* - \hat{\theta}_t + e_t)^2 - \sum_{t=1}^{T} e_t^2\right\} \to_{a.s.} 0. \tag{5}$$

where e_t is the random Gaussian noise at time t, distributed as $\mathcal{N}(0,1)$. In practice, θ_t^* usually dynamically varies with time. As a simple example, let us assume only one abrupt change point, in the sense that $\theta_t^* = \theta_1$ for $1 \leq t \leq T/2$, and $\theta_t^* = \theta_2$ for $T/2 < t \leq T$, and $\theta_1 \neq \theta_2$. For illustration, Fig. 1 shows the case of $\theta_1 = 5$ and $\theta_2 = 6$, and a realization of observations for $T = 500$ with unit variance. Is it possible to construct an estimator $\hat{\theta}_t$ using $Y_{1:t-1}$ so that (5) can be achieved, without knowing the location of the change point in advance?

If a data analyst routinely assumes *one* unknown parameter $\theta_t^* = \theta^*$, and estimates it at each time using maximum likelihood estimator $\hat{\theta}_t = \sum_{i=1}^{t} Y_i/t$, the left term in (5) is almost surely lower bounded by a positive constant. Thus the optimality cannot be achieved. This is rigorously summarized in Proposition 1. We note that the results can be largely generalized to other cases with different segment lengths and multiple abrupt changes. Moreover, similar situations recur if classical Bayesian predictive estimation is used instead of MLE, for instance $\hat{\theta}_t \triangleq E(\theta | Y_1, \ldots, Y_{t-1})$. We do not pursue the details here.

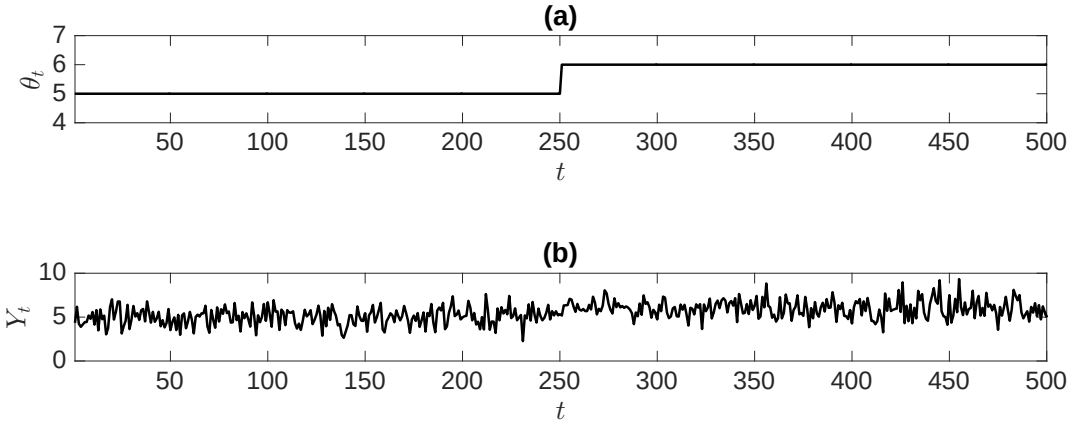

Fig. 1. Example 1: independent Gaussian random variables with an abrupt change in their means.

Proposition 1. *For Example 1, if we use the maximum likelihood estimator (MLE) for $\hat{\theta}_t$, we have*

$$\liminf_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T}(\theta_t^* - \hat{\theta}_t + e_t)^2 - \sum_{t=1}^{T} e_t^2\right\} > 0, \quad a.s.$$

Example 2 (Gaussian with many unknown abrupt changes in mean). *A less trivial version of Example 1 can be constructed when the number of change points in $\{\theta_t^*\}_{t=1}^T$ grows as a function of T. Specifically, we generated $\lfloor T^{(1/3)} \rfloor$ number of change points, whose locations are randomly generated from $1, \ldots, T$ with $T = 500$ according to the uniform distribution without replacement, and the Gaussian means in each segment are uniformly generated from $[5, 6]$. For illustration, a realization of time-varying means and the corresponding observations with unit variance are plotted in Fig. 2.*

To sequentially predict, classical change detection methods along with parameter inference in the latest detected segment can adequately handle the situation of Example 1. But they may fail to be applicable to the current setting (as we shall show in numerical experiments). Again, we would like to achieve the oracle stated by Equation (5) in this setting.

Example 3 (Unknown distributions with unknown abrupt changes). *Suppose that Y_t at different t's are independent random variables on $[0, 1]$, with true density functions in the form of*

$$g_t^*(y) = \begin{cases} 3.25 - (c_t - y) & \text{if } c_t - 0.25 \le y < c_t \\ 3.25 - (y - c_t) & \text{if } c_t \le y < c_t + 0.25 \\ 0.25 & \text{otherwise} \end{cases} \tag{6}$$

where $c_t \in [0.25, 0.75]$. In other words, the true densities are in the form of a constant plus a triangle centered at c_t. Suppose that the sequence $c_1, \ldots, c_T$ is piecewise constant, with abrupt changes at $\lfloor T/2^k \rfloor, k = 1, \ldots, \lfloor \log_2(T) \rfloor$, with values in each segment alternating between 0.25 and 0.75. Fig. 3(a) shows how c_t's vary with time which index data densities, and Fig. 3(b) plots one realization of observations, given that $T = 500$.

If we are to use logarithmic scoring rule, can we come up with $\hat{g}_t$'s that achieve the oracle bound (recall (3)) in the sense that

$$\frac{1}{T}\sum_{t=1}^{T}\{-\log \hat{g}_t(Y_t) + \log g_t^*(Y_t)\} \to_{a.s.} 0 \quad ? \tag{7}$$

It is a challenging prediction task, since the locations and number of abrupt changes, as well as the form of densities are completely unknown. Our proposed methodology will solve this challenge under some mild assumptions on g_t^*'s, e.g. we will require these to belong to various function spaces and to satisfy some mathematical properties such as Lipschitz continuity condition. We will also assume that the number of change points is sub-linear in T.

Example 4 (Autoregression with both unknown abrupt changes and smooth variations). *In this example, we consider time-dependent data. Suppose that the underlying data generating model is a first order time-varying autoregression $Y_t \sim [\theta_t^* Y_{t-1}, 1]$ (not necessarily Gaussian) with parameter space $(-1, 1)$. The true parameter θ_t^* is time-varying in the following manner. At first it follows a linear trend, from $\theta_1^* = 0.9$ to $\theta_{T/2}^* = -0.9$, subsequently it has an abrupt switch to a constant $\theta_{T/2+1}^* = \cdots = \theta_T^* = 0.8$. Fig. 4 shows the sequence $\{\theta_t^*\}_{t=1}^T$ and a realization of observations, with $T = 500$.*

If we use the following square score (commonly used

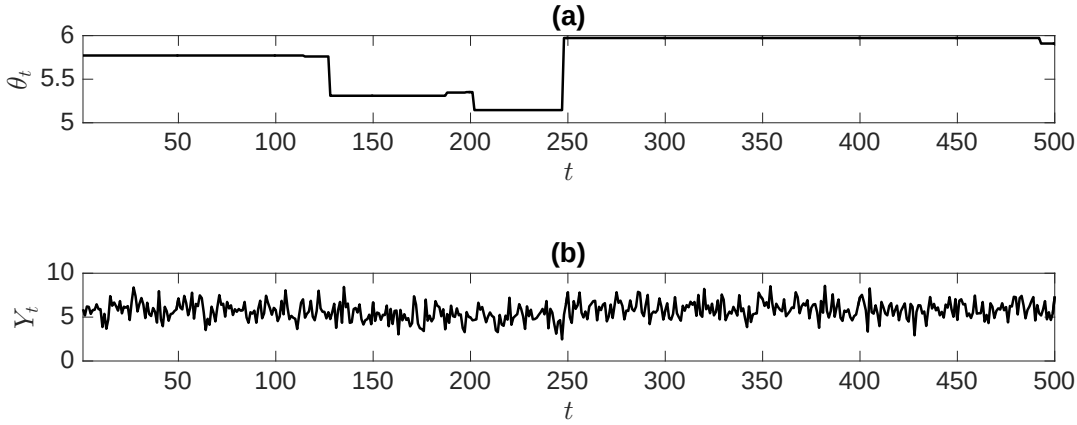

Fig. 2. Example 2: independent Gaussian random variables with growing number of abrupt changes in their means.

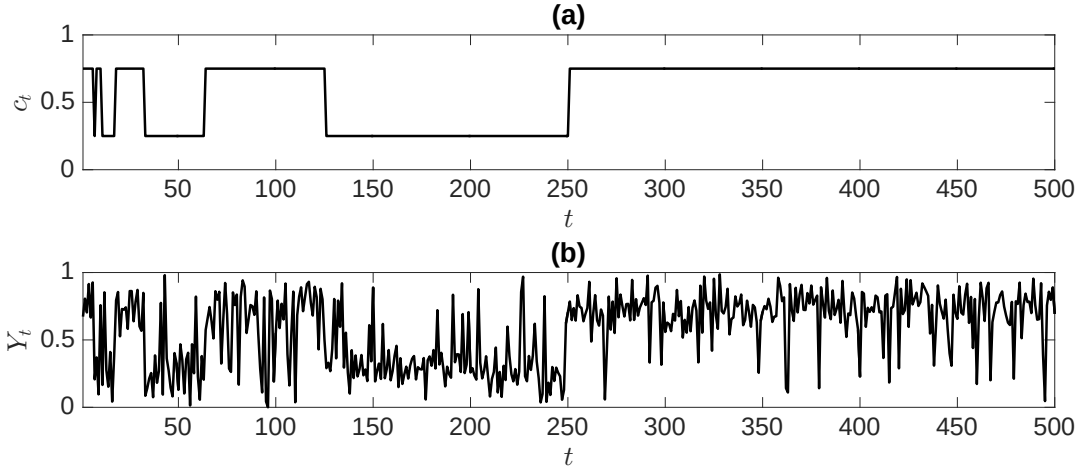

Fig. 3. Example 3: independent random variables with densities varying with time.

for linear models),

$$s(g_{\hat{\theta}_t}, Y_t) = \{Y_t - \mathbb{E}_*(Y_t \mid Y_{1:t-1})\}^2 = (Y_t - \hat{\theta}_t Y_{t-1})^2,$$

then achieving the oracle (3) amounts to

$$\frac{1}{T}\left\{\sum_{t=1}^{T}(Y_t - \hat{\theta}_t Y_{t-1})^2 - \sum_{t=1}^{T} e_t^2\right\} \to_{a.s.} 0. \quad (8)$$

where again e_t is the random Gaussian noise at time t, distributed as $\mathcal{N}(0,1)$. Achieving (8) is not a trivial task, considering that the intercept and slope of each underlying linear trend, the number of trends, and their locations are unknown to data analysts in advance. The example can be generalized to more complex cases where there is a larger number of different linear trends, or even nonlinear (e.g. quadratic and cubic) trends. Our methodology will address these types of challenges using a simplified framework, and provides rigorous theoretical guarantees.

C. Scoring rules and "oracle" predictive performance

The logarithmic scoring rule has been extensively employed in statistical inference, information theory, and machine learning, since, among many other reasons, it appears naturally from the Kullback-Leibler divergence. For instance, maximum likelihood estimation is equivalent to minimizing the expected logarithmic scoring rule: $\hat{\theta}_{MLE} = \arg\min_\theta \int_{\mathcal{Y}} \{-\log g_\theta(y)\} g^*(y) dy$.

In spite of the widespread use of the logarithmic scoring rule, when measuring predictive performance and model evidence, many other scoring rules are also favored due to various reasons. Motivated by the works of [38], [39], we will consider the following general class of scoring rules in our evaluation of predictive performance.

Definition 1 (Proper scoring rule). *Denoting by g^* the true data-generating density function, a scoring rule $s : (g, y) \mapsto s(g, y)$ is said to be "proper" if $\int_{\mathcal{Y}} s(g, y) g^*(y) dy \geq \int_{\mathcal{Y}} s(g^*, y) g^*(y) dy$ for any density*

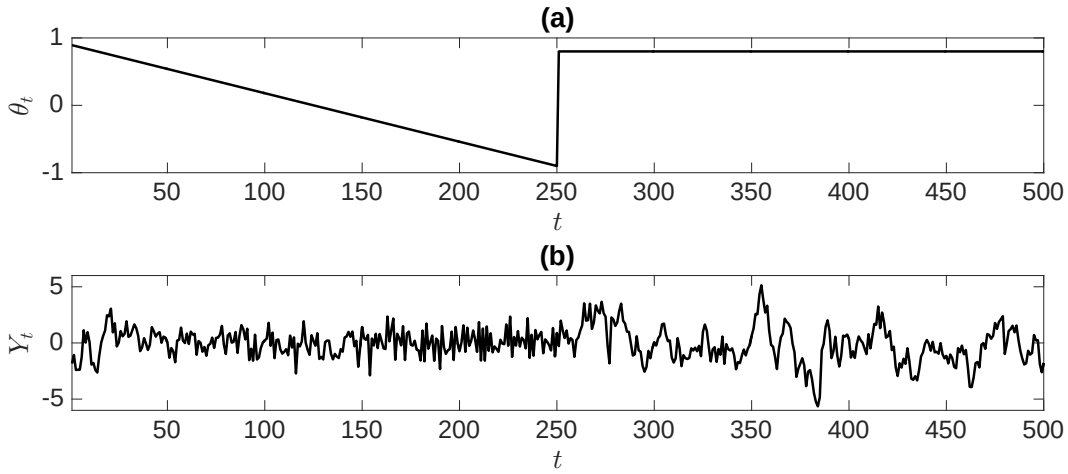

Fig. 4. Example 4: a first order autoregressive process whose coefficient first follows a linear trend, and then abruptly switches to a constant.

function g, and "strictly proper" if the equality happens only if $g = g^*$ (almost everywhere).

Being proper ensures that the user has an incentive to provide a predictive distribution close to g^*, and ideally g^* itself. Two other examples of proper scoring function are

$$s(g, y) = (y - \int \tilde{y} g(\tilde{y}) d\tilde{y})^2,$$

and

$$s(g, y) = -g(y) + \frac{1}{2} \int g(\tilde{y})^2 d\tilde{y},$$

whenever they are well defined for all $g \in \mathcal{G}$. A counterexample of a non-proper function is given by $s(g, y) = -g(y)$.

Next we will prove that for any proper scoring rules, the predictive performance cannot be better than when the true data generating processes is known advance (e.g. with the assistance of a genie). To this end, we define a regularity condition on random variables.

Definition 2 (Tail-exponential random variables). *A nonnegative random variable X is said to be tail-exponential (TE) with parameters λ, a, b ($\lambda > 0, a, b \geq 0$), denoted by $X \sim TE(\lambda; a, b)$, if it satisfies*

$$\mathrm{pr}\big\{X > a + \delta\big\} \leq b \exp(-\delta \lambda^{-1}) \tag{9}$$

for any constant $\delta > 0$.

Roughly speaking, the TE class consists of the absolute values of all random variables that have exponentially decaying densities in the tail. Examples are the truncated Gaussian, Gamma, Exponential, Chi-square, and bounded nonnegative random variables.

We next make the following assumptions on the model class, scoring function and data sequences.

Assumption 1. *Let $\mathcal{G} \subseteq \mathscr{G}$ be the model class (set of density functions) which is a subset of function space $\mathscr{G}$ with a metric $d_{\mathscr{G}}$. There exists a fixed constant $0 < c_{\mathcal{G}} < \infty$ such that $\sup_{g,\tilde{g} \in \mathcal{G}} d_{\mathscr{G}}(g, \tilde{g}) \leq c_{\mathcal{G}}$. In other words, the diameter of $\mathcal{G}$ is bounded.*

Assumption 2. *For all $g, \tilde{g} \in \mathcal{G} \subseteq \mathscr{G}$, $|s(g, Y_t) - s(\tilde{g}, Y_t)| \leq Z(Y_t) \cdot d_{\mathscr{G}}(g, \tilde{g})$, where $Z(\cdot)$ is a nonnegative measurable function such that for all $t = 1, \ldots, T$, $Z(Y_t) \sim TE(\lambda; a, b)$ for some fixed constants $\lambda > 0, a \geq 0$ and $b \geq 0$.*

Assumption 3. *For all $t = 1, \ldots, T$, there exists a fixed constant $c_Y > 0$, such that for all $g_t^* \in \mathcal{G} \subseteq \mathscr{G}$, if $Y_t \sim g_t^*$, then $\mathbb{E}_*\{s(g_t^*, Y_t)^2\} \leq c_Y$.*

Theorem 1 (Oracle performance). *Suppose that Assumptions (1), (2) and (3) hold. Suppose that $g_t^* \in \mathcal{G}$ is a sequence of data generating densities of Y_t for $t = 1, \ldots, T$, and that $s(\cdot, \cdot)$ is a proper scoring rule. Then, for any arbitrary sequence of densities $g_t \in \mathcal{G}$, we have*

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \Big\{ s(g_t, Y_t) - s(g_t^*, Y_t) \Big\} \geq 0, \quad a.s. \tag{10}$$

Proof: The proof is given in the Appendix.

Assumptions (1), (2) and (3) are satisfied by many model classes, such as the Gaussian model class $\mathcal{N}(\theta_t^*, 1)$ with θ_t^* in a compact set, and the model classes in Examples 3 and 4 (as we will show in the sequel).

Remark 1 (The choice of metric). *Clearly an appropriately selected metric $d_{\mathscr{G}}$ can greatly simplify the prediction task. We note that $d_{\mathscr{G}}$ can be relaxed to be a pseudo-norm without affecting the underlying technical proofs. For instance, suppose that the quadratic scoring rule $s(g, Y) = (Y - E_g(Y))^2$ is selected, where*

$\mathbb{E}_g(Y) = \int yg(y)dy$ denotes the mean under density $g(\cdot)$. Then

$$|s(g,Y) - s(\tilde{g},Y)| = |\mathbb{E}_g(Y) - \mathbb{E}_{\tilde{g}}(Y)| \times |2Y - \mathbb{E}_g(Y) - \mathbb{E}_{\tilde{g}}(Y)|$$

Suppose that $\sup_{g\in\mathcal{G}} |\mathbb{E}_g(Y)| < \infty$ and $|Y|$ is tail-exponential with the same parameters over $g \in \mathcal{G}$, and we define $d_{\mathscr{G}}(g,\tilde{g}) = |\mathbb{E}_g(Y) - \mathbb{E}_{\tilde{g}}(Y)|$. Then it is easy to prove that $\mathbb{E}_g(Y^4) < \infty$, and Assumptions (1), (2), and (3) hold.

In the above definition, $d_{\mathscr{G}}$ is a pseudo-norm since $\mathbb{E}_g(Y) = \mathbb{E}_{\tilde{g}}(Y)$ does not necessarily imply $g = \tilde{g}$. This fact can be used to greatly simplify the algorithms to be developed later on (as mathematically we will only need to apply ε-nets to a quotient space).

Remark 2 (Time dependency). *We note that the above assumptions do not require observations at different time steps to be independent. In other words, g_t^* may or may not depend on $Y_{1:t-1}$, as long as $g_t^* \in \mathcal{G}$. In general, it is more cumbersome to verify Assumptions (1), (2) and (3) if Y_t's are dependent. Additionally, the definition of a proper scoring rule clearly does not require the data to be independent. In fact, this definition can be extended to requiring*

$$\mathbb{E}_*\Big\{s(g(\cdot \mid Y_{1:t-1}), Y_t)\Big\} \geq \mathbb{E}_*\Big\{s(g^*(\cdot \mid Y_{1:t-1}), Y_t)\Big\}.$$

For brevity, we have written $s(g_t(\cdot \mid Y_{1:t-1}), Y_t)$ and $s(g_t^*(\cdot \mid Y_{1:t-1}), Y_t)$ respectively as $s(g_t, Y_t)$ and $s(g_t^*, Y_t)$ in the statement of Theorem 1.

An interesting dependent case is the autoregressive models. Consider for instance the AR(1) model in Example 4. Suppose that the initial observation Y_1, noises e_t's, and coefficients θ_t^*'s are bounded and satisfy $\mathbb{E}_*|Y_1| \leq c, \mathbb{E}_*|e_t| \leq c$, $|\theta_t^*| \leq q$ for some constant $c > 0$ and $0 < q < 1$. Then it can be verified that $\mathbb{E}_*|Y_t| = \mathbb{E}_*|\theta_t^* Y_{t-1} + e_t| \leq c/(1-q)$ (using mathematical induction). Thus, the model class can be defined by $\mathcal{G} = \{g : y \mapsto (2\pi\sigma^2)^{-1/2} \exp\{-(y-\mu)^2/(2\sigma^2)\}, |\mu| \leq c/(1-q), \sigma \geq \sigma_0\}$, where $\sigma_0 > 0$ is any fixed constant. Moreover, if we choose the quadratic scoring rule $s(g,Y) = \{Y - \mathbb{E}_g(Y)\}^2$, then we may again use the metric outlined in Remark 1.

III. Prediction When Only Unknown Abrupt Changes Occur: General Solution

In this section, we consider optimal prediction assuming that only unknown abrupt changes happen. We propose an efficient algorithm and provide rigorous theories proving that our method achieves oracle performance without first performing inference (change point detection). The model family can be either parametric or nonparametric. This gives a positive answer to the question (Q1) (posed in Section I). The general idea is to first apply ε-nets to cover the function space by countably many elements (if possible). The elements in an ε-net give our approximations of true data generating density functions (at a time epoch). We then adaptively assign the weights of the elements upon the arrival of each new observation. By appropriately choosing a vanishing rate for $\varepsilon > 0$ and other tuning parameters (as functions of sample size), we will prove the asymptotic optimality in the sense of (3).

Recall that the model class $\mathcal{G}$ is a set of data generating densities. We assume that $\mathcal{G}$ is totally bounded in the metric space $(\mathscr{G}, d_{\mathscr{G}})$. This assumption holds for wide classes of density functions. We note that the Arzelà-Ascoli Theorem (resp. Kolmogorov-Riesz Compactness Theorem) gives a necessary and sufficient condition for a set to be totally bounded in the space of continuous real-valued functions with sup norm (resp. in L^p spaces with L^p norm). We will give more specific examples in the sequel. We will also need the following definition of Kolmogorov and Tikhomirov [41, Chapter 7].

Definition 3 (ε-net and function bases). *A set $U \subseteq \mathscr{G}$ is said to be an ε-net for the set $\mathcal{G}$, if any point of the set $\mathcal{G}$ is located at a distance no greater than ε from some point of the set U. Moreover, for a set $\mathcal{G} \subseteq \mathscr{G}$, we define its ε-entropy $H_\varepsilon(\mathcal{G})$ as the minimum of natural logarithms of the cardinalities of all ε-nets of $\mathcal{G}$. Following Kolmogorov and Tikhomirov, we refer to the elements of an ε-net as function bases.*

The intuition for our approach is given next. Given a totally bounded model class $\mathcal{G} \subseteq \mathscr{G}$, we first discretize it by applying an ε-net $U \subseteq \mathscr{G}$. Let the function base (see above definition) be denoted by $g_i^{(\varepsilon)}$ for $i = 1, \ldots, N$. By definition we have $\log(N) = H_\varepsilon(\mathcal{G})$ (if U is the smallest ε-net). Under appropriate smoothness conditions of the scoring rule, each function base will be thought of as a representative of all the functions within its ε-neighborhood (since all the elements of its neighborhood will have close predictive performance as measured by the score). Based on the sequentially observed data y_t, our prediction for the data generating distribution proceeds by assigning weights to the function bases, and updating them upon arrival of data. This is done in a manner that will quickly track abrupt changes in the underlying data generating densities.

A. Algorithmic description

An algorithmic description of our prediction procedure is summarized in Algorithm 1. We then discuss the related work, examples, and provide rigorous theoretical analysis.

We note that when $\eta = 1$, $\alpha = 0$, and the scoring rule is the negative log likelihood of the data, Algorithm 1

Algorithm 1 Kinetic Prediction with Discretized Function Space

input Discretization parameter $\varepsilon > 0$ and a model class $\mathcal{G} \subset \mathscr{G}$, data $\{y_t : t = 1, \ldots, T\}$ observed sequentially, learning parameter $\eta > 0$, mixing parameter $\alpha \in [0, 1)$.

output $\{\boldsymbol{p}_t \in \mathbb{R}^N : t = 1, \ldots, T\}$ (predictive distributions over the function bases), and $\{\hat{g}_t : t = 1, \ldots, T\}$ (predicted data generating densities).

Choose an ε-net for $\mathcal{G}$ (preferably with minimal cardinality), and obtain function base $\{g_1^{(\varepsilon)}, \ldots, g_N^{(\varepsilon)}\}$ (with cardinality N as described above).

Initialization: $w_{0,1} = \cdots = w_{0,N} = 1/N$;

for $t = 1 \to T$ **do**

  Predict $\hat{g}_t$ according to the distribution $\{\boldsymbol{p}_t\}$ (using one of the methods outlined in Theorem 2 below) with $p_{t,i} = (\sum_{j=1}^N w_{t-1,j})^{-1} w_{t-1,i}$, $i = 1, \ldots, N$;

  Observe (incoming data) y_t and compute $v_{t,i} = w_{t-1,i} \exp\{-\eta\, s(g_i^{(\varepsilon)}, y_t)\}$ for each $i = 1, \ldots, N$.

  Let $w_{t,i} = (1 - \alpha) v_{t,i} + \alpha N^{-1} \sum_{j=1}^N v_{t,j}$ for each $i = 1, \ldots, N$.

end for

becomes the standard Bayesian posterior update. Smaller η and properly chosen nonzero α give a "tempering" effect on the weight updating, offering more flexibility/tolerance for potential underlying changes, while producing different rates of convergence.

Remark 3 (Relation with non-statistical online learning). *Algorithm 1 is similar to the fixed share algorithm in the online learning literature [42], [43], where prediction is non-statistical and is evaluated in terms of "regret". Instead of assumptions on a specific data generating process, it measures the forecaster's performance through the baseline of "experts". At each time before the arrival of a new data point, each expert "gives" a prediction, and the forecaster average the experts' predictions. The goal here is to keep the cumulative "loss" (which is often assumed to be bounded) close to that of the best expert. Although non-statistical learning theory is apparently distinct from statistical prediction tasks in terms of formulations and goals, there do exist many technical relations. Suppose we apply an ε-net to a model class $\mathcal{G}$. We may think of our function bases (resp. statical scoring rule) as N "experts" (resp. learning loss) in online learning.*

Similarly, our continuous-updating procedure (to be) presented in Section IV can be thought of as an extension of classical expert learning from countably many to uncountably many experts.

Algorithm 1 is very simple to implement. In practice, the difficulty mainly comes from determining a good ε-net of the density function space. For parametric model class, this can be done by directly discretizing the parameter space. For nonparametric model class, we may design more sophisticated ε-nets. Some examples are given below.

Example 5 (Parametric model class). *Let θ denote the parameter for the data generating process. Then we may then write g_t (resp. g_t^*) as g_{θ_t} (resp. $g_{\theta_t^*}$) to emphasize dependence on the parameter. Suppose that the parameter space Θ is a close interval $[\underline{\theta}, \overline{\theta}]$ in $\mathbb{R}$. We can construct an N element ε-net by letting $\varepsilon = (\overline{\theta} - \underline{\theta})/(2N)$ and by uniformly dividing the interval $[\underline{\theta}, \overline{\theta}]$ into equal length segments. Each function base then corresponds to $\theta_i^{(\varepsilon)} = \underline{\theta} + (2i - 1)\varepsilon$, for $i = 1, 2, \ldots, N$.*

Example 6 (Nonparametric: Lipschitz class). *Let $\mathcal{G}$ be the set of density functions $g(\cdot)$ defined on the closed interval $\mathcal{Y} = [\underline{y}, \overline{y}]$ satisfying*

$$|g(y) - g(y')| \le L|y - y'|$$

for some Lipschitz constant $L > 0$, $g(y) \ge c$, and $g(\underline{y}) = c$ for some nonnegative constant c.

Let $\tilde{\mathcal{G}}$ denote the set of nonnegative functions satisfying the above conditions, except that the elements are not required to have ℓ_1-norms equal to one. Clearly, $\mathcal{G} \subset \tilde{\mathcal{G}}$. Consider the metric induced by $\|\cdot\|_\infty$ norm, i.e., $d(g_1, g_2) \triangleq \sup_{y \in \mathcal{Y}} |g_1(y) - g_2(y)|$. Based on an adaptation of Kolmogorov's ε-corridor [41, Page 94], we can construct an ε-net of $\tilde{\mathcal{G}}$ denoted by U'. The construction is given as follows. For brevity, we let $\varepsilon' \triangleq \varepsilon/2L$ and $|\mathcal{Y}| \triangleq \overline{y} - \underline{y}$. We let B denote all vectors with entries 0 or 1, and length $h_\varepsilon \triangleq \lceil |\mathcal{Y}|/\varepsilon' \rceil$. For each $\boldsymbol{b} = [b_1, \ldots, b_{h_\varepsilon}] \in B$, we define the function $f_{\boldsymbol{b}} : \mathcal{Y} \to \mathbb{R}$ by

$$f_{\boldsymbol{b}}(\underline{y}) = c,$$
$$f_{\boldsymbol{b}}(y) = f_{\boldsymbol{b}}(\underline{y} + (j-1)\varepsilon') + (2b_j - 1)(y - (j-1)\varepsilon')$$

for $\underline{y} + (j-1)\varepsilon' < y \le \underline{y} + j\varepsilon'$, for $j = 1, \ldots, h_\varepsilon$. We then define

$$\tilde{U} = \{f_{\boldsymbol{b}} : \boldsymbol{b} \in B,\ f_{\boldsymbol{b}}(y) \ge 0 \text{ for all } y \in \mathcal{Y}\}.$$

In other words, $\tilde{U}$ consists of nonnegative curves that starts from $(\underline{y}, c)$, moving linearly for every ε-length, with slope L or $-L$ indexed by binary sequences. A demo is drawn in Fig. 5. By similar arguments as Kolmogorov's ε-corridor, it can be proved that $\tilde{U}$ is an ε-net of $\tilde{\mathcal{G}}$, implying

$$H_\varepsilon(\tilde{\mathcal{G}}) \le \log\{\mathit{card}(\tilde{U})\} \le h_\varepsilon \log 2 \qquad (11)$$

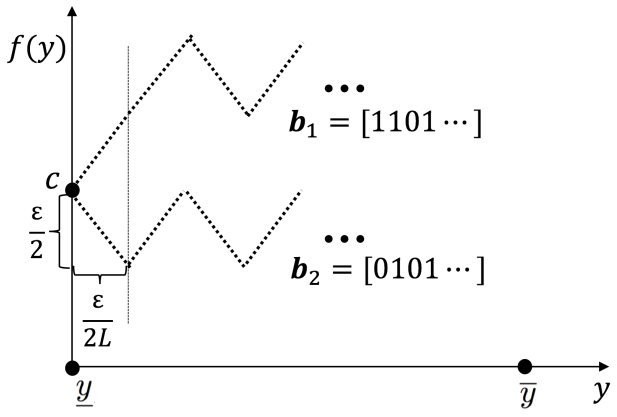

Fig. 5. Illustration of two corridor-shaped function bases.

for small ε.

We define U as the "projection" of $\tilde{U}$ on $\mathscr{G}$, given by

$$U \triangleq \left\{ f = \frac{\tilde{f}}{I_{\tilde{f}}} : \tilde{f} \in \tilde{U} \right\}, \quad \text{where } I_{\tilde{f}} \triangleq \int_{y \in \mathcal{Y}} \tilde{f}(y) dy.$$

Next, we show that U is an $c'\varepsilon$-net of $\mathcal{G}$ for some fixed positive constant c'. The idea of proof is that if for each $f \in \mathcal{G}$, its ε-neighborhood contains some $\tilde{f} \in \tilde{U}$, then the integral of $\tilde{f}$ should be close to one, implying that its normalized version is also close to f.

In fact, for each $f \in \mathcal{G}$, since $\mathcal{G} \subset \tilde{\mathcal{G}}$ and $\tilde{U}$ is an ε-net of $\tilde{\mathcal{G}}$, there exists $\tilde{f} \in \tilde{U}$ that satisfies $d(f, \tilde{f}) \leq \varepsilon$. Therefore, we have

$$\left| I_{\tilde{f}} - 1 \right| \leq \int_{y \in \mathcal{Y}} |\tilde{f}(y) - f(y)| dy \leq \varepsilon |\mathcal{Y}| \tag{12}$$

From (12) and elementary inequalities, we have

$$\begin{aligned} d\left(f, \frac{\tilde{f}}{I_{\tilde{f}}}\right) &\leq d(f, \tilde{f}) + d\left(\tilde{f}, \frac{\tilde{f}}{I_{\tilde{f}}}\right) \\ &\leq \varepsilon + \|\tilde{f}\|_\infty \frac{\varepsilon |\mathcal{Y}|}{1 - \varepsilon |\mathcal{Y}|} \\ &\leq \varepsilon + (c + L|\mathcal{Y}|) \frac{\varepsilon |\mathcal{Y}|}{1 - \varepsilon |\mathcal{Y}|} \\ &\leq c' \varepsilon \end{aligned}$$

for small ε, where $c' \triangleq 1 + 2(c + L|\mathcal{Y}|)|\mathcal{Y}|$ is a fixed constant. Therefore, from (11) we have

$$H_\varepsilon(\mathcal{G}) \leq h_{\varepsilon/c'} \log 2 = O(\varepsilon^{-1})$$

for small ε.

Example 7 (Nonparametric: Sobolev class). *Let $\mathcal{G}$ be the set of all density functions $g(\cdot)$ defined on the closed interval $\mathcal{Y} = [\underline{y}, \overline{y}]$ satisfying $\|g\|_\infty \leq c_1$, with $(k-1)$-th derivative absolutely continuous on $\mathcal{Y}$ and satisfying $\int_{y \in \mathcal{Y}} (g^{(k)}(x))^2 dx \leq c_2$ for some fixed $k \in \mathbb{N}, c_1, c_2 \in \mathbb{R}$.*

Let $\tilde{\mathcal{G}}$ denote set of functions satisfying the same conditions as above, except that the elements are not required to be integrated to one and be nonnegative. Using the result from [44, 19.10], it can be proved that $H_\varepsilon(\tilde{\mathcal{G}}) = O(\varepsilon^{-1/k})$ for small ε, under the metric induced by $\|\cdot\|_\infty$ norm. Using a similar "projection" technique to that proposed in Example 6, we can prove that $H_\varepsilon(\mathcal{G}) = O(\varepsilon^{-1/k})$ for small ε.

B. Theoretical Analysis

For the presentation of our results, we will need a technical definition.

Definition 4 (ε-predictable). *A sequence of density functions $g_1^*, \ldots, g_T^* \in \mathcal{G} \subseteq \mathscr{G}$ is ε-predictable with respect to a sequence of positive integers $\{M_T\}$, if there exists a fixed constant $\beta \in (0, 1/4)$ and a deterministic sequence $\{\varepsilon_T\}$ such that*

$$\lim_{T \to \infty} \left(T^\beta \varepsilon_T + \frac{M_T H_{\varepsilon_T}(\mathcal{G})}{T^{1-4\beta}} \right) = 0. \tag{13}$$

The main result of this section is given in the following theorem.

Theorem 2 (Prediction performance under abrupt changes). *Suppose that Assumptions (1), (2), and (3) hold, and that the true data generating distribution sequence $\{g_t^*\}_{t=1}^T$ has at most $M_T - 1$ abrupt changes by time T.*

(i) Suppose that $\{g_1^{(\varepsilon)}, \ldots, g_N^{(\varepsilon)}\}$ is an ε-net of $\mathcal{G}$, and that $\beta > 0$ is an arbitrary constant chosen such that $T^\beta > \max\{c_{\mathcal{G}} a, a\}$ (where $c_{\mathcal{G}}, a$ were defined in Assumptions (1) and (2)). If we choose α and η by

$$\alpha = \frac{M_T - 1}{T - 1}, \quad \eta = \sqrt{\frac{8 Q_{T,N}}{T^{1+2\beta}}}, \tag{14}$$

where

$$Q_{T,N} \triangleq M_T \log N + (T-1) H\left(\frac{M_T - 1}{T - 1}\right),$$

then Algorithm 1 outputs $\boldsymbol{p}_t : t = 1, \ldots, T$ such that

$$\begin{aligned} &\sum_{t=1}^T \sum_{i=1}^N p_{t,i} \; s(g_i^{(\varepsilon)}, Y_t) - \sum_{t=1}^T s(g_t^*, Y_t) \\ &\leq \sqrt{2^{-1} T^{1+2\beta} Q_{T,N}} + T^{1+\beta} \varepsilon. \end{aligned} \tag{15}$$

with probability at least $1 - C_1 T \exp(-C_2 T^\beta)$, where C_1, C_2 are fixed constants not depending on T.

(ii) Suppose that $\{g_t^*\}_{t=1}^T$ is ε-predictable, then there exist N (depending on T) and function bases

$g_{1,T}^{(\varepsilon)}, \ldots, g_{N,T}^{(\varepsilon)}$ *such that the output of Algorithm 1 satisfies*

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left\{\sum_{i=1}^{N} p_{t,i}\ s(g_{i,T}^{(\varepsilon)}, Y_t) - s(g_t^*, Y_t)\right\} \leq 0 \quad a.s. \tag{16}$$

(iii) Under all the above assumptions, suppose that $s(g, y)$ is convex in its first argument, and we use $\hat{g}_t = \sum_{i=1}^{N} p_{t,i} g_{i,T}^{(\varepsilon)}$ to predict at time t. Then

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left\{s(\hat{g}_t, Y_t) - s(g_t^*, Y_t)\right\} = 0 \quad a.s. \tag{17}$$

(iv) Under all the above assumptions, suppose that we independently draw J_t from $1, \ldots, N$ with probabilities specified in $\boldsymbol{p}_t$, and use $\hat{g}_t = g_{J_t,T}^{(\varepsilon)}$ as predictor at each time t. Then

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left\{s(\hat{g}_t, Y_t) - s(g_t^*, Y_t)\right\} = 0 \quad a.s. \tag{18}$$

Proof: The proof is given in the Appendix.

Remark 4. *Theorem 2 Part (i) gives a non-asymptotic bound on the sum $\sum_{i=1}^{N} p_{t,i}\ s(g_i^{(\varepsilon)}, y_t)$, the average predictive score over all function bases at each time. The bound then leads to an asymptotic inequality in Part (ii).*

Parts (iii) and (iv) can be implemented for prediction in practice. The predictor in Part (iii) averages the function bases according to their predictive weights, and in Part (iv) it is randomly drawn from the function bases according to their predictive weights. Under reasonable assumptions, the scores of both predictors achieve the oracle bound asymptotically (recall Theorem 1). From extensive numerical studies, we conjecture that the predictor in Part (iii) exhibits less prediction variance than that of Part (iv). On the other hand, it requires the convexity of the score function which is not needed for the predictor of Part (iv).

With only abrupt changes, the underlying density function sequence is piece-wise constant. As we introduced before, classical change detection framework often requires that changes happen far away from each other. In contrast, our kinetic prediction method does not have such a requirement, and it can automatically approach the oracle predictive performance. Moreover, our method incorporates both nonparametric and parametric model classes into a unified framework.

C. The inference-prediction dilemma

We have asked before "whether it is possible to achieve optimal prediction without inferring the number and locations of abrupt change points first?". To this point, we have partially answered this question affirmatively for the case when the unknown sequence of densities comes from a size-controlled (in terms of Kolmogorov-Tikhomirov ε-entropy) model class with a reasonable number of abrupt changes.

The following proposition sheds further light on this question.

Proposition 2 (Prediction without inference). *Under the assumptions of Theorem 2, there exists no statistical procedure that is always consistent in selecting the number of abrupt change points.*

Proof: The proof is given in the Appendix.

We note that the distinction between the tasks of *inference* and *prediction* originates from different statistical objectives. The task of prediction focuses on minimizing some statistical scoring function, whether or not the model class is well-specified. In contrast, the task of inference typically focuses on accurate identification of the data generating model (from a well-specified model class). It is a common practice to apply inference first, and then use the obtained estimator (from either frequentist or Bayesian perspective) for prediction. However, it is not clear that this two-stage approach is always the right way.

Interestingly, such inference-prediction dilemma has been also directly or implicitly observed in several domains. For example, in high dimensional regression analysis, it has been shown that near-oracle prediction loss is possible even the data-generating variables cannot be consistently selected [19], [45]. In model selection, there have been many debates on whether Akaike information criterion (AIC) [20] or Bayesian information criterion (BIC) [21], which also represent a broader range of criteria, should be used. It is known that in a nested well-specified model class, BIC-like criteria is typically consistent in selecting the true model, while AIC suffers a non-vanishing overfitting probability. However, AIC is shown to exhibit optimal predictive power in misspecified settings, while BIC is not [23]. A comprehensive review in that direction can be found in [46].

We note that for change point analysis, questions **(Q1)** and **(Q2)** (see the introduction) are legitimate, whenever we are only interested in predictive performance.

IV. Prediction under unknown Abrupt Changes: Parametric Solution

In the above, we have discussed our kinetic prediction framework for data generating distributions in general

function spaces under abrupt changes. Applying an ε-net, we can "represent" potentially infinite dimensional space of the model class by countably many function bases ("representatives"). Under mild smoothness conditions, we proved that by applying our prediction strategy to the discretized function bases, the oracle predictive performance can be asymptotically approached. Although this "discretization technique" can be applied to both non-parametric and parametric function spaces, when we are working with a fully parametric model class, we may perform prediction directly with the continuous space. This can be seen as the limiting case when $\varepsilon \to 0$, and the partition produces uncountably many arbitrarily close representatives of the space. Our kinetic prediction methodology for this scenario aims to automatically search the space (using Markov chain particles) and update the best performing parameters sequentially.

More specifically, suppose the parametric data generating distribution is $g_\theta(\cdot)$, indexed by a parameter $\theta \in \Theta \subseteq \mathbb{R}^n$, where the parameter space Θ is compact. We shall then identify each density function with a parameter θ. For example, the true data generating densities $g_t^*(\cdot)$ can be described by the underlying parameter sequence θ_t^*, the scoring rule $s(g, y)$ can be rewritten as $s(\theta, y)$, and the task of making predictions of the data generating process is then turned into the prediction of the time-varying parameter θ_t^*, when there are no ambiguities.

We will assume the model class to be well-specified, i.e. the true data generating parameter $\theta_t^* \in \Theta$, $\forall t = 1, \ldots, T$, and the parameter space Θ is a hypercube in $\mathbb{R}^n$ (also called an $n-$dimensional cube). In particular, $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_n$, with each Θ_j being a one-dimensional closed interval: $\Theta_j = [\underline{\theta}_j, \overline{\theta}_j] \in \mathbb{R}$, $j = 1, \ldots, n$. We use $|\cdot|$ to denote the Lebesgue measure of the parameter space (or the volume of a hypercube here), namely $|\Theta_j| = (\overline{\theta}_j - \underline{\theta}_j)$ and $|\Theta| = \prod_{j=1}^n |\Theta_j|$. We will predict by working with this continuous parameter space directly, assuming that the underlying true parameter sequence $\{\theta_t^*\}_{t=1}^T$ contains unknown abrupt changes. Our goal is to achieve the oracle performance as previously defined.

A. Algorithmic description

We summarize the pseudocode in Algorithm 2. At each time step t, before the next observation y_t is revealed, the algorithm gives a *predictive distribution* $p_t(\theta)$ of the underlying parameter θ_t^*, based on the past $y_{1:t-1}$ and scoring functions. The point predictor $\hat{\theta}_t$ for θ_t^* can then be generated based on the predictive distribution. For instance, it could be chosen to be the mean of $p_t(\theta)$, or a random draw according to it (See Parts (iii) and (iv) of Corollary 1).

In Algorithm 2, the updating step (19) resembles that of Bayesian update, with the only difference in learning parameter η. A discussion on this has been made in Subsection III-A. The mixing step (20) is crucial as it redistributes belief on the parameter space. At an informal level, it provides the potential for those parameters with little evidence in the past to quickly gain more evidence in the future whenever they become the true parameters (after abrupt changes). The mixing parameter α is related to the frequency of abrupt changes occurring in the sequence. The less frequent changes occur, the less $\alpha \geq 0$ is needed to redistribute the belief. This can also be seen from $\alpha^{opt} = \frac{M_T - 1}{T-1}$ in the following Proposition 3.

B. Theoretical analysis

We next discuss analogs of assumptions made in Section II-C needed for theoretical analysis. Assumption (1) is automatically satisfied if the parameter space Θ is a bounded hypercube, and if we define the metric by $d_{\mathscr{G}}(g_\theta, g_{\tilde{\theta}}) \triangleq \|\theta - \tilde{\theta}\|_2$. We thus introduce the following analogs of Assumptions (2) and (3) in the parametric setting.

Assumption 2′. *The parameter space $\Theta \subseteq \mathbb{R}^n$ is a hypercube. For all $\theta, \tilde{\theta} \in \Theta \subseteq \mathbb{R}^n$, we have $|s(\theta, Y_t) - s(\tilde{\theta}, Y_t)| \leq Z(Y_t) \cdot \|\theta - \tilde{\theta}\|_2$ for all $t = 1, \ldots, T$, where $Z(\cdot)$ is a nonnegative measurable function such that for all $t = 1, \ldots, T$, $Z(Y_t) \sim TE(\lambda; a, b)$ for some fixed constants $\lambda > 0, a \geq 0$ and $b \geq 0$.*

Assumption 3′. *For all $t = 1, \ldots, T$, there exists a fixed constant $c_Y > 0$ that does not depend on t, such that for all $\theta_t^* \in \Theta \subseteq \mathbb{R}^n$, if $Y_t \sim g_{\theta_t^*}$, then $\mathbb{E}_*\{s(\theta_t^*, Y_t)^2\} \leq c_Y$.*

The following theorem provides a performance guarantee for the parametric kinetic prediction.

Theorem 3 (Finite prediction bound). *Suppose that Assumption (2′) holds, and we let*

$$\Delta \triangleq \sup_{\theta, \tilde{\theta} \in \Theta} \|\theta - \tilde{\theta}\|_2.$$

For the kinetic prediction Algorithm 2, we define the average predictive score

$$\bar{s}(\theta, y_t) \triangleq \int_\Theta s(\theta, y_t) p_t(\theta) d\theta$$

for a realization y_t at time t. If the underlying parameter sequence $\{\theta_t^*\}_{t=1}^T$ contains at most $M_T - 1$ ($M_T \geq 1$)

Algorithm 2 Kinetic Prediction with Continuous Parameter Space

Input: Compact parameter space $\Theta \subseteq \mathbb{R}^n$, data $\{y_t, t = 1, \ldots, T\}$ observed sequentially, learning parameter $\eta > 0$, and mixing parameter $\alpha \in [0, 1)$.

Output: Predictive distribution of the unknown parameter $\{p_t(\theta), t = 1, \ldots, T\}$ and predicted sequence $\{\hat{\theta}_t\}_{t=1}^T$.

1: Initialization: $f_0(\theta) = 1$, $\forall \theta \in \Theta$;
2: **for** $t = 1 \to T$ **do**
3: Predict parameter $\hat{\theta}_t$ according to the predictive distribution $p_t(\theta) = \{\int_\Theta f_{t-1}(\tilde{\theta})d\tilde{\theta}\}^{-1} f_{t-1}(\theta)$;
4: After receiving y_t, compute the score function $s(\theta, y_t)$ for all $\theta \in \Theta$;
5: Update:

$$\tilde{f}_t(\theta) = f_{t-1}(\theta) \cdot e^{-\eta s(\theta, y_t)} \tag{19}$$

6: Mix:

$$f_t(\theta) = (1 - \alpha)\tilde{f}_t(\theta) + \alpha \frac{\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}}{|\Theta|} \tag{20}$$

7: **end for**

abrupt changes, then for any value $d > a$ we have

$$\sum_{t=1}^T \bar{s}(\theta, Y_t) - \sum_{t=1}^T s(\theta_t^*, Y_t) \le \frac{1}{\eta}\Bigg[(T - M_T)\ln\frac{1}{1-\alpha} + (M_T - 1)\ln\frac{1}{\alpha} - (T - M_T)\ln\left(1 + \frac{\alpha}{1-\alpha}r_T^{(n)}\right) - \sum_{t=T-M_T+1}^T \ln(r_t^{(n)})\Bigg] + \frac{\eta T d^2 \Delta^2}{8} \tag{21}$$

with probability at least $1 - bT\exp\left[-(d-a)\lambda^{-1}\right]$, where

$$r_t^{(n)} \triangleq \prod_{j=1}^n \frac{1 - \exp(-\eta t d|\Theta_j|)}{\eta t d|\Theta_j|} \in (0, 1). \tag{22}$$

Proof: The proof is given in the Appendix.

It can be proved that $\{r_t^{(n)}\}_{t=1}^T$ is a monotonically decreasing sequence in t. Thus

$$\sum_{t=T-M_T+1}^T \ln(r_t^{(n)}) \ge M_T \ln(r_T^{(n)}).$$

Also we have $(T - M_T)\ln\{1 + \alpha r_T^{(n)}/(1-\alpha)\} \ge 0$. In light of these, the inequality (21) can be written in a slightly weaker but more compact form:

$$\sum_{t=1}^T \bar{s}(\theta, y_t) - \sum_{t=1}^T s(\theta_t^*, y_t) \le \frac{1}{\eta}\Bigg[(T - M_T)\ln\frac{1}{1-\alpha} + (M_T - 1)\ln\frac{1}{\alpha} - M_T \ln(r_T^{(n)})\Bigg] + \frac{\eta T d^2 \Delta^2}{8}. \tag{23}$$

In practice we want to choose good values for parameters α and η to run Algorithm 2 such that the regret bound is small. The following results give some insight on how to select the parameters.

Proposition 3 (Choice of parameters). *Regardless of the learning parameter η, for the regret bound in (21) to be minimized, the optimal mixing parameter is given by*

$$\alpha^{opt} = \frac{M_T - 1}{(T-1)(1 - r_T^{(n)})}.$$

For the bound in (23) to be minimized, the optimal mixing parameter is exactly $\alpha^{opt} = (M_T - 1)/(T - 1)$, using which (23) reduces to

$$\sum_{t=1}^T \bar{s}(\theta, y_t) - \sum_{t=1}^T s(\theta_t^*, y_t) \le \frac{1}{\eta}\Bigg[(T-1)H\left(\frac{M_T - 1}{T-1}\right) - M_T \ln(r_T^{(n)})\Bigg] + \frac{\eta T d^2 \Delta^2}{8} \tag{24}$$

where $H(\cdot)$ is the binary entropy function. The optimal value for the learning parameter η has no close form, but can be calculated by numerical methods.

Corollary 1 (Asymptotic prediction performance). *Suppose that Assumption (2′) holds, $\{\theta_t^*\}_{t=1}^T$ is the true data generating parameter sequence, and the model class represented by Θ is well-specified, then:*

(i) If Assumption (3′) holds, and the scoring rule $s(\cdot, \cdot)$ is proper, then for any parameter sequence $\{\theta_t\}_{t=1}^T$, we have

$$\liminf_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^T s(\theta_t, Y_t) - \sum_{t=1}^T s(\theta_t^*, Y_t)\right\} \ge 0 \quad a.s.$$

(ii) If the true parameter sequence $\{\theta_t^*\}_{t=1}^T$ has at most $M_T - 1$ abrupt changes by time T, and $M_T = O(T^\gamma)$ with $\gamma \in [0, 1)$, then Algorithm 2 outputs $p_t(\theta)$ that satisfies

$$\limsup_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^T \bar{s}(\theta, Y_t) - \sum_{t=1}^T s(\theta_t^*, Y_t)\right\} \le 0 \quad a.s.$$

given mixing parameter $\alpha_T = (M_T - 1)/(T - 1)$, and learning parameter $\eta_T = O(T^{-\beta})$ where $0 < \beta < (1-\gamma)/2$.

(iii) Under the above assumptions (in Parts (i) and (ii)), if the scoring rule $s(\cdot,\cdot)$ is convex in its first argument $\theta \in \Theta$, and the predicted parameter $\hat{\theta}_t$ from Algorithm 2 is given by $\hat{\theta}_t = \mathbb{E}_{p_t}(\theta) = \int_\Theta \theta p_t(\theta) d\theta$, then

$$\lim_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} = 0 \quad a.s.$$

(iv) Under the above assumptions (in Parts (i) and (ii)), if the predicted parameter $\hat{\theta}_t$ from Algorithm 2 is independently drawn from the predictive distribution $p_t(\cdot)$, then

$$\lim_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} = 0 \quad a.s.$$

Proof: The proof is given in the Appendix.

Remark 5 (Discussion on the theoretical results). *Theorem 3 and Proposition 3 give a predictive score bound on the difference between the average score produced by our kinetic algorithm and that produced by a genie (who knows that the true parameter sequence is piece-wise constant with M_T segments). Recall that the genie's performance is the oracle bound if $s(\cdot,\cdot)$ is a proper scoring rule. Thus our prediction method is guaranteed not to perform too badly, for any known data horizon T. Corollary 1 shows that our prediction scheme asymptotically approaches the oracle bound, as long as the number of abrupt changes grows sub-linearly with time.*

In the proposed kinetic algorithms, we implicitly require the knowledge of data horizon T. This is a reasonable assumption in many applications. For example, in high-frequency trading, the data horizon for predicting the prices of a second-bar financial asset is typically chosen to be one day. In some other applications such as trajectory prediction or weather forecasting, prediction needs to made continuously, so the data horizon is not prescribed. In the second case, the proposed methodology is not directly applicable, since appropriately chosen tuning parameters α, η require the knowledge of T (see Corollary 1.(ii)). We conjecture that this issue can be solved by a "doubling trick". Specifically, we may start with an initial data horizon T_0 and the corresponding α_{T_0}, η_{T_0}; Every time the current size of data t is doubled, e.g., at $t = 2T_0$, we update the parameters to $\alpha_{2T_0}, \eta_{2T_0}$ according to the doubled data horizon.

Recall that in Corollary 1.(ii), the mixing parameter $\alpha = (M_T - 1)/(T-1)$ also depends on M_T, the number of abrupt changes up to and including time step T. In practice this could be set from a data analyst's prior information. Otherwise, an appropriate α can be searched from a grid of $[0, 1]$, based on the corresponding average predictive score. An alternative approach is to postulate $\alpha = T^\rho$ with $\rho \in [0, 1)$ being an unknown parameter, and then apply Monte Carlo methods in an analogous manner to that of [33]. Our numerical studies show that the predictive power of the kinetic prediction is not very sensitive to a wide range of choices of α, η.

C. Efficient Monte Carlo implementation

In contrast with Algorithm 1, where the weight updates can be directly evaluated, the implementation of Algorithm 2 is more cumbersome. In particular, the numerical calculation of the integral $\int_\Theta \tilde{f}_t(\tilde{\theta}) d\tilde{\theta}$ in the mixing step (20) may be computationally difficult in practice. Thus we need to develop efficient methods to evaluate the continuous predictive distribution functions $p_t(\theta)$ and related integrals, as well as methods for sampling from the predicted distributions. Motivated by the resemblance of our methods to Bayesian approach, we propose to use Monte Carlo methods for an efficient implementation.

Our idea is to first recast Algorithm 2 as a special case of state space models, and then apply particle filter techniques. In fact, we can rewrite the mixing step (20) as

$$\begin{aligned} f_t(\theta) &= (1-\alpha)\tilde{f}_t(\theta) + \alpha \frac{\int_\Theta \tilde{f}_t(\tilde{\theta}) d\tilde{\theta}}{|\Theta|} \\ &= \int_\Theta \tilde{f}_t(\tilde{\theta}) K(\theta|\tilde{\theta}) d\tilde{\theta} \end{aligned}$$

where

$$K(\theta|\tilde{\theta}) \triangleq (1-\alpha)\delta(\theta) + \alpha/|\Theta|$$

which is a transition kernel from $\tilde{\theta}$ to θ, and $\delta(\cdot)$ is the Dirac delta function in the space Θ. This can be seen as the continuous state transition kernel from $\tilde{\theta}$ to θ in a hidden Markov model.

Interestingly, the above formulation gives a natural Bayesian interpretation. Instead of a frequentist perspective that assumes a certain number of change points in the true data generating parameter sequence $\{\theta_t^*\}_{t=1}^T$, we now think of $\{\theta_t\}_{t=1}^T$ as hidden states that follow some transitional law. The output $p_t(\cdot)$ in Algorithm 2 is then interpreted as the predictive distribution $p(\theta_t|y_{1:t-1})$.

Following the above Bayesian interpretation, we design the following three-step particle filter for implementing Algorithm 2. We start by drawing initial particles (or samples) with equal weights from a uniform distribution on Θ, and then perform the following three steps at each time step:

(1) Update the weights based on $\exp(-\eta s(\cdot, y_t))$;
(2) Resample the particles according to their weights;
(3) Move the particles according to the kernel $K(\cdot,\cdot)$.

The last step is easily implemented by either keeping a particle unchanged with probability $(1-\alpha)$, or replacing it with a random draw from $\mathrm{Unif}(\Theta)$ with probability α.

Despite its simplicity, the above implementation of particle filter suffers two well known problems, namely the sample degeneracy and impoverishment [47]. In practice, if α is chosen to be too small, the moving step (3) doesn't change much of the samples' locations, so the particle filter is very similar to a sequential importance sampling algorithm with resampling [48]. It is foreseeable (and experiments also show) that although the samples could still follow the changes in underlying parameters, the underlying weights will quickly concentrate on a small number of samples. And after resampling, most of the samples fall in the same location in the parameter space. In order to resolve this issue, we borrow some ideas from the iterated batch importance sampling (IBIS) algorithm [49], and propose the following Algorithm 3.

In the above algorithm, resampling is triggered when ESS (Effective Sample Size, estimated from the particle weights) drops below a threshold (step (9)). There are various ways to do resampling, such as multinomial resampling, residual resampling, and systematic resampling [50]. It is then followed by a rejuvenation step, which moves the particles according to some Markov kernel K' which leaves the current distribution invariant. A typical kernel that we use is a Metropolis-Hastings (MH) kernel [51], which proceeds as follows:

(a) Sample $\tilde{\theta} \sim q(\cdot|\bar{\theta}_t^i)$, where q is the proposal distribution of our choice [52] (for example, we can choose q to be an independent Gaussian distribution with mean and covariance estimated from samples);
(b) Compute

$$A(\tilde{\theta}|\bar{\theta}_t^i) = \min\left(1, \frac{\tilde{f}_t(\tilde{\theta})q(\bar{\theta}_t^i|\tilde{\theta})}{\tilde{f}_t(\bar{\theta}_t^i)q(\tilde{\theta}|\bar{\theta}_t^i)}\right);$$

(c) With probability $A(\tilde{\theta}|\bar{\theta}_t^i)$, set $\theta_{t+1}^i = \tilde{\theta}$, otherwise set $\theta_{t+1}^i = \bar{\theta}_t^i$.

The above step only requires point evaluation of the (unnormalized) targeting distribution $\tilde{f}_t(\theta)$. A nontrivial part in such evaluation is to calculate integrals $\int_\Theta \tilde{f}_\tau(\tilde{\theta})d\tilde{\theta}$ for $\tau = 1, 2, \ldots, t-1$, according to (19) and (20). Fortunately, these integrals can be estimated as a by-product in our algorithm. In fact, for the quantities in Step (6), direct calculations show that

$$\tilde{R}_t \underset{N\to\infty}{\longrightarrow} \frac{\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}}{\int_\Theta \tilde{f}_{t-1}(\tilde{\theta})d\tilde{\theta}} \quad (\text{for } t \geq 2), \quad \text{and}$$

$$\tilde{R}_1 \underset{N\to\infty}{\longrightarrow} \int_\Theta \tilde{f}_1(\tilde{\theta})d\tilde{\theta} \quad \text{in probability.}$$

We therefore obtain

$$\hat{R}_t \underset{N\to\infty}{\longrightarrow} \int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta} \quad \text{in probability}$$

for a given t, which gives a consistent estimator of the integral $\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}$. Therefore, using these estimators $\{\hat{R}_\tau\}_{\tau=1}^{t-1}$, we can evaluate $\tilde{f}_t(\theta)$ based on recursive equations (19) and (20) in Algorithm 2 for any $\theta \in \Theta$, so that an MH rejuvenation can be done after resampling to increase sample diversity. It would be interesting to study the asymptotic regime where N diverges as some function of T, so that the propagated errors from Monte Carlo approximation are always negligible. We leave that for future research. The complexity of the recursive procedure of $\tilde{f}_t(\theta)$ function evaluation for all the samples at time t is $O(Nt)$, so the overall complexity of Algorithm 3 is $O(NT^2)$.

V. Prediction under Smooth Variations as well as Abrupt Changes

In practice, we often encounter situations when the data generating distributions in various epochs of time vary smoothly. By "smooth variations," we mean small but persistent changes of densities during a certain time epoch. In those cases, it may not be a good idea to approximate density functions $g_1, g_2, \ldots$ by a number of functions with occasional abrupt changes, since changes (either small or abrupt) happen "all the time". However, if such small variations can be well-approximated by *locally* linear, quadratic, or other deterministic patterns (illustrated in Fig. 6), it is possible to reduce the dimension by treating a sequence of parameters (with a given change pattern) as "one parameter". Following this intuition, we propose the following concept which we refer to as "function flow". As a result, we show that it is still possible to achieve optimal prediction under smooth variations as well as abrupt changes, by adapting our previous methodology.

Definition 5 (Function/parameter flow). *A function flow f on a set of function bases $\mathcal{G}$ is defined to be a map $f : \mathbb{N} \to \mathcal{G}$. If each $g \in \mathcal{G}$ is parameterized with $\theta \in \Theta$, we also call f a parameter flow, and write $f : \mathbb{N} \to \Theta$. The set of all the flows is denoted by $\mathscr{F}$.*

Consider, for example, independent Gaussian $\mathcal{N}(\mu_t, 1)$ with $\mu_t \in [0, 1]$, $t = 1, 2, \ldots$. Suppose that we have an ε-net bases $\mu_1 = \varepsilon, \mu_2 = 2\varepsilon$, etc. If the true means follow a linear trend in a certain period, then it can be approximated by $\mu_t = (a + bt)\varepsilon$ for some integers a, b. It would be easier to apply our kinetic procedure to *two* unknown a, b instead of too many unknown μ_t's. We can define a set of parameter flows $f_\varepsilon : t \mapsto \varepsilon t$ for a grid of ε's, and then apply kinetic prediction procedure to the flows (in an analogous manner to Algorithm 1).

Algorithm 3 Sequential Monte Carlo for Kinetic Prediction (SMC-KP) in Compact Parameter Spaces

Input: Compact parameter space $\Theta \subseteq \mathbb{R}^n$, data $\{y_t\}_{t=1}^T$ observed sequentially, learning parameter $\eta > 0$, mixing parameter $\alpha \in (0,1)$, number of particles N, and ESS threshold $c \in [0,1]$.

Output: $(W_t^i, \theta_t^i)_{i=1}^N$ as weighted samples from the predictive distribution $\{p_t(\theta)\}_{t=1}^T$ defined in Algorithm 2.

1: Initialization: sample θ_1^i independently from $\mathrm{Unif}(\Theta)$, and let $W_1^i = N^{-1}$, for all $i = 1, \ldots, N$.
2: **for** $t = 1 \to T$ **do**
3: Use the weighted samples $(W_t^i, \theta_t^i)_{i=1}^N$ to approximate the predictive distribution $p_t(\theta)$.
4: Receive/Read y_t.
5: Update $w_t^i = W_t^i \cdot e^{-\eta s(\theta_t^i, y_t)}$ for all $i = 1, \ldots, N$;
6: Calculate $\tilde{R}_t = \sum_{i=1}^N w_t^i$, $\hat{R}_t = \prod_{\tau=1}^t \tilde{R}_\tau$ (to be used in Step (11));
7: Normalize $W_{t+1}^i = w_t^i / \sum_{i=1}^N w_t^i$. Note that $(W_{t+1}^i, \theta_t^i)_{i=1}^N$ approximates the density $\frac{\tilde{f}_t(\theta)}{\int_\Theta \tilde{f}_t(\tilde{\theta}) d\tilde{\theta}}$.
8: Calculate $ESS = 1/\big(\sum_{i=1}^N (W_t^i)^2\big)$;
9: **if** $ESS < cN$ **then**
10: **Resample:**
Resample $(W_{t+1}^i, \theta_t^i)_{i=1}^N$ (e.g. using multinomial distribution) to obtain equally weighted samples $(W_{t+1}^i = N^{-1}, \bar{\theta}_t^i)_{i=1}^N$;
11: **Rejuvenate/Move:**
Draw $\theta_{t+1}^i \sim K'(\cdot|\bar{\theta}_t^i)$ where K' is an MCMC kernel targeting at the density $\frac{\tilde{f}_t(\theta)}{\int_\Theta \tilde{f}_t(\tilde{\theta}) d\tilde{\theta}}$. Our choice of the kernel K' is a Metropolis-Hastings (MH) type, which uses results from step (6) and is explained in detail below this Algorithm.
12: **else**
13: $W_{t+1}^i = W_{t+1}^i$, $\theta_{t+1}^i = \theta_t^i$;
14: **end if**
15: Move θ_{t+1}^i according to the transition kernel $K(\cdot|\theta_{t+1}^i) \triangleq (1-\alpha)\delta(\cdot) + \alpha/|\Theta|$ in the following way: with probability $1-\alpha$, let $\theta_{t+1}^i = \theta_{t+1}^i$, and with probability α, let $\theta_{t+1}^i \sim \mathrm{Unif}(\Theta)$, for all $i = 1, \ldots, N$.
16: **end for**

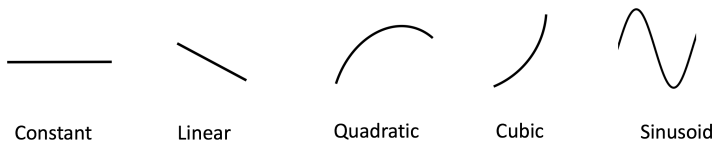

Fig. 6. Some examples of local deterministic trends of parameters (for a parametric model class).

In other words, at each time $t = 1, 2, \ldots$, each base g_n in Algorithm 1 is replaced with some $f(t)$, a flow f evaluated at t.

We note that each function base in Definition 3 can be regarded as a special flow, which takes a constant value as time elapses. As was in Definition 4, the size of all the flows need to be restricted in order to guarantee optimal prediction. Our key ideas and corresponding notation in this section are illustrated in Fig. 7.

In the remaining of this section, for simplicity we only present our methodology by considering a parametric model class with unknown linear trends in the parameters. For further simplicity, we only consider a one-dimensional parameter space $\Theta \subseteq [\underline{\theta}, \overline{\theta}]$ in our presentation. Extension to richer classes of function flows, higher dimensional parameter spaces, and non-parametric models is straightforward.

We denote the parametric model class as $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}$ is equipped with a metric d_Θ, and define the metric $d_{\mathscr{G}}(g_\theta, g_{\tilde{\theta}}) \triangleq d_\Theta(\theta, \tilde{\theta})$. We let $\mathcal{G}^{(\varepsilon)}$ be an ε-net of $\mathcal{G}$, given by

$$g_j^{(\varepsilon)} = g_{\theta_j^{(\varepsilon)}}, \text{ with } \theta_j^{(\varepsilon)} = \underline{\theta} + j d_\Theta(0, \varepsilon) \tag{25}$$

for $j = 0, \ldots, N_\varepsilon - 1$, $N_\varepsilon = \lceil d_\Theta(\overline{\theta}, \underline{\theta}) / d_\Theta(0, \varepsilon) \rceil$. Clearly, $\Theta^{(\varepsilon)} \triangleq \{\theta_j^{(\varepsilon)} : j = 0, \ldots, N_\varepsilon - 1\}$ is an ε-net of Θ. With a slight abuse of notation, we shall sometimes refer to a parametric density g_θ as parameter θ, and to $s(g_\theta, y)$ as $s(\theta, y)$ in the sequel.

Definition 6 (Linear parameter flow). *A linear parameter flow f over $\Theta^{(\varepsilon)}$ is defined to be a map $f : \mathbb{N} \to \Theta^{(\varepsilon)}$, such that*

$$t \mapsto f(t) = \theta_{(\kappa + \zeta t) \bmod N_\varepsilon} \tag{26}$$

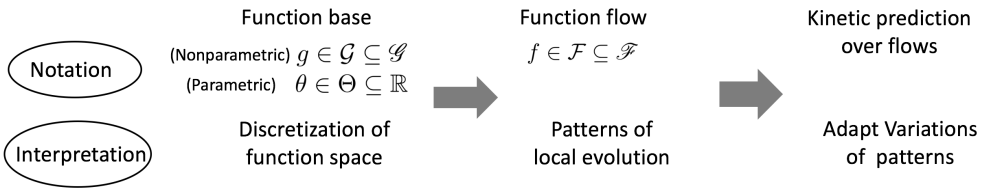

Fig. 7. Illustration of some notation in kinetic prediction.

where $\kappa \in \{0, 1, \ldots, N_\varepsilon - 1\}$ *and* $\zeta \in \{-\lfloor N_\varepsilon/2 \rfloor, \ldots, -1, 0, 1, \ldots, \lfloor N_\varepsilon/2 \rfloor\}$ *are constants that index* f*. The set of all the linear flows is denoted by* $\mathcal{F}$*.*

Suppose that ε is very small, then for any linear trend in some time epoch, its starting point and the slope can be well approximated by some θ_κ and $\zeta\varepsilon$, respectively. Moreover, if the path of true parameters is approximately segment-wise linear, we may be able to achieve optimal prediction by applying kinetic prediction over all the linear parameter flows. All the linear flows may be categorized into three types, illustrated in Fig. 8(a), (b), and (c). These respectively illustrate parameter base (constant flow), parameter flow from small to large values, and parameter flow from large to small values.

In view of the above set-up, the previously presented algorithms and theories can be extended in a straightforward manner. An example is given in Algorithm 4 which has a computational complexity $O(N) = O(\varepsilon^{-2})$ at each time t. Next, we provide performance guarantees for this algorithm.

Before we proceed to the theoretical results, we make the following assumption which slightly extends Assumption (2′).

Assumption 2′′. *The parameter space* $\Theta \subseteq \mathbb{R}^n$ *is a one-dimensional interval. For all* $\theta, \tilde{\theta} \in \Theta \subseteq \mathbb{R}^n$*, we have* $|s(\theta, Y_t) - s(\tilde{\theta}, Y_t)| \leq Z(Y_t) \cdot d_\Theta(\theta, \tilde{\theta})$ *for all* $t = 1, \ldots, T$*, where* $Z(Y_t)$ *is similarly defined as in Assumption (2′), and* $d_\Theta(\theta, \theta') \triangleq \|\theta - \tilde{\theta}\|_2^u$ *with* $u \in (0, 1]$ *being a fixed constant.*

In the sequel, we suppose that $0 = t_0 < t_1 < \cdots < t_{M_T} = T$, and define for any set of M_T linear functions

$$L_m : D_m \to \Theta, \quad D_m \triangleq \{t_{m-1} + 1, \ldots, t_m\}$$

$(m = 1, \ldots, M_T)$ the distance to the true data generating parameters

$$\Delta_T(L_1, \ldots, L_{M_T}, D_1, \ldots, D_{M_T}) \triangleq \sum_{m=1}^{M_t} \sum_{t \in D_m} d_\Theta(L_m(t), \theta_t^*).$$

We further define the minimal segment-wise linear approximation error as

$$\Delta_T \triangleq \min_{\substack{L_1, \ldots, L_{M_T}, \\ D_1, \ldots, D_{M_T}}} \Delta_T(L_1, \ldots, L_{M_T}, D_1, \ldots, D_{M_T}). \tag{28}$$

Note that Δ_T implicitly depends on M_T.

Theorem 4. *Assume that Assumptions (1), (2′′), and (3) hold.*

(i) Suppose that $\{f_1, \ldots, f_N\}$ *are linear flows on* $\mathcal{G}^{(\varepsilon)}$*, and that* $\beta > 0$ *is an arbitrarily chosen constant. If we choose* α *and* η *as in (14), then Algorithm 4 outputs* $\boldsymbol{p}_t : t = 1, \ldots, T$ *such that*

$$\sum_{t=1}^{T} \sum_{n=1}^{N_\varepsilon} p_{t,n}\ s(\theta_n^{(\varepsilon)}, Y_t) \leq \sum_{t=1}^{T} s(\theta_t^*, Y_t) + \sqrt{2^{-1} T^{1+2\beta} Q_{T,N}} + 2T^{2+\beta}\varepsilon + T^\beta \Delta_T \tag{29}$$

holds with probability at least $1 - C_1 T \exp(-C_2 T^\beta)$*, for some fixed constants* C_1 *and* C_2*.*

(ii) Moreover, if there exists a fixed constant $\beta \in (0, 0.2]$ *satisfying*

$$M_T = O(T^{1-5\beta}), \quad \Delta_T = o(T^{1-\beta}), \tag{30}$$

and we choose $\varepsilon = T^{-\nu}$ *for any fixed* $\nu > 1 + \beta$ *in (25), then*

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left\{ \sum_{n=1}^{N_\varepsilon} p_{t,n}\ s(\theta_n^{(\varepsilon)}, Y_t) - s(\theta_t^*, Y_t) \right\} \leq 0 \quad a.s. \tag{31}$$

(iii) Suppose that $s(g_\theta, y)$ *is convex in* θ*, and we use* $\hat{\theta}_t = \sum_{n=1}^{N_\varepsilon} p_{t,n} \theta_n^{(\varepsilon)}$ *for prediction at time* t*. Then*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left\{ s(\hat{\theta}_t, Y_t) - s(\theta_t^*, Y_t) \right\} = 0 \quad a.s.$$

(iv) Suppose that we independently generate J_t *from multinomial distribution with probability* $\boldsymbol{p}_t$*, and use* $f_{\theta_{J_t}}$ *for prediction at each time* t*. Then*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left\{ s(\theta_{J_t}^{(\varepsilon)}, Y_t) - s(\theta_t^*, Y_t) \right\} = 0 \quad a.s. \tag{32}$$

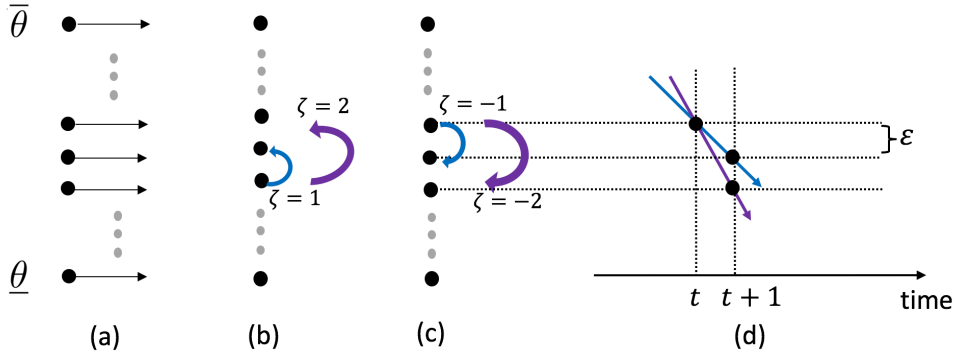

Fig. 8. Illustration of three types of linear flows $t \mapsto f(t) = \theta_{(\kappa+\zeta t) \bmod N_\varepsilon}$: (a) $\zeta = 0$, (b) $\zeta > 0$, (c) $\zeta < 0$, and (d) different slopes.

Algorithm 4 Sequential prediction for time series (with 1-D parameter space)

input $\{y_t : t = 1, \ldots, T\}$, $\varepsilon > 0$, $\eta > 0$, $\alpha \in [0, 1]$
output $\{\boldsymbol{p}_t : t = 1, \ldots, T\}$ (predictive weights over the function bases)
1: **Initialize** $N_\varepsilon = \lfloor (b-a)/\varepsilon \rfloor$ (where $[a, b]$ is the parameter space), $\boldsymbol{w}_0 = [1/N_\varepsilon, \ldots, 1/N_\varepsilon]^{\mathrm{T}}$, $N = N_\varepsilon + 2N_\varepsilon \lfloor N_\varepsilon/2 \rfloor$, $\boldsymbol{w}_0^{\mathscr{F}} = [1/N, \ldots, 1/N]^{\mathrm{T}}$
2: Let $f_1, \ldots, f_N$ be the N linear parameter flows.
3: **for** $t = 1 \to T$ **do**
4: Normalize the predictive weights $p_{t,n} = (\sum_{j=1}^{N_\varepsilon} w_{t-1,j})^{-1} w_{t-1,n}$, $n = 1, \ldots, N_\varepsilon$.
5: Obtain y_t and compute $v_{t,n} = w_{t-1,n}^{\mathscr{F}} \exp\{-\eta\, s(f_n(t), y_t)\}$ for each $n = 1, \ldots, N$.
6: Let $w_{t,n}^{\mathscr{F}} = (1-\alpha) v_{t,n} + \alpha N^{-1} \sum_{j=1}^{N} v_{t,j}$.
7: Aggregate the predictive weights from $\boldsymbol{w}_t^{\mathscr{F}}$ to $\boldsymbol{w}_t$, namely for each $n = 1, \ldots, N_\varepsilon$,

$$w_{t,n} = \sum_{j \in A(n)} w_{t,j}^{\mathscr{F}}, \text{ where } A(n) = \{j : 1 \le j \le N, f_j(t) = \theta_n^{(\varepsilon)} \text{ (the } n\text{th parameter base) } \}. \tag{27}$$

8: **end for**

Remark 6. *We emphasize that N_ε is the number of bases while N is the number of flows, so $N_\varepsilon \neq N$, and their relation is described in Line 1 of Algorithm 4. Theorem 4 can be regarded as an extension of Theorem 3. We note that in the above definition of D_m, it is possible for D_m to consist of a single point t_m as a degenerate linear function). In general, smooth variations of parameters (or parametric densities) can be regarded as concatenation of parameter flows with some combinatorial structures.*

Remark 7 (Infill Asymptotics). *In the above theorem, assumption (30) only requires that Δ_T is sublinear in T. In this section, we provide a wide class of the true parameter paths $\{\theta_t^* : t = 1, \ldots, T\}$ that satisfies the required assumption. We consider the following characterization. Suppose that $\Lambda : [0, 1] \to \Theta$ is a continuous function, and*

$$\theta_t^* = \Lambda(t/T). \tag{33}$$

In other words, Λ approximately contains all the time evolution of true parameters from 1 to T. As T tends to infinity, for a any given $u_0 \in (0, 1)$, more and more data generated from $u_t = t/T \in [u_0 - \varepsilon, u_0 + \varepsilon]$ close to u_0 are observed. This type of "infill statistics" has been used before in the study of a rigorous asymptotic treatment of locally stationary processes [29], [53].

It has been well known that smooth functions can be well approximated by B-splines [54]. Let $[0, 1]$ be partitioned into J equal-sized intervals $\{I_j\}_{j=1}^J$. Let $\mathfrak{F}$ denote the space of polynomial splines of degree $\ell \ge 1$, consisting of functions $g(\cdot)$ satisfying 1) the restriction of $g(\cdot)$ to each interval is a polynomial of degree ℓ, and 2) $g(\cdot) \in C^{\ell-1}[a, b]$ ($\ell - 1$ times continuously differentiable). There exists a normalized B-spline basis $\{b_j\}_{j=1}^{J+\ell}$ for $\mathfrak{F}$. Suppose each considered (non)linear function f has kth derivative, $f^{(k)}$, and satisfies the Holder condition with exponent ρ: $|f^{(k)}(x) - f^{(k)}(x')| \le M|x - x'|^\rho$ for all $x, x' \in [0, 1]$, where $k \le \ell$ is a nonnegative integer, $\rho \in (0, 1]$ such that $k + \rho > 0.5$, and $M > 0$. Define the norm $\|f\|_\infty = \sup_{u \in [0,1]} |f(u)|$ for a any continuous function $f : [0, 1] \to \mathbb{R}$. Standard results on splines imply that there exists $F(\cdot) \in \mathfrak{F}$ with $J + 1$ B-spline basis such that

$$\|\Lambda - F\|_\infty = O\big(J^{-(\ell+\rho)}\big).$$

Then the key assumption (30) required by Theorem 4 is satisfied. Let $J = M_T$. It is easy to prove that if M_T

is chosen such that

$$\frac{M_T}{T^{1-5\beta}} = O(1), \quad \frac{TM_T^{-(1+\rho)}}{T^{1-\beta}} = o(1) \tag{34}$$

then the key assumption (30) required by Theorem 4 is satisfied. It suffices to choose $M = T^w$, with $\beta/(1+\rho) < w \leq 1 - 5\beta$. Clearly, a valid w can always be found by choosing small enough β. In other words, as long as nature generates the true parameter sequence from a smooth function in an "infill" manner, assumption (30) is easily satisfied. It is not difficult to show that (33) can be extended to

$$\theta_t^* = \Lambda(t/T) + \zeta_t$$

with ζ_t satisfying $T^\beta \sup_{t=1,\ldots,T} \zeta_t \to 0$ as $T \to \infty$.

VI. Numerical Experiments

In this section, we present numerical experiments to demonstrate our theoretical results and wide applicability of the proposed methodology using both synthetic and real-world data. Throughout the experiments, we fix the following tuning parameters. In view of Theorem 2 and Theorem 4, we set $\varepsilon = T^{-1/3}$ for Algorithm 1, $\varepsilon = |\Theta|T^{-\nu}$ ($\nu = 5/4$) for Algorithm 4; We use $M_T = \lfloor T^{1/3} \rfloor$, $\beta = 1/8$ for both algorithms, and α, η accordingly defined in (14). In view of Corollary 1, we use $\eta = 10T^{-1/3}$ and different α's in Algorithm 3. Here, $|\Theta|$ denotes the Lebesgue measure of the parameter space (for parametric models), and it is used to enhance the predictive performance under limited data (according to our empirical studies). In practice, a data analyst may achieve better predictive performance of kinetic prediction by fine tuning the above parameters for any particular dataset.

A. Synthetic data experiment: abrupt changes in mean

In this subsection, we consider prediction under abrupt changes in mean for random Gaussian observations, introduced in Examples 1 and 2. In both examples, our kinetic prediction was realized by running Algorithm 3. For comparison, we implemented two "first-infer-then-predict" change detection algorithm, one by employing quadratic loss and BIC [18] into a multi-window detection method (denoted as MW), the other one by using CUSUM. Since the CUSUM was originally designed for detecting a single change, we adapted it to handle multiple changes in a way similar to multiple change point algorithm. In particular, in an online fashion whenever a change point is detected, we reset the CUSUM algorithm and make prediction based on the latest detected segment. We shall also compare with the standard Bayesian procedure (which corresponds to $\eta = 1$ and $\alpha = 0$).

Recall that in Example 1, data are generated according to a Gaussian distribution with unit variance and a changing mean parameter at different time. We set $T = 500$, the true mean sequence $\theta_t^* = 5$ for $1 \leq t \leq T/2$, and $\theta_t^* = 6$ for $T/2 < t \leq T$. We run Algorithm 3 for kinetic prediction with the parameter space $[4, 7]$ and $\alpha = 1/(T-1)$. The score function is set to be $s(\theta, y) = (y-\theta)^2/2$, which is the logarithmic scoring rule for the Gaussian data. To predict at time t using the change detection technique, we first detect changes on the batch data $Y_{1:t-1}$, and then use the estimated mean in the last segment as the predicted value for time t. The predicted parameter at $t = 1$ is randomly drawn from the parameter space since there is no observation yet. Fig. 9 shows the predicted mean sequences from different prediction methods, along with the true mean sequence, for a single realization of data shown in Fig. 1. Since Algorithm 3 generates a predictive distribution instead of a single value for the unknown parameter at each time step, we use the expected value of that distribution as our predictor in the figure. As we can see, both kinetic prediction and classical change detection can predict well given one abrupt change in the unknown parameter sequence, while kinetic prediction behaves more smoothly because it has long memories making itself more robust to data outliers. The standard Bayesian updating procedure fails to follow the true sequence after the abrupt change, and its predictor converges to $(\theta_1^* + \theta_T^*)/2$. In order to further compare our kinetic prediction and the change detection method in terms of cumulative scores, we repeat the experiment for 10 times for different random observation sequences, and evaluate the "average additional score over the oracle" given by the left-hand-side term in (5), for $t = 1, \ldots, T = 500$. Here, we only show results for the MW change detection method for better visualization because its performance is close to the adapted CUSUM. The average performance from the repeated experiments is summarized in Fig. 10, which shows that the average additional score goes to zero for each prediction method, and the kinetic prediction outperforms the change detection scheme in that it generally attains a lower score with less variance.

In Example 2, instead of only one change point in the Gaussian mean, the number of abrupt changes in the true parameter sequence is set to be $\lfloor T^{(1/3)} \rfloor$ for a given time T. The locations of the change points are uniformly random between $[2, \ldots, T]$, and the true mean on each constant segment is uniformly generated from the interval $[5, 6]$. Algorithm 3 is run for kinetic prediction with parameter space $[5, 6]$, $\alpha = \lfloor T^{(1/3)} \rfloor/(T-1)$, and the logarithmic scoring rule. The standard Bayesian updating is also run by Algorithm 3 with $\eta = 1$ and $\alpha = 0$. Fig. 11 shows the histograms of the sampled particles, at certain time steps, for a single realization of

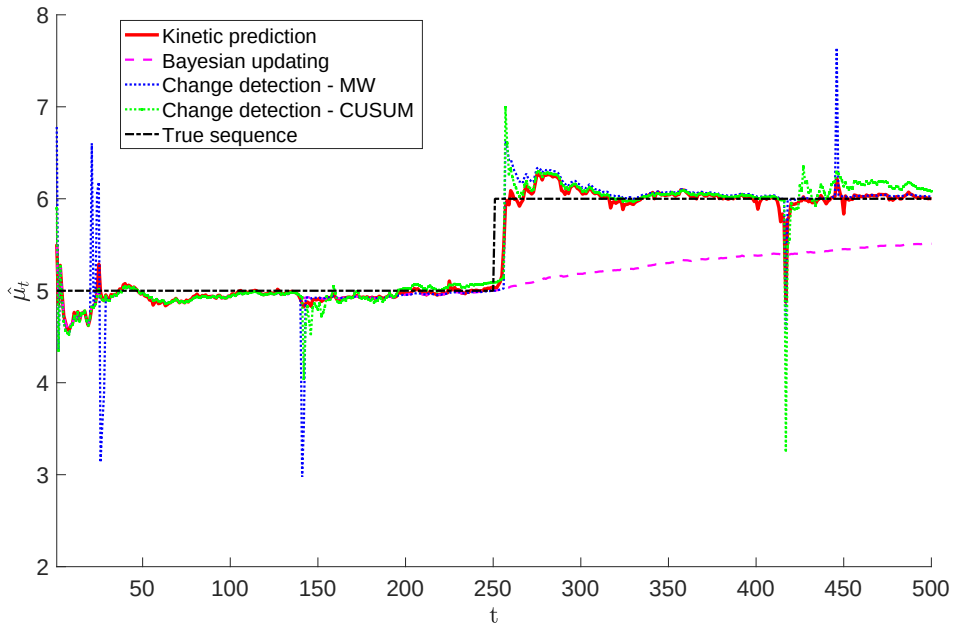

Fig. 9. Synthetic data experiment in Subsection VI-A, Example 1: comparison of different methods in sequentially predicting the Gaussian mean sequence with one abrupt change point.

the parameter and observation sequence with $T = 500$ (as shown in Fig. 2). As we can see, the predictive distributions from the kinetic prediction can quickly concentrate around the location of the true parameter after abrupt changes, while the standard Bayesian update fails to do so. We also repeat the experiments for 10 times for each $T = 50, 100, 150, 250, 500$, and plot the average additional score over the oracle for each T for both the kinetic prediction and MW change detection method in Fig. 12. Our kinetic prediction is comparable to the change detection methods in this experiment.

B. Synthetic data experiment: abrupt changes in nonparametric densities

The purpose of this experiment is to demonstrate the application of kinetic prediction to nonparametric model classes, where densities are not necessarily differentiable and are only known to belong to the Lipschitz class. We generate synthetic data from Example 3. A realization of data was shown in Fig. 3. Note that in this example, Assumption (2) is satisfied under logarithmic scoring rule, due to the elementary inequality

$$|\log x - \log y| \le \frac{|x-y|}{\min\{x,y\}}$$

for any two positive numbers x and y. By implementing Algorithm 1 and the strategy discussed in Example 6, we obtain Fig. 13 which shows the sequential predictive weights and snapshots of the predictors in Theorem 2(iii) at time $t = 250, 500$. As we can see from the plot, the distributional changes in predictive weights mostly capture the true change points (especially around $t = 250$ when there are sufficient data before and after that change). To show the convergence, we also repeat the experiments for different T's ($T = 50, 100, 150, 250, 500$) and estimate the difference between the average score of our predictor and the oracle (i.e. the left-hand-side term in (7)). The estimates and their standard errors are shown in Fig. 14. As our theory expects, the difference goes to zero as T becomes large.

C. Synthetic data experiment: abrupt changes and smooth variations in mean

The purpose of this experiment is to demonstrate the application of kinetic prediction to parametric models with both abrupt changes and smooth variations. We consider the independent Gaussian model $Y_t \sim \mathcal{N}(\mu_t, 1)$, with parameter space $\Theta = [-10, 10]$. The time-varying means μ_t's consist of four segments: first a quadratic trend, then two linear trends with different slopes, followed by a cosine pattern. Each switch from one segment to another is abrupt. The mathematical formula describing the changes are

$$\mu_t = \begin{cases} 5 - T^{-2}(t - T_1/2)^2 & \text{if } t \le T_1 \\ -7 + 60(t - T_1)/T & \text{if } T_1 < t \le T_2 \\ -7 + 20(t - T_2)/T & \text{if } T_2 < t \le T_3 \\ 2 + 5\sin\{6\pi(t - T_3)/(T - T_3)\} & \text{otherwise} \end{cases}$$

where $T_k \stackrel{\Delta}{=} \lfloor kT/4 \rfloor$ for $k = 1, 2, 3$.

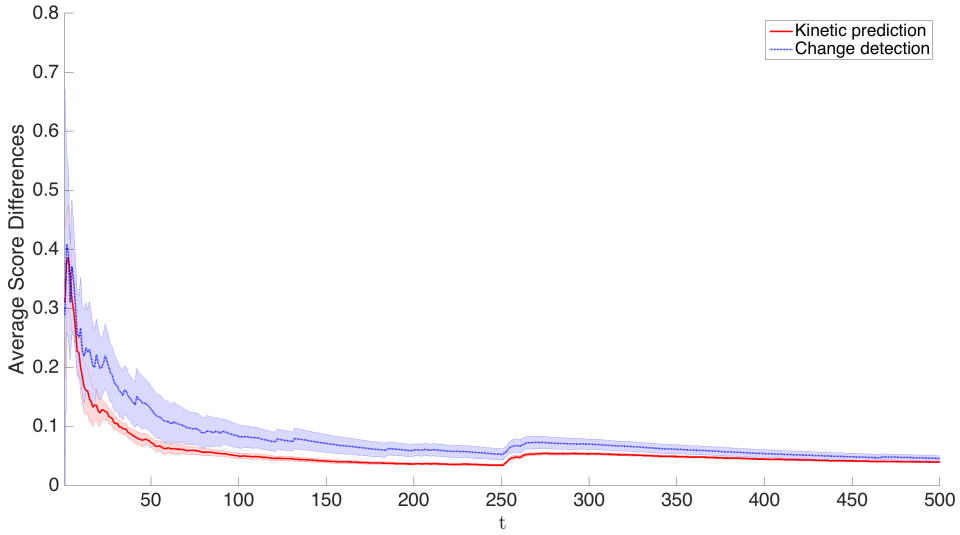

Fig. 10. Synthetic data experiment in Subsection VI-A, Example 1: comparison of the kinetic predictor and a change-detection predictor on the average additional score over the oracle. The result is the average of 10 repeated experiments and the shaded region describes the $+1/-1$ standard errors.

We used quadratic score $(Y_t - \hat{\theta}_t)^2/2$ to evaluate predictive performance in the algorithm. The multiplier $1/2$ is not theoretically essential, but it has the natural interpretation of negative log-likelihood of Gaussian observations with unit variance. By implementing Algorithm 4, we obtain Fig. 15(a) which shows the sequential predictive weights. As we can see from the plot, our predictive weights capture the true trends, and switch promptly after abrupt changes occur. Fig. 15(b) plots our predictor versus true mean at each time step, along with the realization of data. The prediction well matches the truth in general. We also repeat the experiments for different T's ($T = 50, 100, 150, 250, 500$) and estimate the difference between the average score of our predictor and the oracle (i.e. the left term in (5)) at the time step $t = T$. The estimates and their standard errors are plotted in Fig. 16, showing the convergence to zero.

D. Synthetic data experiment: abrupt changes and smooth variations in time-varying autoregression

The purpose of this experiment is to demonstrate the application of kinetic prediction to non-i.i.d. data. We consider the time-varying autoregression in Example 4. A realization of data was shown in Fig. 4. By implementing Algorithm 4 with the parameter space $\Theta = [-1, 1]$, we obtain Fig. 17(a) which shows the sequential predictive weights at each time step. As we can see from the plot, the masses of our predictive weights concentrate along the true flow of parameters (autoregressive coefficients) for $t = 1, \ldots, T/2$. The true parameter abruptly switches to the constant 0.8 for $t > T/2$, and our predictive weights successfully capture the change shortly afterwards. The catch-up to 0.8 is hesitating at the beginning; but when there are sufficient data as the evidence of switch, the masses concentrate more on 0.8 as t grows large. Fig. 17(b) shows the true and predicted conditional means at each time step. We can see that the prediction matches the truth very well.

We also repeat the experiments for different T's ($T = 50, 100, 150, 250, 500$) and estimate the difference between the average score of our predictor and the oracle (i.e. the left-hand-side term in (8)). The estimates and their standard errors are shown in Fig. 18. As our theory expects, the difference goes to zero as T becomes large.

E. Real data experiment: discovering time-varying cointegration

The purpose of this experiment is to apply kinetic prediction proposed in Section IV to discover time-varying cointegration [55], [56] in financial applications. The panel data we used (denoted by $\boldsymbol{Y}$) consist of 390 stock prices of Apple Inc. (denoted by $Y_{1,t}$) and Alphabet Inc. Class A (denoted by $Y_{2,t}$), collected every half hour starting from November 17, 2015. The data were standardized using a mean and standard deviation that were calculated using historical data collected before that date. We note that linear transformations of raw data do not cause essential difference in Algorithms 2 and 3, but they make it easier to plot. The pre-processed data are shown in Fig. 19(a). Each of $Y_{1,t}, Y_{2,t}$ is a process integrated of order 1, according to the augmented Dickey-Fuller test under 0.01 significance level.

Suppose that there exist two deterministic series $a_{0,t}, a_{1,t}$ with a few unknown abrupt changes, such that

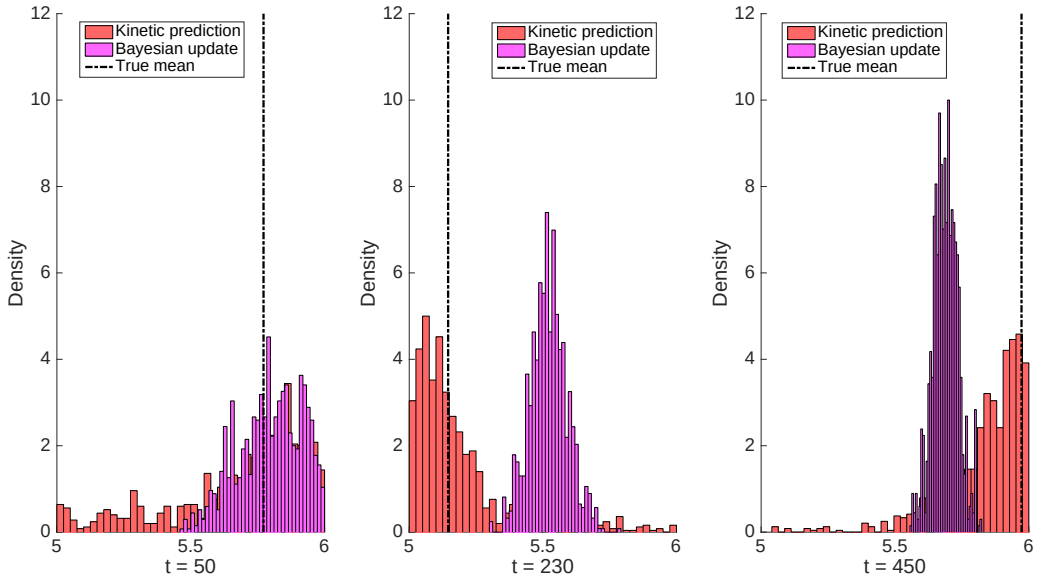

Fig. 11. Synthetic data experiment in Subsection VI-A, Example 2: predictive distribution of the Gaussian mean from kinetic prediction and ordinary Bayesian updating at some time steps, along with the true mean marked by the black dash-dot line.

$Y_{2,t} - (a_{0,t} + a_{1,t}Y_{1,t})$ is stationary. This time-varying cointegration may be more reasonable than a fixed-parameter cointegration, considering potential regime changes in the financial market. We are primarily interested in a sequential context, where the estimates of $a_{0,t}, a_{1,t}$ depend only on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{t-1}$, and we hope that "the predictive residues"

$$R_t \triangleq Y_{2,t} - (\hat{a}_{0,t} + \hat{a}_{1,t}Y_{1,t}) \tag{35}$$

become stationary. We first apply Algorithm 3 with the quadratic score $s : (\theta, y_t) \mapsto (y_{2,t} - a_0 - a_1 y_{1,t})^2/2$, parameter space $a_{0,t}, a_{1,t} \in [-2, 2]$, and default tuning parameters as described at the beginning of Section VI. The means and standard deviations of parameters computed from our predictive distribution at each time step are plotted in Fig. 19(c)&(d). We use those means as $\hat{a}_{0,t}, \hat{a}_{1,t}$ and compute the predictive residues in (35), and plot the sequence in Fig. 19(b) (abbreviated as "Kinetic"). For comparison, we also apply classical recursive least squares method to estimate $\hat{a}_{0,t}, \hat{a}_{1,t}$, and plot the corresponding predictive residues in Fig. 19(b) (abbreviated as "RLS"). For numerical stability, all the results are computed starting from $t = 6$. From Fig. 19, we can see that the parameters undergo significant regime changes over time, especially for the intercept term. The residue sequence given by "Kinetic" is stationary, while that given by "RLS" is not, according to the Dickey-Fuller test under 0.01 significance level. Visually, the former sequence does look more stationary as they fluctuate more frequently around zero (marked by a black dash).

We plot the average score $\sum_{i=1}^{t} R_i^2/t$ given by two methods at each time step t given that $T = 390$ in Fig. 20(a). We also compute the average score at the last time step, namely $\sum_{t=1}^{T} R_t^2/T$ for $T = 50, 100, 150, 250, 390$ (with Algorithm 3 running on corresponding tuning parameters), and plot the score of "RLS" minus that of "Kinetic" in Fig. 20(b). The results show that "Kinetic" is dominantly better than "RLS" method also in terms of scores.

F. Real data experiment: predicting stochastic volatilities

The purpose of this experiment is to apply kinetic prediction to predict stochastic volatilities in a financial market. Forecasting volatility plays an important role in risk management and asset allocation. To model financial time series with time-varying volatility, the autoregressive conditional heteroskedasticity (ARCH) models [57] and the generalized ARCH (GARCH) [58] have been commonly adopted. We refer to [59] for an extensive literature on stochastic volatility models.

We collected 500 daily stock prices of SPDR S&P 500 ETF (SPDR), denoted by s_t, from Jan 6, 2007 to 31 Dec 08. Let $Y_t = 10^2 \log(s_t/s_{t-1})$ be the so-called log-returns (where the scaling is done for numerical convenience). The squared log-returns are shown in Fig. 21(a) in black dash. We note that the squared return on an asset at one time step (assuming a zero mean return) can be interpreted as a conditionally unbiased estimator of the

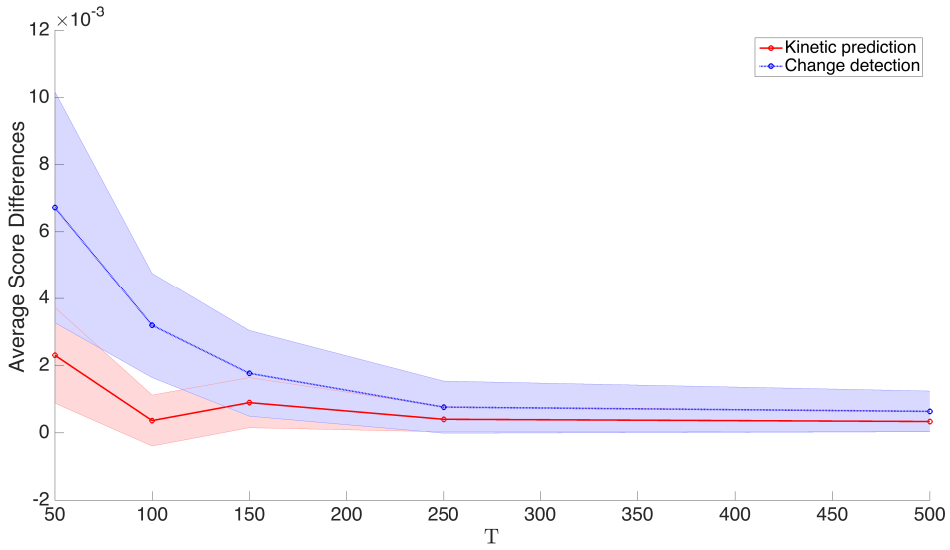

Fig. 12. Synthetic data experiment in Subsection VI-A, Example 2: comparison of the kinetic predictor and a change-detection predictor on the average additional score over the oracle, where each point was computed as the average of 10 repeated experiments and the shaded region describes the $+1/-1$ standard errors.

true but unobserved conditional variance of the asset. After some exploratory studies, we adopt the following GARCH(1,1) as the parametric model class,

$$v_t^2 = a_0 + a_1 Y_{t-1}^2 + b_1 v_{t-1}^2, \quad Y_t = v_t e_t, \tag{36}$$

where e_t's are i.i.d. $\mathcal{N}(0,1)$ noises. In other words, the true (but unobservable) volatility v_t at each time step t only depends on the squared observation Y_{t-1} and the true volatility in the last time step. We used negative log-likelihood as the score function. In other words, $s : (\theta, y_t) \mapsto (\log h_t + h_t^{-1} y_t^2)/2$ is used, where h_t is the predicted volatility at time step t corresponding to $\theta \triangleq [a_0, a_1, b_1]$. Here, h_t can be recursively computed using (36), θ, and $v_1 \triangleq Y_1^2$. Rigorously speaking, the above score is predictive log-likelihood, since h_t is computed using all the data before time step t in a sequential setting [60]. It is worth noting that in the volatility forecasting literature, it is also common to write a score function in the form of $L(v_t^2, h_t)$, and use volatility proxies such as the squared return Y_t^2 in place of v_t^2. Our scoring function is equivalent to the so called "QLIKE" loss functions, which is proved to be robust in the sense of [61, Definition 1]. Squared score in the form of $(Y_t^2 - h_t)^2$ is also commonly used, but it is more sensitive to extreme observations and the level of volatility of returns. For other widely-used score functions used to evaluate conditional variance forecasting, we refer to [61] and the references therein.

We suppose that the true parameters $a_{0,t}, a_{1,t}, b_{1,t}$ at each time step t may not be all the same (e.g. during a financial crisis). We first apply Algorithm 3 with parameter space $a_{0,t}, a_{1,t}, b_{1,t} \in [0,1]$ (which is standard for GARCH), and default tuning parameters as described at the beginning of Section VI. The means and standard deviations of parameters computed from our predictive distribution at each time step are plotted in Fig. 21(b)(c)(d). We use those means to compute the predictive volatilities, and plot the sequence in Fig. 21(a) (abbreviated as "Kinetic"). For comparison, we also apply a fixed-parameter GARCH(1,1) to estimate the parameters at each time t, using $y_1, \ldots, y_{t-1}$. We then plot the corresponding predictive volatilities in Fig. 21(a) (abbreviated as "GARCH").

We plot the average scores of two methods at each time step t given $T = 500$ in Fig. 22(a). We also compute the average score at the last time step $t = T$, for $T = 50, 100, 150, 250, 500$ (with Algorithm 3 running on corresponding tuning parameters), and plot the scores of "GARCH" minus those of "Kinetic" in Fig. 22(b). The results show that "Kinetic" is dominantly better than "GARCH" method in terms of scores.

One anonymous reviewer pointed out an interesting observation that the difference between the average scores of our method and GARCH seems to decrease to zero (from Fig. 22), while the gap between our method and RLS (from Fig. 20) does not decrease to zero. An insight on why GARCH closes the gap for large T comes from Fig. 21(b)(c)(d), which indicates that there exists no significant changes in parameters after around $t = 50$. As a result, the average performance of GARCH becomes closer to the proposed method as time elapses. Likewise, Fig. 19(c)(d) indicates multiple abrupt changes throughout $t = 1, \ldots, 390$, thus the proposed method is constantly much better than the RLS method as shown

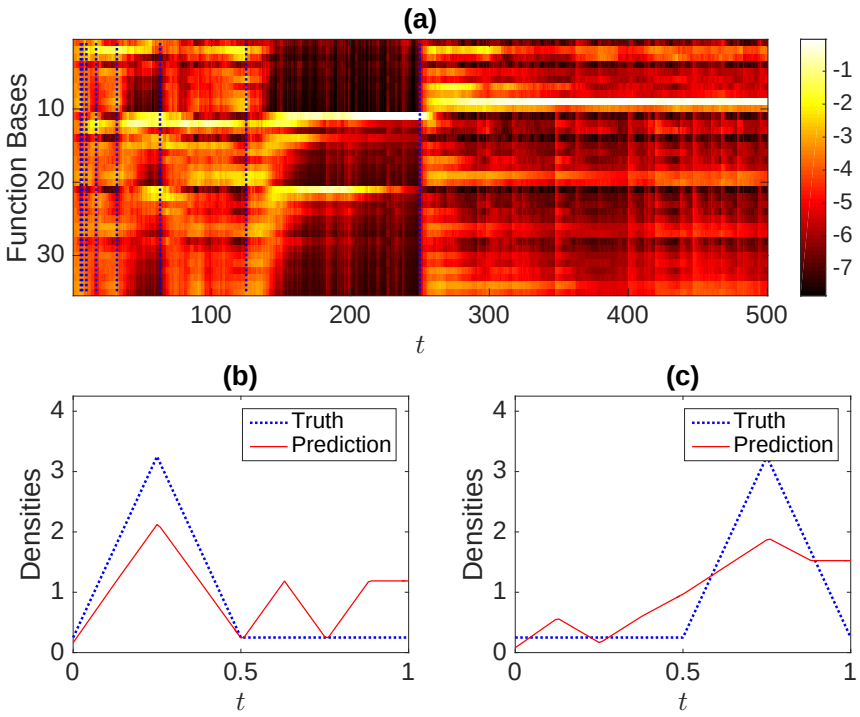

Fig. 13. Synthetic data experiment in Subsection VI-B: (a) the heat map showing the sequential predictive weights (in log scale) over the function bases at each time step, along with the true locations of abrupt changes marked in blue dashes, (b)&(c) true and predicted density functions at $t = 250$ and $t = 500$.

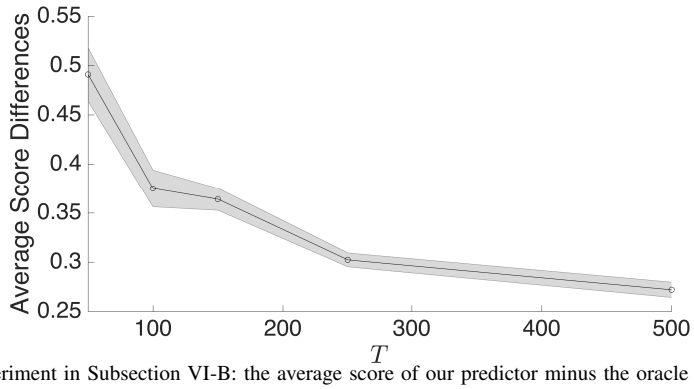

Fig. 14. Synthetic data experiment in Subsection VI-B: the average score of our predictor minus the oracle for different T's, where each point was computed from 10 repeated experiments and the shaded region describes the $+1/-1$ standard errors.

in Fig. 20.

VII. Conclusion

It is common to arrange economic data in the form of a time series (which must be sequentially analyzed), with the eventual goal of optimally predicting future values. To handle potential unknown abrupt changes and smooth variations in the underlying data generating processes, we proposed a novel methodology to approach the oracle predictive performance. The general idea is to first apply an ε-net, and then properly update the predictive weights over function bases. For the parametric case, we also proposed a generic algorithm that directly runs over a continuous parameter space without the need of discretization. Moreover, to capture frequent drifts in parameters, we extended the above methodology by proposing the concept of "function flows". The idea is that smooth changes can usually be approximated by many locally deterministic trends, such as linear functions (in time t). Our methodology is applicable to a wide range of model classes and scoring rules. We hope this work sheds some new light on the relation between inference and prediction in general as well.

Future work can be done in the following directions. 1) When we were concerned with parametric models, the parameter space is assumed to be compact. In practice,

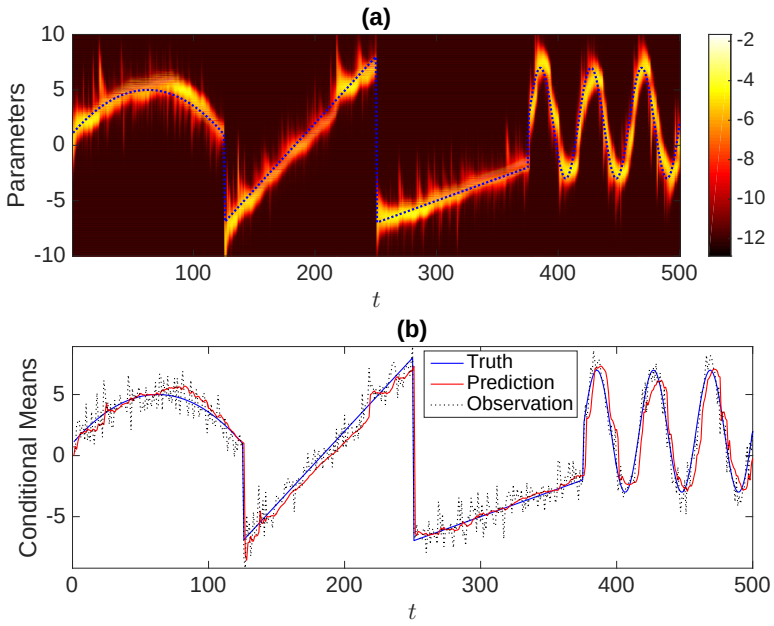

Fig. 15. Synthetic data experiment in Subsection VI-C: (a) the heat map showing the sequential predictive weights (in log scale) over the parameter bases, along with the true parameters at each time step marked in blue dashes, and (b) the true and predicted mean at each time step, along with the observed data.

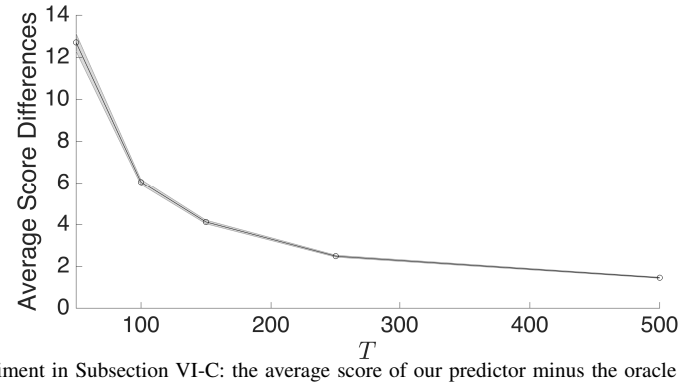

Fig. 16. Synthetic data experiment in Subsection VI-C: the average score of our predictor minus the oracle for different T's, where each point was computed from 10 repeated experiments and the shaded region describes the $+1/-1$ standard errors.

the parameter space can be unbounded, or not known in advance. For unbounded parameter space, a possible solution is to transform parameters to a bounded range, and then apply the proposed technique. For unknown parameter space, we may perform parameter estimation from repetitive subsampling of data, and form an empirical parameter space from the estimated parameters. Another general solution is to adaptively expand the current parameter space as more data are observed. The main challenge lies in the characterization of optimal bias-variance tradeoff. 2) We believe that the existing performance bounds can be tighter for high-dimensional parameter space, given prior knowledge such as sparsity in parameter components, or patterned switching of change points (e.g. seasonality or Markovity).

APPENDIX

APPENDIX: TECHNICAL PROOFS

Proofs for Section II

Proof of Proposition 1

Proof: The left-hand-side term in (5) is

$$\frac{1}{T}\sum_{t=1}^{T}(\theta_t^*-\hat{\theta}_t)^2+\frac{1}{T}\sum_{t=1}^{T}2(\theta_t^*-\hat{\theta}_t)e_t \qquad \text{(A.37)}$$

Since $\{2(\theta_t^*-\hat{\theta}_t)e_t\}$ is a martingale difference sequence with bounded variance, we obtain $T^{-1}\sum_{t=1}^{T}2(\theta_t^*-\hat{\theta}_t)e_t \to_{a.s.} 0$ by a martingale convergence theorem [62,

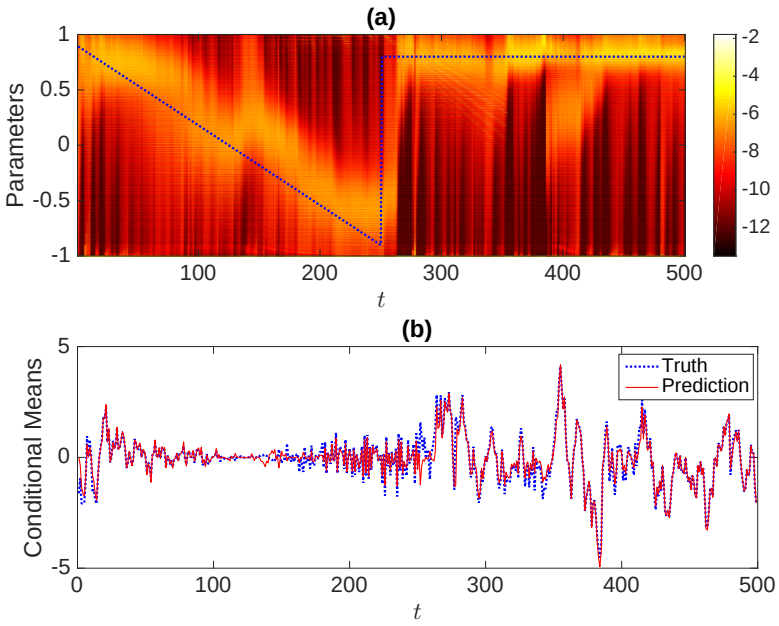

Fig. 17. Synthetic data experiment in Subsection VI-D: (a) the heat map showing the sequential predictive weights (in log scale) over the parameter bases, along with the true parameters at each time step marked in blue dashes, and (b) the true and predicted conditional mean at each time step.

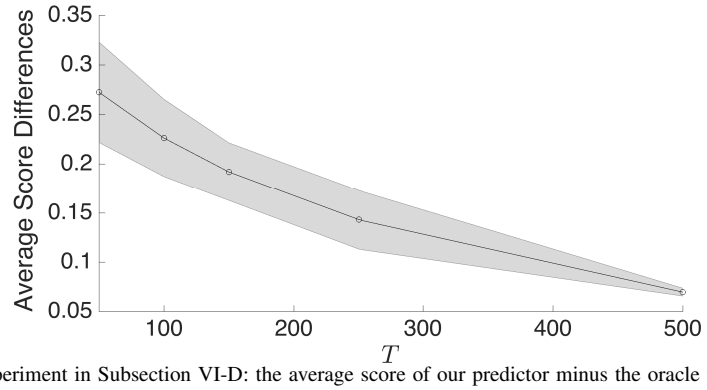

Fig. 18. Synthetic data experiment in Subsection VI-D: the average score of our predictor minus the oracle for different T's, where each point was computed from 10 repeated experiments and the shaded region describes the $+1/-1$ standard errors.

Theorem 1]. It remains to prove that

$$\liminf_{T\to\infty} T^{-1}\sum_{t=1}^{T}(\theta_t^*-\hat{\theta}_t)^2>0, \quad a.s. \tag{A.38}$$

We arbitrarily choose a fixed $\delta\in(0,|\theta_1-\theta_2|/2)$. We let $T_1=\lfloor T/2\rfloor$ for simplicity. By the law of large numbers, for all sufficiently large T, we have for all $t=\lfloor 3T/4\rfloor,\ldots,T$

$$\left|\frac{1}{T_1}\sum_{i=1}^{T_1}Y_i-\theta_1\right|<\delta, \text{ and}$$
$$\left|\frac{1}{t-T_1}\sum_{i=T_1+1}^{t}Y_i-\theta_2\right|<\delta, \quad a.s. \tag{A.39}$$

By direct calculations and the triangle inequality, (A.39) implies that

$$|\hat{\theta}_t-\theta_2|=\left|\frac{1}{2}(\theta_1-\theta_2)+\frac{T_1}{T}\frac{1}{T_1}\sum_{i=1}^{T_1}(Y_i-\theta_1)+\right.$$
$$\left.\frac{t-T_1}{T}\frac{1}{t-T_1}\sum_{i=T_1+1}^{t}(Y_i-\theta_2)\right|\geq\frac{1}{2}|\theta_1-\theta_2|-\delta.$$

Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}(\theta_t^*-\hat{\theta}_t)^2\geq\frac{1}{T}\sum_{t=\lfloor 3T/4\rfloor}^{T}(\theta_2-\hat{\theta}_t)^2$$
$$\geq\frac{1}{4}\left(\frac{1}{2}|\theta_1-\theta_2|-\delta\right)$$

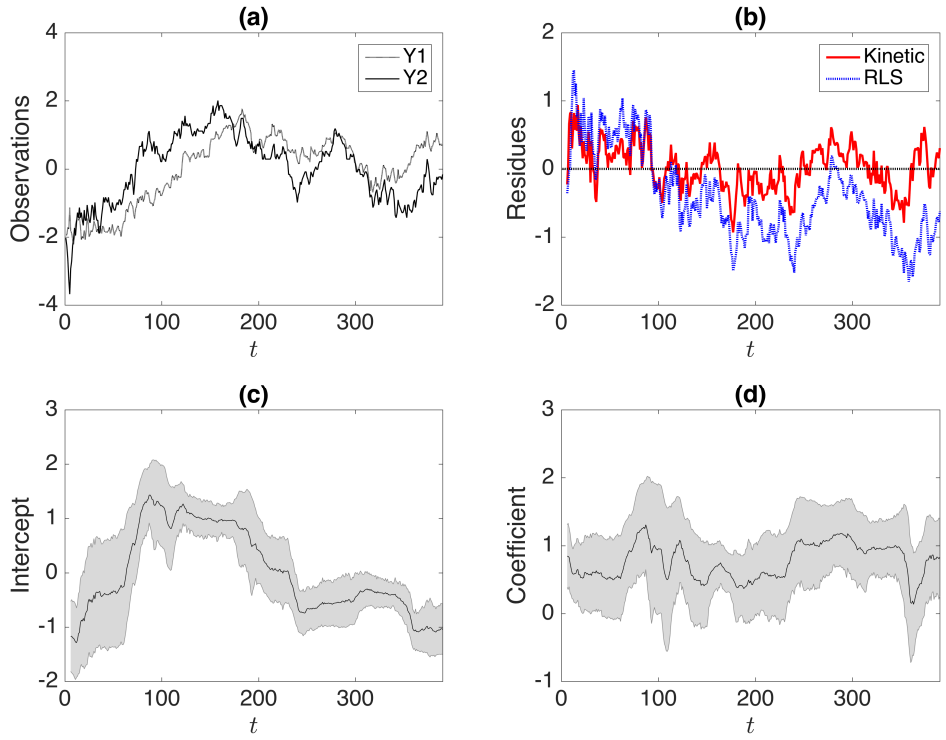

Fig. 19. Real data experiment in Subsection VI-E: (a) the preprocessed observations, (b) the cointegration residues of two prediction methods, (c)&(d) the sequential predictive means of the intercept a_0 and coefficient a_1 in cointegration, along with their +1/-1 standard deviations.

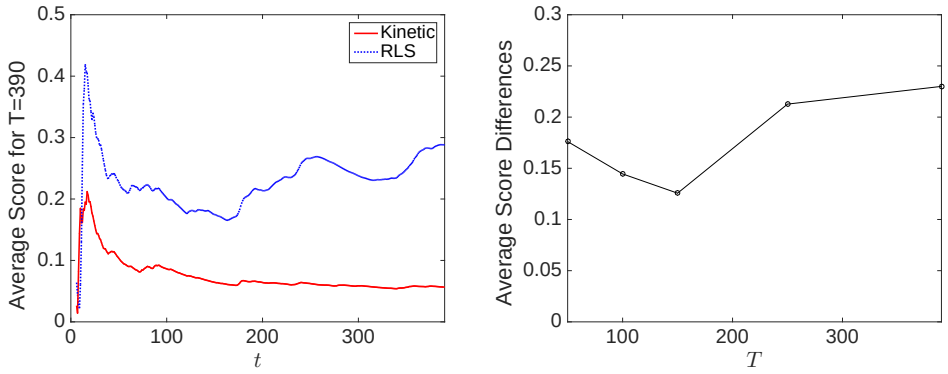

Fig. 20. Real data experiment in Subsection VI-E: (a) the average scores of two methods at each time step t for fixed $T = 390$, (b) the average score of the recursive least squares method minus that of the kinetic method for different T's.

for all sufficiently large T almost surely. This implies (A.38). ■

Proof of Theorem 1

Proof:

We first prove that for an arbitrarily given sequence of data generating densities $g_t \in \mathcal{G}$, $t = 1, \ldots, T$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\Big[s(g_t, Y_t) - \mathbb{E}_*\{s(g_t, Y_t) \mid Y_{1:t-1}\}\Big] \to_{a.s.} 0, \tag{A.40}$$

where $\mathbb{E}_*$ refers to the expectation with respect to the true joint distribution of $Y_1, \ldots, Y_T$. We note that the conditioning on $Y_{1:t-1}$ is to emphasize the potential dependency of data (see Remark 2). A more rigorous notation would replace $s(g_t, Y_t)$ with $s(g_t(\cdot \mid Y_{1:t-1}), Y_t)$.

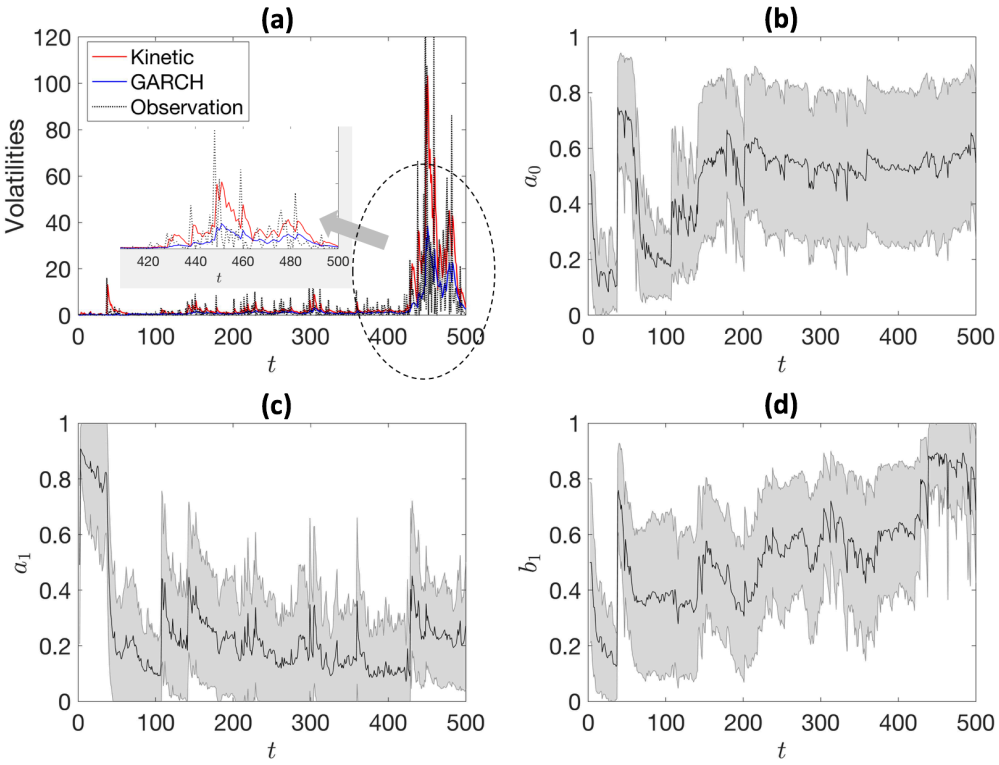

Fig. 21. Real data experiment in Subsection VI-F: (a) the predicted volatilities by kinetic and recursive GARCH methods, along with the squared log returns, (b)-(d) the sequential predictive means of the three coefficients a_0, a_1, b_1, along with their +1/-1 standard deviations.

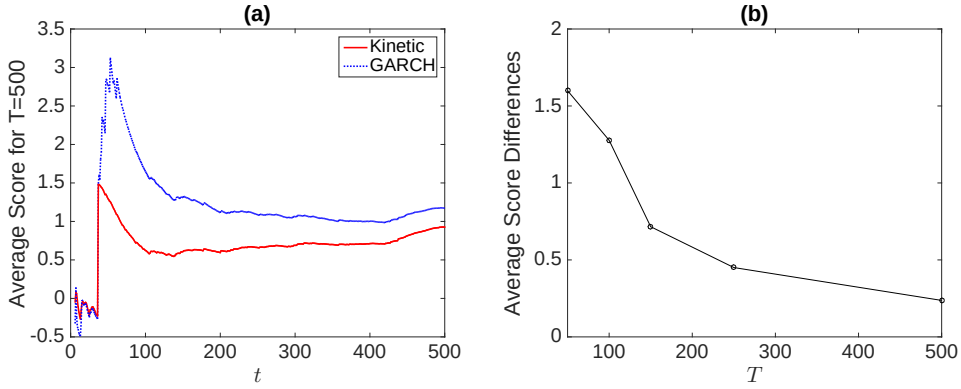

Fig. 22. Synthetic data experiment in Subsection VI-F: (a) the average scores of two methods at each time step t for fixed $T = 500$, (b) the average score of the recursive GARCH method minus that of the kinetic method for different T's.

To prove (A.40), we first show that the summands in (A.40) form a martingale difference sequence. In fact,

$$\mathbb{E}_*\Big\{ s(g_t, Y_t) - \mathbb{E}_*\{s(g_t, Y_t) \mid Y_{1:t-1}\} \;\Big|\; Y_{1:t-1} \Big\} = 0$$

for each $t = 1, 2, \ldots$. For a martingale difference sequence $X_1, X_2, \ldots$, Kolmogorov's strong law of large numbers for martingales states that if $\sum_{t=1}^{\infty} t^{-2} E X_t^2 < \infty$, then $T^{-1}\sum_{t=1}^{T} X_t \to_{a.s.} 0$ [62], [63]. Thus, to obtain (A.40) it is sufficient to prove that $\mathbb{E}_*\{\max_{g\in\mathcal{G}} s(g, Y_t)^2\}$ is upper bounded by a fixed constant.

It follows from Assumption (2) that for some constant c_Z, $EZ_t^2 \le c_Z$ and $EZ_t \le \sqrt{c_Z}$ for all t, where $Z_t = Z(Y_t)$ as defined in Assumption (2). Thus,

$$\mathbb{E}_*\Big\{ \max_{g\in\mathcal{G}} |s(g, Y_t) - s(g_t^*, Y_t)|^2 \Big\} \le E\Big\{ Z_t^2 \max_{g\in\mathcal{G}} d_{\mathscr{G}}(g, g_t^*)^2 \Big\} \le c_Z c_{\mathcal{G}}^2. \tag{A.41}$$

where $c_{\mathcal{G}}$ is defined in Assumption (1). By triangle

inequality, for any $g \in \mathcal{G}$ we have

$$|s(g,Y_t)^2 - s(g_t^*,Y_t)^2| \le |s(g,Y_t) - s(g_t^*,Y_t)|^2 + 2|s(g_t^*,Y_t)| \cdot |s(g,Y_t) - s(g_t^*,Y_t)|,$$

It then follows from Cauchy inequality that

$$\begin{aligned}&\mathbb{E}_*\left\{\max_{g\in\mathcal{G}} |s(g,Y_t)^2 - s(g_t^*,Y_t)^2|\right\}\\ &\le 2\sqrt{\mathbb{E}_*|s(g_t^*,Y_t)|^2}\times\\ &\quad\sqrt{\mathbb{E}_*\left\{\max_{g\in\mathcal{G}} |s(g,Y_t) - s(g_t^*,Y_t)|^2\right\}}\\ &\le c' \end{aligned} \tag{A.42}$$

where $c' \stackrel{\Delta}{=} c_Z c_{\mathcal{G}}^2 + 2\sqrt{c_Y \cdot c_Z c_{\mathcal{G}}^2}$ is a fixed constant, and c_Y is given by Assumption (3).

Combining (A.41) and (A.42), we obtain

$$\begin{aligned}&\mathbb{E}_*\left\{\max_{g\in\mathcal{G}} s(g,Y_t)^2\right\} \le \mathbb{E}_*\left\{s(g_t^*,Y_t)^2\right\} +\\ &\mathbb{E}_*\left\{\max_{g\in\mathcal{G}} |s(g,Y_t)^2 - s(g_t^*,Y_t)^2|\right\} \le c_Y + c'. \end{aligned} \tag{A.43}$$

which concludes the proof for (A.40). The proof of (10) follows directly from (A.40), $\mathbb{E}_*\{s(g_t,Y_t)\} \ge \mathbb{E}_*\{s(g_t^*,Y_t)\}$ (by the definition of proper scoring rule), and the identity

$$\begin{aligned}&\sum_{t=1}^T \left\{s(g_t,Y_t) - s(g_t^*,Y_t)\right\}\\ =&\sum_{t=1}^T \left\{s(g_t,Y_t) - \mathbb{E}_*\{s(g_t,Y_t)\}\right\} +\\ &\sum_{t=1}^T \left\{\mathbb{E}_*\{s(g_t,Y_t)\} - \mathbb{E}_*\{s(g_t^*,Y_t)\}\right\} -\\ &\sum_{t=1}^T \left\{s(g_t^*,Y_t) - \mathbb{E}_*\{s(g_t^*,Y_t)\}\right\}.\end{aligned}$$

■

Proofs for Section III

First, we introduce some technical lemmas.

Lemma 1. *If X_1, X_2 are nonnegative random variables that satisfy $X_1 \sim$ TE$(\lambda; a, b)$ and $X_2 \le cX_1$ almost surely for some constant $c > 0$. Then*

$$X_2 \sim \text{TE}(c\lambda; ca, b). \tag{A.44}$$

Proof: For any constant $\delta > 0$, we have

$$\begin{aligned}\text{pr}\{X_2 > ca + \delta\} &\le \text{pr}\{X_1 > a + \delta c^{-1}\}\\ &\le b\exp\{-\delta(c\lambda)^{-1}\}.\end{aligned}$$

This concludes the proof. ■

In the sequel, we define $g_i^{(\varepsilon)}$, $p_{t,i}$, and $w_{t,i}$ as they were in Algorithm 1. We let $W_t \stackrel{\Delta}{=} \sum_{i=1}^N w_{t,i}$.

Lemma 2. *Suppose that Assumptions (1), (2), (3) hold. Suppose that Algorithm 1 is run with $\alpha = 0$ and any initial weights such that $W_0 \le 1$. Then for each $T \ge 1$ and any $\beta > 0$ such that $T^\beta > c_{\mathcal{G}} a$,*

$$\sum_{t=1}^T \sum_{i=1}^N p_{t,i}\ s(g_i^{(\varepsilon)}, Y_t) \le -\frac{1}{\eta}\log(W_T) + \frac{\eta}{8}T^{1+2\beta} \tag{A.45}$$

holds with probability at least $1 - c_1 T\exp(-c_2 T^\beta)$ for some fixed constants $c_1, c_2 > 0$ (the randomness comes from $\{Y_t\}_{t=1}^T$).

Proof: This is a variant of Lemma 5.1 of [43]. From Assumptions (1) (2) (3), and Lemma 1, it is easy to see that for each t,

$$\sup_{g,\tilde{g}\in\mathcal{G}} |s(g,Y_t) - s(\tilde{g},Y_t)| \sim \text{TE}(c_{\mathcal{G}}\lambda; c_{\mathcal{G}}a, b).$$

Define the event $E_t = \{\omega \in \Omega : \sup_{g,\tilde{g}\in\mathcal{G}} |s(g,Y_t(\omega)) - s(\tilde{g},Y_t(\omega))| \le T^\beta\}$. By Definition (2), if $T^\beta > c_{\mathcal{G}}a$,

$$\text{pr}(\overline{E_t}) \le b\exp\left[-\{T^\beta - c_{\mathcal{G}}a\}(c_{\mathcal{G}}\lambda)^{-1}\right]$$

where $\overline{E_t}$ denotes the complement of event E_t. This implies that the event $\cap_{t=1}^T E_t$ holds with probability at least $1 - c_1 T\exp(-c_2 T^\beta)$ for some fixed constants $c_1, c_2 > 0$.

Now we place ourselves on $\cap_{t=1}^T E_t$. It remains to prove (A.45). For this we could directly apply the proving techniques as in [43, Lemma 5.1] by treating each $g_i^{(\varepsilon)}$ as an expert and $s(g_i^{(\varepsilon)}, Y_t)$ as the loss function for the expert, which gives

$$\begin{aligned}&\sum_{t=1}^T \sum_{i=1}^N p_{t,i}\ s(g_i^{(\varepsilon)}, Y_t) \le -\frac{1}{\eta}\log(W_T) +\\ &\frac{\eta T}{8}\left[\sup_{t=1,\ldots,T}\left\{\sup_{g\in\mathcal{G}} s(g,Y_t) - \inf_{g\in\mathcal{G}} s(g,Y_t)\right\}\right]^2.\end{aligned}$$

On the events $\cap_{t=1}^T E_t$, we have that for all $t = 1,\ldots,T$,

$$\sup_{g\in\mathcal{G}} s(g,Y_t) - \inf_{g\in\mathcal{G}} s(g,Y_t) = \sup_{g,\tilde{g}\in\mathcal{G}} |s(g,Y_t) - s(\tilde{g},Y_t)| \le T^\beta,$$

which concludes the proof.

■

Lemma 3. *Suppose that Assumptions (1), (2), (3) hold. Suppose the true data generating density sequence $\{g_t^*\}_{t=1}^T$ has at most $M_T - 1$ abrupt changes, and let $m = 1,\ldots,M_T$ index the constant stages $\{g_t^*\}_{t=1}^T$. Denote the time index set of the m-th stage as $D_m =$*

$\{t_{m-1}, t_{m-1}+1, \cdots, t_m - 1\}$, *where* $1 < t_1 < t_2 < \cdots < t_{M_T - 1}$ *are the changes time points, and we define* $t_0 = 1$ *and* $t_{M_T} = T+1$. *For each constant stage* m, *let the true data generating density be* $g^*_{(m)}$, *and the closest function base be* $g^{(\varepsilon)}_{(m)}$. *In other words,* $g^*_t = g^*_{(m)}$ *whenever* $t \in D_m$, *and that* $d_{\mathscr{G}}(g^*_{(m)}, g^{(\varepsilon)}_{(m)}) \leq \varepsilon$. *Then if we run Algorithm 1 with* $\alpha \in (0,1)$ *and any initial weights such that* $W_0 \leq 1$, *for each* $T \geq 1$ *and any* $\beta > 0$ *such that* $T^\beta > c_{\mathcal{G}} a$, *we will have*

$$\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\ s(g_i^{(\varepsilon)}, y_t) - \sum_{m=1}^{M_T}\sum_{t\in D_m} s(g^{(\varepsilon)}_{(m)}, Y_t) \leq \frac{M_T}{\eta}\log N - \frac{1}{\eta}\log\{\alpha^{M_T-1}(1-\alpha)^{T-M_T}\} + \frac{\eta}{8}T^{1+2\beta}$$

holds with probability at least $1 - c_1 T\exp(-c_2 T^\beta)$, *where* c_1, c_2 *are the same as those defined in Lemma 2.*

Proof: Lemma 3 can be proved using Lemma 2 and a direct adaptation of [43, Theorem 5.2]. ∎

Proof of Theorem 2

Proof:

(i) With the same assumptions and notations as in Lemma 3, we obtain

$$\begin{aligned}&\sum_{m=1}^{M_T}\sum_{t\in D_m} s(g^{(\varepsilon)}_{(m)}, Y_t) - \sum_{t=1}^{T} s(g^*_t, Y_t)\\ &= \sum_{m=1}^{M_T}\sum_{t\in D_m}\{s(g^{(\varepsilon)}_{(m)}, Y_t) - s(g^*_{(m)}, Y_t)\}\\ &\leq \sum_{m=1}^{M_T}\sum_{t\in D_m} Z_t\varepsilon \leq\ Z^{(T)}T\varepsilon \qquad \text{(A.46)}\end{aligned}$$

where $Z_t = Z(Y_t) \sim \text{TE}(\lambda; a, b)$ are nonnegative random variables defined in Assumption (2), and $Z^{(T)} \triangleq \max\{Z_t : t = 1, \ldots, T\}$. Simple calculation gives that

$$\begin{aligned}&\text{pr}(Z^{(T)} > T^\beta) = \text{pr}\Big(\cup_{t=1}^{T}\{Z_t > T^\beta\}\Big)\\ &\leq \sum_{t=1}^{T}\text{pr}(Z_t > T^\beta)\\ &\leq Tb\exp\{-(T^\beta - a)\lambda^{-1}\} \qquad \text{(A.47)}\end{aligned}$$

for all $T > a^{-\beta}$. From (A.47), the right-hand side of (A.46) is less than $T^{1+\beta}\varepsilon$ with probability at least $1 - c_3 T\exp(-c_4 T^\beta)$ for some fixed constants $c_3, c_4 > 0$. Combining this with Lemma 3, we obtain that

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\ s(g_i^{(\varepsilon)}, y_t) - \sum_{t=1}^{T} s(g^*_t, y_t)\\ &= \sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\ s(g_i^{(\varepsilon)}, y_t) - \sum_{m=1}^{M_T}\sum_{t\in D_m} s(g^{(\varepsilon)}_{(m)}, Y_t)\\ &+ \sum_{m=1}^{M_T}\sum_{t\in D_m} s(g^{(\varepsilon)}_{(m)}, Y_t) - \sum_{t=1}^{T} s(g^*_t, Y_t)\\ &\leq \frac{M_T}{\eta}\log N - \frac{1}{\eta}\log\{\alpha^{M_T-1}(1-\alpha)^{T-M_T}\} + \\ &\quad \frac{\eta}{8}T^{1+2\beta} + T^{1+\beta}\varepsilon \qquad \text{(A.48)}\end{aligned}$$

holds with probability at least $1 - c_1 T\exp(-c_2 T^\beta) - c_3 T\exp(-c_4 T^\beta) \geq 1 - C_1 T\exp(-C_2 T^\beta)$, where c_1, c_2 were defined in Lemma 3, and $C_1 \triangleq c_1 + c_3$, $C_2 \triangleq \min\{c_2, c_4\}$. Choose α and η as in (14), then the inequality (A.48) becomes (15).

(ii) We arbitrarily pick up ε_T that satisfies (13), and apply an ε_T-net on $\mathcal{G}$, denoted by $\{g_i^{(\varepsilon)} : i = 1, \ldots, N\}$ with $\log N = H_{\mathcal{G}}(\varepsilon_T)$. Recall the binary entropy inequality

$$H(q) \leq 2\log(2)\sqrt{q(1-q)} \leq 2\log(2)\sqrt{q}. \qquad \text{(A.49)}$$

Thus, the right hand side of inequality (15) divided by T equals

$$\begin{aligned}&\sqrt{2^{-1}T^{-1+2\beta}Q_{T,N}} + T^\beta\varepsilon_T \leq \sqrt{2^{-1}}\times\\ &\sqrt{T^{-1+2\beta}M_T H_{\mathcal{G}}(\varepsilon_T) + 2\log(2)T^{-1/2+2\beta}M_T^{1/2}} + \\ &T^\beta\varepsilon_T \qquad \text{(A.50)}\end{aligned}$$

which converges to zero given (13). Therefore, (i) implies that

$$\frac{1}{T}\sum_{t=1}^{T}\Big\{\sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_{i,T}, Y_t) - s(g^*_t, Y_t)\Big\} \leq o(1)$$

holds with probability at least $1 - C_1 T\exp(-C_2 T^\beta)$.
From $\sum_{T=1}^{\infty} C_1 T\exp(-C_2 T^\beta) < \infty$ and using Borel-Cantelli lemma, we can prove (16).

(iii) The result follows directly from from (16) and the convexity

$$s(\hat{g}_t, Y_t) \leq \sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_{i,T}, Y_t),$$

combined with Theorem 1.

(iv) First, we note that

$$B_t \triangleq s(g^{(\varepsilon)}_{J_{t,T}}, Y_t) - \sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_{i,T}, Y_t)$$

is a martingale difference sequence with respect to (J_t, Y_t). We already proved (A.43) which implies that $E|B_t| < \infty$. Because J_t is independently generated from $\{1, \ldots, N\}$ with probability $p_{t,i}$ which is determined before time t, we have

$$\mathbb{E}_*(B_t \mid J_{1:t-1}, Y_{1:t-1}) = \mathbb{E}_*(B_t \mid Y_{1:t-1}) = \mathbb{E}_*\Big\{ \big[s(g^{(\varepsilon)}_{J_t,T}, Y_t) - \sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_{i,T}, Y_t) \big] \mid Y_{1:t-1} \Big\} = 0.$$

Next, we prove that $E(B_t \mid J_{1:t-1}, Y_{1:t-1})^2$ is upper bounded by a constant. Recall that $p_{t,i}, i = 1, \ldots, N$ are in the σ-field generated by $(J_{1:t-1}, Y_{1:t-1})$. We have

$$\begin{aligned}
&\mathbb{E}_*(B_t \mid J_{1:t-1}, Y_{1:t-1})^2 \\
&= var(B_t \mid J_{1:t-1}, Y_{1:t-1}) = var(B_t \mid Y_{1:t-1}) \\
&= \mathbb{E}_*\Big\{ var(B_t \mid Y_{1:t}) \mid Y_{1:t-1} \Big\} + \\
&\qquad var\Big\{ \mathbb{E}_*(B_t \mid Y_{1:t}) \mid Y_{1:t-1} \Big\} \qquad \text{(A.51)} \\
&= \mathbb{E}_*\Big\{ var(B_t \mid Y_{1:t}) \mid Y_{1:t-1} \Big\} \\
&= \mathbb{E}_*\Big\{ \sum_{j=1}^{N} p_{t,j} \Big(s(g^{(\varepsilon)}_j, Y_t) - \\
&\qquad \sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_i, Y_t) \Big)^2 \mid Y_{1:t-1} \Big\} \\
&\le \sum_{j=1}^{N} p_{t,j} \times 2\mathbb{E}_*\Big\{ s(g^{(\varepsilon)}_i, Y_t)^2 + \\
&\qquad \Big(\sum_{i=1}^{N} p_{t,i}\ s(g^{(\varepsilon)}_i, Y_t) \Big)^2 \mid Y_{1:t-1} \Big\} \\
&\le 4\mathbb{E}_*\Big\{ \max_{g \in \mathcal{G}} s(g, Y_t)^2 \Big\} \le C \qquad \text{(A.52)}
\end{aligned}$$

for some constant C, where (A.51) is from the Eve's law, and (A.52) is due to the bound in (A.43). Finally, by a martingale convergence theorem [62, Theorem 1], we have $T^{-1} \sum_{t=1}^{T} B_t \to_{a.s.} 0$. Combining this with (16) and Theorem 1, we obtain (18).

■

Proof of Proposition 2

Proof:

We prove by construction. Suppose that the distributions of $Y_{1:T}$ (denoted by P_1) contains two abrupt change points $1 < t_1 < t_2 < T$ such that $Y_{t_1+1}, \ldots, Y_{t_2}$ are i.i.d. distributed, with distributions different from that of Y_{t_1} and Y_{t_2+1}. Now consider the distribution of $Y_{1:T}$ (denoted by P_2) that replace the distribution of each $Y_{t_1+1}, \ldots, Y_{t_2}$ with that of Y_{t_1}. We let $\#(P_k)$ denote the number of change points of the distribution for $k = 1, 2$. Let $\psi(Y_{1:T})$ denote the selected number of change points by any selection procedure (denoted by ψ). Using Le Cam's method, we obtain

$$\max_{k=1,2} P_k\big\{ \#(P_k) \neq \psi(Y_{1:T}) \big\} \ge \frac{1}{2} - \frac{1}{2} d_{\text{TV}}(P_1, P_2) \qquad \text{(A.53)}$$

where d_{TV} denotes the total variation distance. Pinsker's inequality gives

$$\begin{aligned}
d_{\text{TV}}(P_1, P_2) &\le \sqrt{\frac{1}{2} d_{\text{KL}}(P_1, P_2)} \\
&= \sqrt{\frac{1}{2}(t_2 - t_1) d_{\text{KL}}(g^*_{t_2}, g^*_{t_1})}. \qquad \text{(A.54)}
\end{aligned}$$

where d_{KL} denotes the Kullback-Leibler divergence. If procedure ψ is consistent in selecting the number of abrupt change points, then $\max_{k=1,2} P_k\big\{ \#(P_k) \neq \psi(Y_{1:T}) \big\} \to 0$. Inequalities (A.53) and (A.54) imply that consistent selection is not possible as long as

$$(t_2 - t_1)\ d_{\text{KL}}(g^*_{t_2}, g^*_{t_1}) \le q < 2 \qquad \text{(A.55)}$$

for some positive constant q. Since condition (A.55) does not violate the assumptions of Theorem 2, the proof is complete. ■

Proofs for Section IV

We first introduce some helpful technical lemmas.

Lemma 4. *([43, Lemma A.1]) Let X be a random variable with $a \le X \le b$. Then for any $s \in \mathbb{R}$,*

$$\ln \mathbb{E}[e^{sX}] \le s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}.$$

Lemma 5. *Suppose that $\Theta \triangleq [a, b] \subseteq \mathbb{R}$ is a compact interval. For any function $f(x) : \Theta \to \mathbb{R}$ that is Lipschitz continuous, i.e.*

$$\forall x_1, x_2 \in \Theta, \quad |f(x_1) - f(x_2)| \le D \cdot |x_1 - x_2|$$

where $D > 0$ is the Lipschitz constant, we have

$$\int_{\Theta} e^{-f(x)} dx \ge e^{-\min_{x \in \Theta} f(x)} \cdot |\Theta| \cdot r$$

where $|\Theta| \triangleq \int_{\Theta} dx = (b - a)$, and

$$r \triangleq \frac{1 - \exp(-D|\Theta|)}{D|\Theta|} \in (0, 1).$$

Proof: Since Lipschitz continuity implies continuity and the function $f(x)$ is defined on a compact space, the minimum value is always attained (by the Extreme Value Theorem).

First consider the case when $\min_{x\in\Theta} f(x) = f(a)$. Partition the interval $[a,b]$ evenly into N pieces, with each piece of length $\epsilon = |\Theta|/N$. Denote the end points of the small intervals as $a = x_0 < x_1 < \cdots < x_N = b$. By Riemann integration theory, we have

$$\int_\Theta e^{-f(x)}dx = \lim_{\epsilon\to 0}\sum_{i=0}^{N-1}\epsilon\cdot e^{-f(x_i)}$$

Using the Lipschitz condition that is satisfied by f, we have that

$$e^{-f(x_i)} = e^{-f(x_0)}e^{f(x_0)-f(x_i)} \geq e^{-f(x_0)}\cdot e^{-D|x_i-x_0|}$$
$$= e^{-f(x_0)}\cdot e^{-D\cdot i\epsilon}.$$

Therefore, we have

$$\begin{aligned}\sum_{i=0}^{N-1}\epsilon\cdot e^{-f(x_i)} &\geq \epsilon\cdot e^{-f(x_0)}\sum_{i=0}^{N-1}e^{-D\cdot i\epsilon}\\ &= \epsilon\cdot e^{-f(a)}\frac{1-e^{-D\cdot N\epsilon}}{1-e^{-D\cdot\epsilon}}\\ &= \epsilon\cdot e^{-f(a)}\frac{1-e^{-D\cdot|\Theta|}}{1-e^{-D\cdot\epsilon}}.\end{aligned}$$

Taking the limit as $\epsilon\to 0$ on both sides, we obtain

$$\begin{aligned}\int_\Theta e^{-f(x)}dx \geq& e^{-f(a)}(1-e^{-D\cdot|\Theta|})\lim_{\epsilon\to 0}\frac{\epsilon}{1-e^{-D\cdot\epsilon}}\\ =& e^{-f(a)}\frac{1-e^{-D\cdot|\Theta|}}{D}\end{aligned}$$

by L'Hopital's rule. By assumption $\min_{x\in\Theta} f(x) = f(a)$, hence the lemma is proved for this case.

For the case when $\min_{x\in\Theta} f(x) = f(b)$, the proof is similar.

Finally consider the situation where we have $\min_{x\in\Theta} f(x) = f(c)$ with some $c\in(a,b)$. Let $\Theta_1 \triangleq [a,c]$ and $\Theta_2 \triangleq [c,b]$. We already proved that

$$\int_{\Theta_1} e^{-f(x)}dx \geq e^{-f(c)}\frac{1-e^{-D\cdot|\Theta_1|}}{D},$$
$$\int_{\Theta_2} e^{-f(x)}dx \geq e^{-f(c)}\frac{1-e^{-D\cdot|\Theta_2|}}{D}.$$

By adding them, we have

$$\int_\Theta e^{-f(x)}dx \geq e^{-f(c)}\frac{2-e^{-D\cdot|\Theta_1|}-e^{-D\cdot|\Theta_2|}}{D}$$

Let $g(c) = e^{-D\cdot|\Theta_1|}+e^{-D\cdot|\Theta_2|}$, where $|\Theta_1| = (c-a)$ and $|\Theta_2| = (b-c)$. Direct calculations show that $g(c)$ is convex in $c\in[a,b]$. So we have

$$g(c)\leq g(a) = g(b) = 1+e^{-D\cdot|\Theta|}$$

Therefore, we still obtain

$$\int_\Theta e^{-f(x)}dx \geq e^{-f(c)}\frac{1-e^{-D\cdot|\Theta|}}{D},$$

where $f(c) = \min_{x\in\Theta} f(x)$, and this is the same as the inequality stated in the lemma.

Furthermore, we can easily show that $r\to 1$ whenever $D|\Theta|\to 0$, and $r\to 0$ whenever $D|\Theta|\to\infty$, and that r is a decreasing function in $D|\Theta|$. ∎

Lemma 6. *Suppose that $\Theta = \Theta_1\times\Theta_2\times\cdots\times\Theta_n\subseteq\mathbb{R}^n$, where each Θ_j is a one-dimensional closed interval: $\Theta_j = [\underline{\theta}_j, \overline{\theta}_j]\in\mathbb{R}$, $j = 1,\ldots,n$. For any function $f(\mathbf{x}):\Theta\to\mathbb{R}$ that is Lipschitz continuous with $D>0$ is the Lipschitz constant, i.e.*

$$\forall\mathbf{x}_1,\mathbf{x}_2\in\Theta,\quad |f(\mathbf{x}_1)-f(\mathbf{x}_2)|\leq D\cdot\|\mathbf{x}_1-\mathbf{x}_2\|_2$$

we have

$$\int_\Theta e^{-f(\mathbf{x})}d\mathbf{x}\geq e^{-\min_{\mathbf{x}\in\Theta} f(\mathbf{x})}\cdot|\Theta|\cdot r^{(n)}$$

where $|\Theta|\triangleq\int_\Theta d\mathbf{x} = \prod_{j=1}^n|\Theta_j| = \prod_{j=1}^n(\overline{\theta}_j-\underline{\theta}_j)$ is the Lebesgue measure of the hypercube Θ, and

$$r^{(n)}\triangleq\prod_{j=1}^n\frac{1-\exp(-D|\Theta_j|)}{D|\Theta_j|}\in(0,1).$$

Proof: Since Lipschitz continuity implies continuity and the function $f(\mathbf{x})$ is defined on a compact space, the minimum value can always be attained from the Extreme Value Theorem. Let $\mathbf{x}^* = [x_1^*, x_2^*,\cdots,x_n^*]^{\mathrm{T}} = \arg\min_{\mathbf{x}\in\Theta} f(\mathbf{x})$.

First, for dimension $j = 1,\ldots,n$, we divide Θ_j into N_j pieces evenly with each piece having length $\epsilon_j = |\Theta_j|/N_j$. Denote the end points of the small intervals in dimension j as $\underline{\theta}_j = x_j^{(0)} < x_j^{(1)} < \cdots < x_j^{(N_j)} = \overline{\theta}_j$. Then by Riemann integration formula, we have

$$\int_\Theta e^{-f(\mathbf{x})}d\mathbf{x} = \lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0}\cdots\lim_{\epsilon_n\to 0}(\epsilon_1\epsilon_2\cdots\epsilon_n)$$
$$\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2}\cdots\sum_{i_n=1}^{N_n}e^{-f\left([x_1^{(i_1)},x_2^{(i_2)},\cdots,x_n^{(i_n)}]^{\mathrm{T}}\right)}.$$

Now fix i_2 to i_n, only let i_1 vary, from lemma 5 we have

$$\lim_{\epsilon_1\to 0}\epsilon_1\sum_{i_1=1}^{N_1}e^{-f\left([x_1^{(i_1)},x_2^{(i_2)},\cdots,x_n^{(i_n)}]^{\mathrm{T}}\right)}$$
$$\geq e^{-f\left([x_1^*,x_2^{(i_2)},\cdots,x_n^{(i_n)}]^{\mathrm{T}}\right)}\cdot\frac{1-e^{-D|\Theta_1|}}{D}.$$

This is because the function $f(\cdot)$ above can be seen as a univariate function $f_1(\cdot):\Theta_1\to\mathbb{R}$, and $e^{-\min_{x_1\in\Theta_1} f_1(x_1)}\geq e^{-f_1(x_1^*)}$. Then we can fix i_3 to i_n, and only let i_2 vary and apply the same reasoning. By repeated application of this procedure, we conclude that

$$\int_\Theta e^{-f(\mathbf{x})}d\mathbf{x}\geq e^{-f(\mathbf{x}^*)}\left[\prod_{j=1}^n\frac{1-e^{-D|\Theta_j|}}{D|\Theta_j|}\right]|\Theta|$$

which is exactly the result in the lemma. ∎

Next we start the analysis of Algorithm 2. We first introduce some additional notations. Let $l_t(\cdot)$ be a function of θ after y_t is revealed, given by $l_t(\theta) = s(\theta, y_t)$. We define $L_{(s,t]}(\cdot) = \sum_{\tau=s+1}^{t} l_\tau(\cdot)$ as the partial cumulative score function, and define $L_{[s,t]}(\cdot)$, $L_{[s,t)}(\cdot)$ similarly.

Lemma 7. *Assume that the parameter space $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_n \subseteq \mathbb{R}^n$ is an $n-$dimensional cube, and the score functions $l_t(\theta) : \Theta \to \mathbb{R}$ are Lipschitz continuous uniformly over $t = 1, \ldots, T$, with the common Lipschitz constant $d > 0$. Then for Algorithm 2, we have*

$$\int_\Theta \tilde{f}_t(\theta)d\theta \geq \max_{\theta\in\Theta} \tilde{f}_t(\theta) \cdot r_t^{(n)} \cdot |\Theta| \tag{A.56}$$

for all $t = 1, 2, \ldots, T$, where

$$r_t^{(n)} = \prod_{j=1}^{n} \frac{1 - \exp(-\eta t d|\Theta_j|)}{\eta t d|\Theta_j|} \in (0, 1) \tag{A.57}$$

for all $t = 1, \ldots, T$. Moreover, $\{r_t^{(n)}\}_{t=1}^T$ is a decreasing sequence in t.

Proof: Let $g_{s,t}(\theta) = e^{-\eta \sum_{i=s}^{t} l_i(\theta)}$, $\theta \in \Theta$, $s, t \in \{1, \ldots, T\}$ and $s \leq t$. We first show that for any $t \in \{1, \ldots, T\}$, $\tilde{f}_t(\theta)$ may be written as $\sum_{s=1}^{t} C_{s,t} g_{s,t}(\theta)$ where $C_{s,t} \geq 0$ are constants. To prove this, we first start with

$$\tilde{f}_1(\theta) = f_0(\theta)e^{-\eta l_1(\theta)} = e^{-\eta l_1(\theta)} = g_{1,1}(\theta)$$

with $C_{1,1} = 1$. Then suppose $\tilde{f}_t(\theta) = \sum_{s=1}^{t} C_{s,t} g_{s,t}(\theta)$ (where $C_{s,t} > 0$) is true for some $t = 1, \ldots, T$. From the procedure (20) of Algorithm 2, we have

$$f_t(\theta) = (1 - \alpha)\tilde{f}_t(\theta) + A_t$$

where $A_t = \alpha|\Theta|^{-1} \int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}$ is a constant regardless of θ (which only depends on $y_{1:t}$ through the score function). Then from procedure (19), we obtain

$$\begin{aligned}
\tilde{f}_{t+1}(\theta) =& f_t(\theta) \cdot e^{-\eta l_{t+1}(\theta)} \\
=& (1 - \alpha)\tilde{f}_t(\theta)e^{-\eta l_{t+1}(\theta)} + A_t e^{-\eta l_{t+1}(\theta)} \\
=& (1 - \alpha) \sum_{s=1}^{t} C_{s,t} g_{s,t}(\theta) \cdot e^{-\eta l_{t+1}(\theta)} + \\
& A_t g_{t+1,t+1}(\theta) \\
=& (1 - \alpha) \sum_{s=1}^{t} C_{s,t} g_{s,t+1}(\theta) + A_t g_{t+1,t+1}(\theta) \\
=& \sum_{s=1}^{t+1} C_{s,t+1} g_{s,t+1}(\theta)
\end{aligned}$$

where

$$C_{s,t+1} = \begin{cases} (1 - \alpha)C_{s,t} & \text{if } s \leq t \\ A_t & \text{if } s = t + 1 \end{cases}.$$

Therefore, we have proved by mathematical induction that for all $t = 1, \ldots, T$, $\tilde{f}_t(\theta) = \sum_{s=1}^{t} C_{s,t} g_{s,t}(\theta)$ with constants $C_{s,t} > 0$ given by the recursion above.

Now, since $l_i(\theta)$ is Lipschitz continuous with Lipschitz constant being d (by assumption), $\eta \sum_{i=s}^{t} l_i(\theta)$ is also Lipschitz continuous with Lipschitz constant $\eta(t - s + 1)d$. Applying Lemma 6, we obtain

$$\int_\Theta g_{s,t}(\theta)d\theta \geq \max_{\theta\in\Theta} g_{s,t}(\theta) \cdot r_{s,t}^{(n)} \cdot |\Theta|$$

with

$$r_{s,t}^{(n)} \triangleq \prod_{j=1}^{n} \frac{1 - \exp(-\eta(t - s + 1)d|\Theta_j|)}{\eta(t - s + 1)d|\Theta_j|} \in (0, 1).$$

Thus, we obtain

$$\begin{aligned}
\int_\Theta \tilde{f}_t(\theta)d\theta =& \int_\Theta \sum_{s=1}^{t} C_{s,t} g_{s,t}(\theta)d\theta \\
=& \sum_{s=1}^{t} C_{s,t} \int_\Theta g_{s,t}(\theta)d\theta \\
\geq& \sum_{s=1}^{t} C_{s,t} \max_{\theta\in\Theta} g_{s,t}(\theta) r_{s,t}^{(n)} |\Theta| \\
=& \sum_{s=1}^{t} \max_{\theta\in\Theta} \{C_{s,t} g_{s,t}(\theta)\} r_{s,t}^{(n)} |\Theta| \\
\geq& \big(\min_{s=1,\ldots,t} r_{s,t}^{(n)}\big) \cdot \sum_{s=1}^{t} \max_{\theta\in\Theta} \{C_{s,t} g_{s,t}(\theta)\} |\Theta| \\
\geq& \big(\min_{s=1,\ldots,t} r_{s,t}^{(n)}\big) \cdot \max_{\theta\in\Theta} \sum_{s=1}^{t} [C_{s,t} g_{s,t}(\theta)] |\Theta| \\
=& \max_{\theta\in\Theta} \tilde{f}_t(\theta) \cdot r_t^{(n)} \cdot |\Theta|
\end{aligned}$$

where

$$\min_{s=1,\ldots,t} r_{s,t}^{(n)} = r_{1,t}^{(n)} = r_t^{(n)}.$$

This is true for any $t = 1, \ldots, T$, and it is easy to verify that the sequence $\{r_t^{(n)}\}_{t=1}^T$ is a monotonically decreasing sequence with $\lim_{t\to 0} r_t^{(n)} = 1$ and $\lim_{t\to\infty} r_t^{(n)} = 0$. ∎

Lemma 8. *Under the same assumptions as in Lemma 7, for Algorithm 2, for all $s, t \in \{0, 1, \ldots, T\}$ and $t \geq s+1$, we have*

$$\begin{aligned}
\frac{\tilde{f}_t(\theta)}{f_s(\theta)} \geq& \Big[(1 - \alpha) + \alpha r_T^{(n)}\Big]^{t-s-1} e^{-\eta \sum_{i=s+1}^{t} l_i(\theta)} \\
\geq& (1 - \alpha)^{t-s-1} e^{-\eta \sum_{i=s+1}^{t} l_i(\theta)},
\end{aligned}$$

where $r_T^{(n)}$ was defined in (A.57).

Proof: Using the result of Lemma 7, we have

$$\begin{aligned} f_t(\theta) =&(1-\alpha)\tilde{f}_t(\theta) + \frac{\alpha}{|\Theta|}\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta} \\ \geq&(1-\alpha)\tilde{f}_t(\theta) + \frac{\alpha}{|\Theta|}\int_\Theta \max_{\theta\in\Theta} \tilde{f}_t(\theta)\cdot r_t^{(n)}\cdot|\Theta| \\ \geq&\{(1-\alpha)+\alpha r_t^{(n)}\}\tilde{f}_t(\theta) \\ =&\{(1-\alpha)+\alpha r_t^{(n)}\}f_{t-1}(\theta)e^{-\eta l_t(\theta)}. \end{aligned}$$

Since $f_t(\theta) > 0$ almost surely for all $\theta \in \Theta$ and $t = 0, 1, \ldots$ (from the proof of Lemma 7), we have

$$\frac{f_t(\theta)}{f_{t-1}(\theta)} \geq [(1-\alpha)+\alpha r_t^{(n)}]e^{-\eta l_t(\theta)}.$$

When $t \geq s+2$, by applying the inequality recursively, we have

$$\frac{f_{t-1}(\theta)}{f_s(\theta)} \geq \prod_{i=s+1}^{t-1}[(1-\alpha)+\alpha r_i^{(n)}]e^{-\eta L_{(s,t-1]}(\theta)}.$$

It follows that

$$\begin{aligned} \frac{\tilde{f}_t(\theta)}{f_s(\theta)} =& \frac{f_{t-1}(\theta)\cdot e^{-\eta l_t(\theta)}}{f_s(\theta)} \\ \geq& \prod_{i=s+1}^{t}[(1-\alpha)+\alpha r_i^{(n)}]e^{-\eta L_{(s,t]}(\theta)}. \end{aligned}$$

When $t = s+1$, we also have

$$\frac{\tilde{f}_t(\theta)}{f_s(\theta)} = e^{-\eta l_t(\theta)} = e^{-\eta L_{(s,t]}(\theta)}.$$

Finally, the proof is complete by using the fact that $r_i^{(n)} \geq r_T^{(n)} > 0$ for all $i = 1, \ldots, T$. ∎

Lemma 9. *Under the same assumptions as in Lemma 7, for Algorithm 2, for all $\theta, \tilde{\theta} \in \Theta$ and $t = 1, 2, \ldots, T$, we have*

$$\frac{f_t(\theta)}{\tilde{f}_t(\tilde{\theta})} \geq \alpha r_t^{(n)}, \tag{A.58}$$

where $r_t^{(n)}$ was defined in (A.57).

Proof: From the mixing step (20) and Lemma 7, we have

$$\begin{aligned} f_t(\theta) \geq&\alpha\frac{\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}}{|\Theta|} \\ \geq&\alpha\cdot r_t^{(n)}\cdot\max_{\theta\in\Theta}\tilde{f}_t(\theta) \geq \alpha\cdot r_t^{(n)}\cdot\tilde{f}_t(\tilde{\theta}) \end{aligned}$$

for any $\tilde{\theta} \in \Theta$. ∎

Proof of Theorem 3

Proof: Under the assumptions of Theorem 3, the assumptions of Lemma 7 hold with probability at least $1 - bT\exp\left[-(d-a)\lambda^{-1}\right]$; we focus on the events that those assumptions hold.

First, we note that

$$\begin{aligned} \int_\Theta f_t(\theta)d\theta =&(1-\alpha)\int_\Theta \tilde{f}_t(\theta)d\theta + \alpha\frac{\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta}}{|\Theta|}\int_\Theta d\theta \\ =&(1-\alpha)\int_\Theta \tilde{f}_t(\theta)d\theta + \alpha\int_\Theta \tilde{f}_t(\tilde{\theta})d\tilde{\theta} \\ =&\int_\Theta \tilde{f}_t(\theta)d\theta. \end{aligned}$$

Denote $Z_t \triangleq \int_\Theta f_t(\theta)d\theta = \int_\Theta \tilde{f}_t(\theta)d\theta$. Then by Lemma 4,

$$\begin{aligned} \ln\frac{Z_t}{Z_{t-1}} =& \ln\frac{\int_\Theta \tilde{f}_t(\theta)d\theta}{\int_\Theta f_{t-1}(\theta)d\theta} = \ln\frac{\int_\Theta f_{t-1}(\theta)e^{-\eta l_t(\theta)}d\theta}{\int_\Theta f_{t-1}(\theta)d\theta} \\ =& \ln\int_\Theta p_t(\theta)e^{-\eta l_t(\theta)}d\theta = \ln\mathbb{E}_{p_t}[e^{-\eta l_t(\theta)}] \\ \leq& -\eta\mathbb{E}_{p_t}[l_t(\theta)] + \frac{\eta^2\left(\sup_{\theta,\tilde{\theta}\in\Theta}|l_t(\theta)-l_t(\tilde{\theta})|\right)^2}{8} \\ \leq& -\eta\bar{l}_t(\theta) + \frac{\eta^2(d\Delta)^2}{8} \end{aligned}$$

where $\bar{l}_t(\theta) \triangleq \mathbb{E}_{p_t}[l_t(\theta)] = \bar{s}(\theta, y_t)$, and $\Delta \triangleq \sup_{\theta,\tilde{\theta}\in\Theta}\|\theta-\tilde{\theta}\|_2$ in the theorem. By telescoping, we obtain

$$\ln\frac{Z_T}{Z_0} \leq -\eta\sum_{t=1}^{T}\bar{l}_t(\theta) + \frac{\eta^2 T(d\Delta)^2}{8}, \tag{A.59}$$

and $\ln Z_0 = \ln\int_\Theta 1d\theta = \ln|\Theta|$.

Next, suppose that the sequence $\{\theta_t^*\}_{t=1}^T$ contains $k \triangleq M_T - 1$ abrupt switches. We denote the change locations and two endpoints by $0 = t_0 < t_1 < t_2 < \cdots < t_{k+1} = T$. Let $\{\theta_{t_1}, \theta_{t_2}, \cdots, \theta_{t_{k+1}}\}$ be the constant parameters in the $k+1$ time segments, namely $\theta_t^* = \theta_{t_j}$ for $t_{j-1} < t \leq t_j$, $j = 1, 2, \ldots, k+1$. Note that we allow that $\theta_{t_{i+1}} = \theta_{t_i}$ for some i, whenever the sequence $\{\theta_t^*\}_{t=1}^T$ has number of switches less than k. Then we have

$$\begin{aligned} \tilde{f}_{t_{k+1}}(\theta_{t_{k+1}}) =&\tilde{f}_{t_1}(\theta_{t_1})\prod_{i=1}^{k}\frac{\tilde{f}_{t_{i+1}}(\theta_{t_{i+1}})}{\tilde{f}_{t_i}(\theta_{t_i})} \\ =&f_0(\theta_{t_1})\frac{\tilde{f}_{t_1}(\theta_{t_1})}{f_0(\theta_{t_1})}\prod_{i=1}^{k}\frac{f_{t_i}(\theta_{t_{i+1}})}{\tilde{f}_{t_i}(\theta_{t_i})}\frac{\tilde{f}_{t_{i+1}}(\theta_{t_{i+1}})}{f_{t_i}(\theta_{t_{i+1}})}. \end{aligned}$$

Applying the results of Lemma 8 and Lemma 9, we

obtain

$$\begin{aligned}
&\tilde{f}_T(\theta_T^*) = \tilde{f}_{t_{k+1}}(\theta_{t_{k+1}}) \\
&\geq 1 \cdot (1-\alpha+\alpha r_T^{(n)})^{t_1-1} \cdot e^{-\eta \sum_{j=1}^{t_1} l_j(\theta_{t_1})} \cdot \\
&\prod_{i=1}^{k} \left[\alpha r_{t_i}^{(n)} \cdot (1-\alpha+\alpha r_T^{(n)})^{t_{i+1}-t_i-1} e^{-\eta \sum_{j=t_i+1}^{t_{i+1}} l_j(\theta_{t_{i+1}})} \right] \\
&= (1-\alpha+\alpha r_T^{(n)})^{t_{k+1}-(k+1)} \cdot \alpha^k \cdot \left(\prod_{i=1}^{k} r_{t_i}^{(n)} \right) \cdot e^{-\eta \sum_{t=1}^{T} l_t(\theta_t^*)} \\
&= (1-\alpha)^{T-k-1} \alpha^k \cdot \left(1 + \frac{\alpha}{1-\alpha} r_T^{(n)} \right)^{T-k-1} \cdot \left(\prod_{i=1}^{k} r_{t_i}^{(n)} \right) \cdot \\
&\quad e^{-\eta \sum_{t=1}^{T} l_t(\theta_t^*)}.
\end{aligned}$$

Applying Lemma 7, we obtain

$$\begin{aligned}
Z_T = \int_\Theta \tilde{f}_T(\theta) d\theta \geq & r_T^{(n)} |\Theta| \max_{\theta \in \Theta} \tilde{f}_T(\theta) \\
\geq & r_T^{(n)} |\Theta| \tilde{f}_T(\theta_T^*),
\end{aligned}$$

which implies

$$\begin{aligned}
\ln \frac{Z_T}{Z_0} &= \ln Z_T - \ln |\Theta| \geq \ln r_T^{(n)} + \ln \tilde{f}_T(\theta_T^*) \\
&\geq (T-k-1) \ln(1-\alpha) + k \ln \alpha + \\
&\quad (T-k-1) \ln \left(1 + \frac{\alpha}{1-\alpha} r_T^{(n)} \right) + \\
&\quad \sum_{i=1}^{k+1} \ln r_{t_i}^{(n)} - \eta \sum_{t=1}^{T} l_t(\theta_t^*). \qquad \text{(A.60)}
\end{aligned}$$

By combining (A.59) and (A.60), and using the fact that $\{r_t^{(n)}\}_{t=1}^T$ is a decreasing sequence, we obtain the bound in (21).

■

Proof of Corollary 1.

Proof:

(i) This is a restatement of Theorem 1 in the parametric case. Proof is similar to that of Theorem 1.

(ii) From the results of Theorem 3 and Proposition 3, we know that under the assumptions here, we have

$$\begin{aligned}
&\frac{1}{T} \left\{ \sum_{t=1}^{T} \bar{s}(\theta, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \right\} \\
&\leq \frac{1}{T \eta_T} \left[(T-1) H\left(\frac{M_T - 1}{T-1} \right) - M_T \ln(r_T^{(n)}) \right] + \\
&\frac{\eta_T (d\Delta)^2}{8}
\end{aligned}$$

holds with probability at least $1 - bT \exp\left[-(d-a)\lambda^{-1}\right]$. Denote right-hand-side bound above as B_T, and the event $\left\{ \omega \in \Omega : \frac{1}{T}\left\{ \sum_{t=1}^{T} \bar{s}(\theta, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \right\} \leq B_T \right\}$ as A_T. Let $d = d_T = O(T^\delta)$ with $\delta \in (0, \beta/2)$. First we have

$$\begin{aligned}
\frac{\eta_T (d_T \sup_{\theta, \tilde{\theta} \in \Theta} \|\theta - \tilde{\theta}\|_2)^2}{8} &= O(T^{-\beta} T^{2\delta}) \\
&= O(T^{-2(\beta/2-\delta)}) \\
&= o(1),
\end{aligned}$$

$$\begin{aligned}
H(\frac{k_T}{T-1}) &= O\Big(H(T^{\gamma-1}) \Big) \\
&= O\Big(T^{\gamma-1} \ln(T^{1-\gamma}) + o(1) \Big) \\
&= O\Big(T^{\gamma-1} \ln(T^{1-\gamma}) \Big).
\end{aligned}$$

Since

$$\begin{aligned}
\ln(r_T^{(n)}) = \sum_{j=1}^{n} \Big\{ & \ln\big(1 - \exp(-\eta_T T d_T |\Theta_j|)\big) - \\
& \ln\big(\eta_T T d_T |\Theta_j|\big) \Big\}
\end{aligned}$$

and $\eta_T T d_T = O(T^{1+\delta-\beta}) \to \infty$ as $T \to \infty$, we have

$$\begin{aligned}
\ln(r_T^{(n)}) &= -\sum_{j=1}^{n} \ln(O(T^{1+\delta-\beta})) \\
&= -o(T^{1+\delta-\beta})
\end{aligned}$$

as n is a constant. Then we obtain

$$\begin{aligned}
B_T = O\Big\{ & T^{\beta+\gamma-1} \ln(T^{1-\gamma}) + T^{\beta+\gamma-1} \ln(T^{1-\beta+\delta}) \Big\} \\
& + o(1) = o(1).
\end{aligned}$$

since $\beta + \gamma < 1$. Therefore $A_T = \left\{ \omega \in \Omega : \frac{1}{T}\left\{ \sum_{t=1}^{T} \bar{s}(\theta, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \right\} \leq o(1) \right\}$, and $\mathbb{P}(A_T) \geq 1 - bT \exp\left[-(d_T - a)\lambda^{-1}\right]$. Furthermore, let $\bar{A}_T$ be the complement event of A_T, we have

$$\begin{aligned}
\sum_{T=1}^{\infty} \mathbb{P}(\bar{A}_T) &\leq \sum_{T=1}^{\infty} bT \exp\left[-(d_T - a)\lambda^{-1}\right] \\
&= \sum_{T=1}^{\infty} O(T \exp(-T^\delta)) < \infty.
\end{aligned}$$

So by Borel-Cantelli lemma, we have that $\mathbb{P}(\limsup_{T\to\infty} \bar{A}_T) = 0$, so $\mathbb{P}(\liminf_{T\to\infty} A_T) = 1$, which means from some time point on, $\frac{1}{T}\left\{ \sum_{t=1}^{T} \bar{s}(\theta, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \right\} \leq o(1)$ will

be true for all the following T's almost surely. Therefore we can conclude that

$$\limsup_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} \bar{s}(\theta, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} \le 0 \quad a.s.$$

(iii) If the scoring rule is convex, from Jensen's inequality we will have

$$s(\hat{\theta}_t, Y_t) \le \bar{s}(\theta, Y_t).$$

Then from the result in part (ii), we obtain

$$\limsup_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} \le 0 \quad a.s.$$

Furthermore from the result in part (i), we have

$$\liminf_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} \ge 0 \quad a.s.$$

Because $\limsup(\cdot) \ge \liminf(\cdot)$, in this case they must be equal and both equal to the limit. So we conclude

$$\lim_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t)\right\} = 0 \quad a.s.$$

(iv) The proof is similar to that of Theorem 2(iv). Basically, from Martingale convergence theorem we obtain

$$\frac{1}{T}\left\{\sum_{t=1}^{T} s(\hat{\theta}_t, Y_t) - \sum_{t=1}^{T} \bar{s}(\theta, Y_t)\right\} \underset{T\to\infty}{\longrightarrow} 0 \quad a.s.$$

which gives the result. ∎

Proofs for Section V

Proof of Theorem 4

Proof:

(i) For notational simplicity, we shall write $d_\Theta(0,\varepsilon) = \varepsilon^{1/u}$ as δ. Let $L_m, D_m, m = 1, \ldots, M_T$ denote the linear functions and segments that achieve the minimum in (28). We first prove the existence of $f_{i_1}, \ldots, f_{i_m} \in \mathcal{F}$ such that

$$\sum_{m=1}^{M_T} \sum_{t\in D_m} s(f_{i_m}(t), Y_t) \le \sum_{t=1}^{T} s(\theta_t^*, Y_t) + 2T^{2+\beta}\varepsilon + T^\beta \Delta_T$$

holds with high probability. Suppose that the linear function $L_m(t)$ for $t \in D_m$ has slope ζ and starts with value $L_m(t_{m-1}+1) = L_m(t_{m-1}) + \zeta$. Without loss of generality, suppose that $\zeta > 0$. There exist positive integers k_m, $n_m \le N_\varepsilon$ such that

$$\begin{aligned} k_m\delta \le \zeta < (k_m+1)\delta, \text{ and} \\ L_m(t_{m-1}) - \delta \le \theta_{(n_m + k_m t_{m-1}) \bmod N_\varepsilon} \\ < L_m(t_{m-1}). \end{aligned} \quad (A.61)$$

We choose i_m such that f_{i_m} is the parameter flow that maps t to $(n_m + k_m t) \bmod N_\varepsilon$. In particular, we obtain from (A.61) that

$$\begin{aligned} &d_\Theta(k_m\delta, \zeta) < \varepsilon, \\ &d_\Theta(f_{i_m}(t_{m-1}), L_m(t_{m-1})) < \varepsilon. \end{aligned} \quad (A.62)$$

The inequality (A.62) means that the starting point (resp. the slope) of the linear trend in D_m can be approximated by a quantizer (resp. linear flow over the quantizers), up to ε-error. Since $L_m(t) \in \Theta$ for all $t \in D_m$, the inequalities in (A.61) imply

$$f_{i_m}(t) = f_{i_m}(t_{m-1}) + (t - t_{m-1})k_m\delta \quad (A.63)$$

for all $t \in D_m$. Therefore, we have from (A.62) and (A.63) that

$$d_\Theta(f_{i_m}(t), L_m(t)) < (t - t_{m-1} + 1)\varepsilon \quad (A.64)$$

for all $t \in D_m$. For illustration, a list of the above notation is summarized in Table I. The same bound (A.64) applies for the cases $\zeta = 0$ and $\zeta < 0$.

TABLE I
SOME NOTATION USED IN THE PROOF OF (I)

t	$t_{m-1}+1$	$\cdots$	t_m
θ_{n+kt}	$\theta_{n+kt_{m-1}} + k\delta$	$\cdots$	$\theta_{n+kt_{m-1}} + (t_m - t_{m-1})k\delta$
$L_m(t)$	$L_m(t_{m-1}) + \zeta$	$\cdots$	$L_m(t_{m-1}) + (t_m - t_{m-1})\zeta$
θ_t^*	$\theta_{t_{m-1}+1}^*$	$\cdots$	$\theta_{t_m}^*$

Using Assumption (2″) and inequality (A.64), we obtain

$$\begin{aligned} &\sum_{m=1}^{M_T} \sum_{t\in D_m} s(f_{i_m}(t), Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \\ &= \sum_{m=1}^{M_T} \sum_{t\in D_m} \{s(f_{i_m}(t), Y_t) - s(\theta_t^*, Y_t)\} \\ &\le \sum_{m=1}^{M_T} \sum_{t\in D_m} Z_t \; d_\Theta(f_{i_m}(t), \theta_t^*) \\ &\le Z^{(T)} \sum_{t\in D_m} \Big(d_\Theta(f_{i_m}(t), L_m(t)) + d_\Theta(L_m(t), \theta_t^*) \Big) \\ &= Z^{(T)} \sum_{m=1}^{M_T} \sum_{t\in D_m} (t - t_{m-1} + 1)\varepsilon + \\ &\quad Z^{(T)}\Delta_T. \end{aligned} \quad (A.65)$$

where $Z_1, \ldots, Z_T \sim \mathrm{TE}(\lambda; a, b)$ are independent random variables, and $Z^{(T)} \triangleq \max_{t=1,\ldots,T} Z_t$. By similar arguments as in the proof of Theorem 2(i), we obtain from (A.65) and elementary inequalities that

$$\sum_{m=1}^{M_T} \sum_{t \in D_m} s(f_{i_m}(t), Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) < 2T^{2+\beta}\varepsilon + T^{\beta}\Delta_T \tag{A.66}$$

with probability at least $1 - c_3 T \exp(-c_4 T^{\beta})$ for some fixed constants $c_3, c_4 > 0$.
Finally, we define $\boldsymbol{p}_t^{\mathscr{F}}$ by

$$p_{t,n}^{\mathscr{F}} = (\sum_{j=1}^{N_\varepsilon} w_{t-1,j}^{\mathscr{F}})^{-1} w_{t-1,n}^{\mathscr{F}}, \quad n = 1, \ldots, N,$$

which denote the normalized predictive weights over N parameter flows. By the definitions of $\boldsymbol{p}_t$ and $\boldsymbol{p}_t^{\mathscr{F}}$, we have for each t that

$$\sum_{n=1}^{N_\varepsilon} p_{t,n}\ s(\theta_n^{(\varepsilon)}, Y_t) = \sum_{n=1}^{N} p_{t,n}^{\mathscr{F}}\ s(f_n(t), Y_t). \tag{A.67}$$

Combining (A.66) and (A.67), and using an adaptation of Lemma 3 (for the flows instead of bases), we obtain that

$$\begin{aligned}
&\sum_{n=1}^{N_\varepsilon} p_{t,n}\ s(\theta_n^{(\varepsilon)}, Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \\
&= \left\{ \sum_{n=1}^{N} p_{t,n}^{\mathscr{F}}\ s(f_n(t), Y_t) - \sum_{m=1}^{M_T} \sum_{t \in D_m} s(f_{i_m}(t), Y_t) \right\} \\
&\quad + \left\{ \sum_{m=1}^{M_T} \sum_{t \in D_m} s(f_{i_m}(t), Y_t) - \sum_{t=1}^{T} s(\theta_t^*, Y_t) \right\} \\
&\le \frac{M_T}{\eta} \log N - \frac{1}{\eta} \log\{\alpha^{M_T - 1}(1-\alpha)^{T - M_T}\} + \\
&\quad \frac{\eta}{8} T^{1+2\beta} + 2T^{2+\beta}\varepsilon + T^{\beta}\Delta_T
\end{aligned} \tag{A.68}$$

holds with probability at least

$$\begin{aligned}
&1 - c_1 T \exp(-c_2 T^{\beta}) - c_3 T \exp(-c_4 T^{\beta}) \\
&\ge 1 - C_1 T \exp(-C_2 T^{\beta})
\end{aligned}$$

where $C_1 \triangleq c_1 + c_3$, $C_2 \triangleq \min\{c_2, c_4\}$. Choose α and η as in (14), inequality (A.68) implies (29).

(ii) By inequality (A.50) and some calculations, the right hand side of inequality (29) is $o(T)$ as long as

$$\frac{M_T \log T^{\nu/u}}{T^{1-2\beta}} + \frac{M_T}{T^{1-4\beta}} + T^{1+\beta-\nu} + \frac{\Delta_T}{T^{1-\beta}} = o(1)$$

as $T \to \infty$. This can be satisfied by assumptions in (30). Therefore,

$$\frac{1}{T} \sum_{t=1}^{T} \left\{ \sum_{n=1}^{N_\varepsilon} w_{t,n}^{\theta}\ s(g_{\theta_n}, Y_t) - s(g_{\theta_t^*}, Y_t) \right\} \le c$$

holds with probability at least $1 - C_1 T \exp(-C_2 T^{\beta})$, for any fixed constant $c > 0$. Finally, from $\sum_{T=1}^{\infty} C_1 T \exp(-C_2 T^{\beta}) < \infty$ and using Borel-Cantelli lemma, we obtain (31).

(iii) & (iv)
The proof is similar to that of Theorem 2(iii)&(iv). ■

REFERENCES

[1] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Stat.*, vol. 16, no. 2, pp. 117–186, 1945.
[2] K. J. Arrow, D. Blackwell, and M. A. Girshick, "Bayes and minimax solutions of sequential decision problems," *Econometrica*, pp. 213–244, 1949.
[3] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.
[4] G. V. Moustakides *et al.*, "Optimal stopping times for detecting changes in distributions," *Ann. Stat.*, vol. 14, no. 4, pp. 1379–1387, 1986.
[5] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory Probab. Its Appl.*, vol. 8, no. 1, pp. 22–46, 1963.
[6] S. Roberts, "A comparison of some control chart procedures," *Technometrics*, vol. 8, no. 3, pp. 411–430, 1966.
[7] M. Pollak, "Optimal detection of a change in distribution," *Ann. Stat.*, pp. 206–227, 1985.
[8] T. Banerjee and G. V. Moustakides, "Minimax optimality of shiryaev-roberts procedure for quickest drift change detection of a brownian motion," *Sequential Analysis*, vol. 36, no. 3, pp. 355–369, 2017.
[9] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.
[10] D. Siegmund, *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.
[11] R. A. Davis, D. Huang, and Y.-C. Yao, "Testing for a change in the parameter values and order of an autoregressive model," *Ann. Statist.*, pp. 282–304, 1995.
[12] C. Incln and G. C. Tiao, "Use of cumulative sums of squares for retrospective detection of change of variance," *J. Amer. Statist. Assoc.*, vol. 89, pp. 913–923, 1994.
[13] G. E. Berkes, I. and L. Horváth, "Testing for changes in the covariance structure of linear processes," *J. Stat. Plan. Inference*, vol. 139, pp. 2044–2063, 2009.
[14] D. Picard, "Testing and estimating change-points in time series," *Adv. Appl. Probab.*, vol. 17, pp. 841–867, 1985.
[15] Y.-C. Yao, "Estimating the number of change-points via schwarz'criterion," *Stat. Probab. Lett.*, vol. 6, no. 3, pp. 181–189, 1988.
[16] E. S. Venkatraman, "Consistency results in multiple change-point problems," Ph.D. dissertation, to the Department of Statistics.Stanford University, 1992.
[17] C. Du, C.-L. M. Kao, and S. Kou, "Stepwise signal extraction via marginal likelihood," *J. Amer. Statist. Assoc.*, vol. 111, no. 513, pp. 314–330, 2016.
[18] J. Ding, Y. Xiang, L. Shen, and V. Tarokh, "Multiple change point analysis: Fast implementation and strong consistency," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4495–4510, 2017.
[19] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc. Ser. B*, pp. 267–288, 1996.

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

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

[22] 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.

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

[24] R. T. Sutton and D. L. Hodson, "Atlantic ocean forcing of north american and european summer climate," *Science*, vol. 309, no. 5731, pp. 115–118, 2005.

[25] T. Dangl and M. Halling, "Predictive regressions with time-varying coefficients," *J. Financ. Econ.*, vol. 106, no. 1, pp. 157–181, 2012.

[26] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[27] M. B. Priestley, "Evolutionary spectra and non-stationary processes," *J. Roy. Statist. Soc. Ser. B*, pp. 204–237, 1965.

[28] Y. Xiang, J. Ding, and V. Tarokh, "Evolutionary spectra based on the multitaper method with application to stationarity test," *to appear in IEEE Trans. Signal Process.*, 2018.

[29] R. Dahlhaus, "On the kullback-leibler information divergence of locally stationary processes," *Stoch. Proc. Appl.*, vol. 62, no. 1, pp. 139–168, 1996.

[30] L. Cohen, *Time-frequency analysis*. Prentice hall, 1995, vol. 778.

[31] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. OUP Oxford, 2012, vol. 38.

[32] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[33] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "Smc2: an efficient algorithm for sequential analysis of state space models," *J. Roy. Statist. Soc. Ser. B*, vol. 75, no. 3, pp. 397–426, 2013.

[34] J. Fan and W. Zhang, "Statistical estimation in varying coefficient models," *Ann. Stat.*, pp. 1491–1518, 1999.

[35] G. E. Primiceri, "Time varying structural vector autoregressions and monetary policy," *Rev. Econ. Stud.*, vol. 72, no. 3, pp. 821–852, 2005.

[36] L. Tian, D. Zucker, and L. Wei, "On the cox model with time-varying regression coefficients," *J. Am. Stat. Assoc.*, vol. 100, no. 469, pp. 172–183, 2005.

[37] Q. Han, J. Ding, E. M. Airoldi, and V. Tarokh, "SLANTS: sequential adaptive nonlinear modeling of time series," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 4994–5005, 2017.

[38] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[39] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Ann. Stat.*, pp. 561–592, 2012.

[40] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, "Bayesian model comparison with the Hyvarinen score: computation and consistency," *J. Am. Stat. Assoc.*, 2018.

[41] A. Shiryayev, *Selected Works of AN Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*. Springer Science & Business Media, 2013, vol. 27.

[42] M. Herbster and M. K. Warmuth, "Tracking the best expert," *Machine learning*, vol. 32, no. 2, pp. 151–178, 1998.

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

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

[45] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

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

[47] T. Li, S. Sun, T. P. Sattar, and J. M. Corchado, "Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches," *Expert Syst. Appl.*, vol. 41, no. 8, pp. 3944–3954, 2014.

[48] J. S. Liu, R. Chen, and T. Logvinenko, "A theoretical framework for sequential importance sampling with resampling," in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 225–246.

[49] N. Chopin, "A sequential particle filter method for static models," *Biometrika*, vol. 89, no. 3, pp. 539–552, 2002.

[50] R. Douc and O. Cappé, "Comparison of resampling schemes for particle filtering," in *IEEE ISPA Conf.*, 2005, pp. 64–69.

[51] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[52] C. P. Robert, *Monte carlo methods*. Wiley Online Library, 2004.

[53] R. Dahlhaus, "Asymptotic statistical inference for nonstationary processes with evolutionary spectra," in *Athens conference on applied probability and time series analysis*. Springer, 1996, pp. 145–159.

[54] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Stat.*, pp. 689–705, 1985.

[55] R. F. Engle and C. W. Granger, "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, pp. 251–276, 1987.

[56] C. W. Granger and H. S. Lee, "An introduction to time-varying parameter cointegration," in *Economic Structural Change*. Springer, 1991, pp. 139–157.

[57] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econom.*, vol. 31, no. 3, pp. 307–327, 1986.

[58] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, pp. 987–1007, 1982.

[59] O. E. Barndorff-Nielsen and N. Shephard, "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *J. Roy. Statist. Soc. Ser. B*, vol. 64, no. 2, pp. 253–280, 2002.

[60] A. P. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *J. Roy. Statist. Soc. Ser. A*, pp. 278–292, 1984.

[61] A. J. Patton, "Volatility forecast comparison using imperfect volatility proxies," *J. Econom.*, vol. 160, no. 1, pp. 246–256, 2011.

[62] M. Csörgo, "On the strong law of large numbers and the central limit theorem for martingales," *Trans. Am. Math Soc.*, vol. 131, no. 1, pp. 259–275, 1968.

[63] J. Mogyoródi, "A central limit theorem for the sum of a random number of independent random variables," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 7, no. 3, pp. 409–424, 1962.