

TAKDE: Temporal Adaptive Kernel Density Estimator for Real-Time Dynamic Density Estimation

Yinsong Wang, Yu Ding, *Senior Member, IEEE*, and Shahin Shahrampour, *Senior Member, IEEE*

Abstract—Real-time density estimation is ubiquitous in many applications, including computer vision and signal processing. Kernel density estimation is arguably one of the most commonly used density estimation techniques, and the use of “sliding window” mechanism adapts kernel density estimators to dynamic processes. In this paper, we derive the asymptotic mean integrated squared error (AMISE) upper bound for the “sliding window” kernel density estimator. This upper bound provides a principled guide to devise a novel estimator, which we name the temporal adaptive kernel density estimator (TAKDE). Compared to heuristic approaches for “sliding window” kernel density estimator, TAKDE is theoretically optimal in terms of the worst-case AMISE. We provide numerical experiments using synthetic and real-world datasets, showing that TAKDE outperforms other state-of-the-art dynamic density estimators (including those outside of kernel family). In particular, TAKDE achieves a superior test log-likelihood with a smaller run-time.

Index Terms—Real-time Density Estimation, Kernel Density Estimation, Adaptive Estimation, Asymptotic Mean Integrated Squared Error.



1 INTRODUCTION

THIS work is concerned with estimation and tracking of dynamic probability density functions in real time, motivated by a nanoscience application. The introduction of *in situ* transmission electron microscope (TEM) technology [1] allows the growth of nanoparticles to be captured in real time and has the potential to enable precise control in nanoparticle self-assembly processes. Part of the underlying nanoscience problem is framed into a learning problem with the following characteristics [2]: (1) Estimation and tracking of a time-varying probability density function that reflects the collective changes across ensembles of the nano objects. (2) It seems inevitable to adopt a non-parametric approach in the density tracking, because there is no settled parametric density function that can adequately describe growth mechanisms in a multi-stage nanoparticle growth process [1], [3]. (3) In order to be useful for in-process decision making, the density estimation and tracking needs to be conducted in real time. By “real-time” we mean that the learning and computation speed ought to be fast enough relative to the imaging rate (or the data arrival rate in general), which is 15 frames per second (fps) in [1]. While the research is motivated by the dynamic nano imaging, we believe that the aforementioned characteristics are rather common in many types of dynamic streaming data, brought forth in various applications by fast-pace data collection capability. The objective of this research is to present one

competitive solution for dynamic density estimation and tracking.

On the subject of density estimation, kernel density estimator has had great success (in terms of accuracy) for static datasets [4]. The direct adaptation of kernel density estimator to dynamic density estimation [5] is infeasible as the memory and computation cost constantly scale with the total number of incoming data points. [6] further shows that even with unlimited computation and storage resources, a traditional kernel density estimator will only be a consistent estimator for a few specific dynamic systems. [2] also shows that traditional kernel density estimation falls short in practice in dynamic density estimation due to limited data availability.

To address the disadvantages of traditional kernel density estimator in dynamic density estimation, most researchers resort to the “sliding window” mechanism [7], [8], [9]. For example, [7] proposed the M-kernel algorithm, where the contribution of each data point in the “sliding window” is approximated as an additional weight added to the kernel density at the closest grid point. This approach manages to keep the memory and computation costs within budget despite the growth of the total number of data points. However, with a poor choice of grid points, it can suffer from either over-fitting or under-fitting. [8] employed cluster kernel and resampling technique to improve the merger performance. This approach uses the exponentially decaying weight scheme to capture the dynamic of the true density. [10] proposed the local region kernel density estimator (LRKDE), where the kernel bandwidth varies in different regions. The regions are divided such that the sum of data variances in each region is minimized. LRKDE also uses a “sliding window” to capture the dynamic of the true density. [9] further improved upon the previous works by using linear interpolation with kernel densities at grid points to

-
- Y. Wang and S. Shahrampour are with the Department of Mechanical and Industrial Engineering at Northeastern University, Boston, MA 02115 USA.
E-mail: wang.yinso@northeastern.edu
E-mail: s.shahrampour@northeastern.edu
 - Y. Ding is with the Wm Michael Barnes '64 Department of Industrial and Systems Engineering at Texas A&M University, College Station, TX 77843 USA.
E-mail: yuding@tamu.edu

approximate the kernel density estimator and then updating the kernel densities at the grid points with data points within a "sliding window".

The "sliding window" kernel density estimators do not only use the data points at the current time stamp, and they take into account older data points for inferring the current distribution. Intuitively, this mechanism provides two improvements that allow the kernel density estimator to work well in dynamic density estimation. First, defining a window size according to the computation and memory limit of the learning machine can alleviate the scalability issue of the kernel density estimator as old data points that are irrelevant to the current distribution can be discarded. Second, including older data points in the window can help alleviate the low data volume issue for most streaming data applications. However, to the best of our knowledge, all "sliding window" kernel density estimators proposed so far focus on modifying the kernel density estimator itself, and less attention has been given to the "sliding window" mechanism. As the only component that addresses the "dynamic" part of dynamic density estimation, there is no answer regarding how this mechanism affects the performance of the estimation.

We note that there also exists another line of works that model the dynamic density transition using a dynamical system with a fixed number of parameters. One class of frameworks is based on Bayesian learning [11], [12], [13], which models the prior with an evolving Dirichlet process called dependent Dirichlet process, where the dependence between a class of Dirichlet processes is defined by a covariate. When using the covariate to describe time, the dependent Dirichlet process can be used to model the evolution of the dynamic distribution. The computation and memory costs are also maintained at a constant level. Another approach [2] couples B-spline with Kalman filter to capture the density evolution with a state space model. It imposes space continuity with B-spline smoothing and time continuity with Kalman filter to develop a fast density estimator for real-time process control. However, these estimators always need a normalization process with numerical operations to return a proper density function. For real-time density estimation tasks that require a model update cycle in the order of sub-second, these methods may not be ideal as we will later show in simulations.

In this paper, we propose the temporal adaptive kernel density estimator (TAKDE), a novel kernel density estimator for real-time dynamic density estimation that is theoretically optimal in terms of the worst-case asymptotic mean integrated squared error (AMISE). For the first time, we derive the AMISE upper bound for the "sliding window" kernel density estimator in a dynamic density estimation context. The minimizer of the upper bound entails a novel sequence for bandwidth selection and data weighting, which forms the basis of TAKDE. We provide numerical experiments on synthetic datasets to support our theoretical claim, and we then use several real-world datasets to show that TAKDE outperforms other state-of-the-art fast dynamic density estimators, such as the B-spline Kalman Filter [2] and KDEtrack [9] in terms of mean test log-likelihood metric. Interestingly, TAKDE also dominates these algorithms in terms of achieving a smaller run-time.

The organization of the paper is as follows. We present in Section 2 the preliminaries, including definitions and notations used throughout the paper. In Section 3, we present the details for TAKDE design, which addresses three important questions, i.e., the selection of window size, bandwidth and the data weights. We provide in Section 4 numerical experiments with synthetic and real datasets to demonstrate the performance of TAKDE. Finally, we draw conclusions and discuss the potential and limitations of TAKDE in Section 5.

2 PRELIMINARIES

2.1 Kernel Density Estimation: A Brief Overview

The kernel density estimator for a given set of data points $\{x_i\}_{i=1}^n$ is as follows

$$\hat{p}(x; \sigma) = \frac{1}{n} \sum_{i=1}^n K_\sigma(x - x_i), \quad (1)$$

where $K_\sigma(\cdot)$ is the kernel function with the bandwidth σ . Throughout this paper, $K(\cdot)$ denotes a standard kernel function with a unit kernel bandwidth. We have that $K_\sigma(x) = \frac{1}{\sigma} K(\frac{x}{\sigma})$. We further impose the following mild assumptions on the kernel function $K(\cdot)$.

Assumption 1. [14] *The bandwidth sequence σ_n (the subscript n shows the dependence of σ to the number of data points) has the following properties*

$$\begin{aligned} \lim_{n \rightarrow \infty} \sigma_n &= 0 \\ \lim_{n \rightarrow \infty} n\sigma_n &= \infty, \end{aligned} \quad (2)$$

which implies that the bandwidth σ_n decays slower than n^{-1} and converges to 0. The standard kernel function $K(\cdot)$ is a bounded, symmetric probability density function with a zero first moment and a finite second moment. That is, the following properties hold

$$\begin{aligned} \int K(x) dx &= 1 \\ \int xK(x) dx &= 0 \\ \int x^2 K(x) dx &< \infty. \end{aligned} \quad (3)$$

The convergence to 0 for bandwidth is rather intuitive, in that when we have infinitely many data points at hand, our estimator can be as flexible as possible without having to be concerned about over-fitting. It is also easy to verify that many commonly used kernels (e.g., the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$) satisfy (3).

2.2 Problem Formulation

In dynamic estimation, the density evolves over time. The evolution might be continuous in nature, but we only observe samples from time to time. Here, we consider the case where the streaming data comes in batches. We first define the dynamic streaming dataset, where we observe one new batch of data points $\mathbf{x}^{(t)} = \{x_i^{(t)} \in \mathbb{R}\}_{i=1}^{n_t}$ at a new time stamp t . This data structure applies to most real-world streaming datasets. An important example is estimating density information in video datasets [2] like the dynamic

nano imaging problem mentioned in the introduction. An image processing tool extracts the sizes of nanoparticles as the sample points for estimating the normalized particle size distribution (NSPD), which is an indicator to anticipate and detect phase changes in the nanoparticle growth. This data structure further applies to many time-series datasets [15]. For the cases where streaming data comes in on a per point basis, one can convert those types of data into our defined structure through combining consecutive data points into batches.

We assume data points $\mathbf{x}^{(t)}$ are generated independently from $p_t(x)$, the true density at time stamp t . Also, the data points $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(t')}$ in different time stamps ($t \neq t'$) are independent from each other. We impose the following assumption on the true density function.

Assumption 2. *The true density function $p_t(x)$ at any time stamp t is twice differentiable, and its second derivative $p_t''(x)$ is continuous and square integrable.*

Assumption 2 is commonly used for continuous density functions [14]. The square integrable condition is necessary as the integrated second order Taylor expansion appears later in the error bound derivation.

Following (1), we write the traditional kernel density estimator of the density $p_t(x)$ as follows

$$\hat{p}_t(x; \sigma) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_\sigma(x - x_i^{(t)}). \quad (4)$$

The "sliding window" kernel density estimator, popularly used in dynamic density estimation [7], [8], [9], takes the following form

$$\hat{h}_t(x) = \sum_{j \in \mathcal{T}_t} \alpha_j^{(t)} \hat{p}_j(x; \sigma_j^{(t)}), \quad (5)$$

where \mathcal{T}_t represents the set of batches within the moving window (memory), \hat{p}_j is defined following (4), and $\alpha_j^{(t)}$ is a non-negative weight sequence that satisfies $\sum_{j \in \mathcal{T}_t} \alpha_j^{(t)} = 1$, to ensure that the output is a proper density function. The window size is T_t , i.e. $|\mathcal{T}_t| = T_t$, so that \mathcal{T}_t can be naturally written as $\mathcal{T}_t = \{t - T_t + 1, \dots, t\}$. The superscripts (t) on α and σ are omitted hereafter for the presentation clarity.

In order to develop a fast real-time estimator, we need to address the following three problems.

Problem 1. *How do we choose the set \mathcal{T}_t to have a good enough "memory" for estimating the density at time t while maintaining real-time processing?*

Problem 2. *How do we design the weight sequence in (5)?*

Problem 3. *How do we devise a kernel bandwidth selector in (4)?*

3 ALGORITHM DESIGN

In this section, we derive the AMISE upper bound for the general "sliding window" kernel density estimator in (5). We then present a novel weight and bandwidth sequence, entailed by the upper bound minimizer (Problems 2-3). We use these sequences to design the TAKDE algorithm.

3.1 Asymptotic Mean Integrated Squared Error Upper Bound

AMISE is a popular metric used to theoretically evaluate the performance of a density estimator [14]. For a given density estimator $\hat{h}(x)$ of a density function $p(x)$, the mean integrated squared error (MISE) is defined as follows

$$\begin{aligned} MISE(\hat{h}, p) &\triangleq \int \mathbb{E}[(\hat{h}(x) - p(x))^2] dx \\ &= \int MSE(\hat{h}, p) dx, \end{aligned} \quad (6)$$

where the expectation is taken with respect to the distributions of data points involved in estimator \hat{h} . MISE is the integration of the mean squared error of the density estimator over the support. [14] shows that the asymptotic expression (with respect to the sample size n) of the MISE for a standard kernel density estimator $\hat{p}(x; \sigma_n)$ with kernel bandwidth σ_n is

$$AMISE(\hat{p}, p) = \frac{R(K)}{n\sigma_n} + \frac{1}{4}\sigma_n^4\mu_2^2(K)R(p''), \quad (7)$$

where

$$\begin{aligned} R(f) &= \int f^2(x) dx, \\ \mu_2(f) &= \int x^2 f(x) dx. \end{aligned} \quad (8)$$

We can see that the conditions in (2) guarantee that AMISE converges to zero as $n \rightarrow \infty$. The MISE and AMISE have been popular measures for characterizing non-parametric density estimators, including binned density estimator [16], kernel density estimator [14], wavelet density estimator [17], and diffusion estimator with a static limit [18]. The exact expression for kernel density estimator can also be derived in the case of specific distributions like Gaussian distribution [14]. However, all these derivations assume that data points in the non-parametric density estimator are samples from a static target density function.

In the following theorem, we derive the theoretical upper bound of AMISE for the "sliding window" kernel density estimator given in (5) in the context of dynamic density estimation. To the best of our knowledge, this is the first AMISE bound for "sliding window" kernel density estimator in estimating the evolving true density $p_t(x)$.

Theorem 1. *Let Assumptions 1-2 hold. The AMISE of a "sliding window" kernel density estimator \hat{h}_t at time t with window size $|\mathcal{T}_t| = T_t$, weight sequence $\{\alpha_i\}_{i=1}^{T_t}$, and bandwidth sequence $\{\sigma_i\}_{i=1}^{T_t}$ has the following upper bound*

$$\begin{aligned} AMISE(\hat{h}_t, p_t) &\leq \sum_{i \in \mathcal{T}_t} \frac{\alpha_i^2}{n_i \sigma_i} R(K) \\ &\quad + (2T_t - 1) \sum_{i \in \mathcal{T}_t} \alpha_i^2 R(b_i^{(t)}) \\ &\quad + \frac{2T_t - 1}{4} \mu_2^2(K) \sum_{i \in \mathcal{T}_t} \alpha_i^2 \sigma_i^4 R(p_i''), \end{aligned} \quad (9)$$

where $b_i^{(j)}(x)$ defines the difference between density functions $p_i(x), p_j(x)$ ($j \geq i$)

$$b_i^{(j)}(x) \triangleq p_i(x) - p_j(x). \quad (10)$$

Proof. We omit the superscript (t) for weight α and bandwidth σ for the presentation clarity. First, recall the definition of $\hat{h}_t(x)$ from (4)-(5), where we have

$$\hat{h}_t(x) = \sum_{i \in \mathcal{T}_t} \alpha_i \hat{p}_i(x; \sigma_i) = \sum_{i \in \mathcal{T}_t} \frac{\alpha_i}{n_i} \sum_{j=1}^{n_i} K_{\sigma_i}(x - x_j^{(i)}). \quad (11)$$

The bias of the estimator can be written as

$$\begin{aligned} B(\hat{h}_t(x)) &\triangleq \mathbb{E}[\hat{h}_t(x) - p_t(x)] \\ &= \mathbb{E}\left[\sum_{i \in \mathcal{T}_t} \frac{\alpha_i}{n_i} \sum_{j=1}^{n_i} K_{\sigma_i}(x - x_j^{(i)}) - p_t(x)\right] \\ &= \sum_{i \in \mathcal{T}_t} \alpha_i \int K_{\sigma_i}(x - y) p_i(y) dy - p_t(x) \\ &= \sum_{i \in \mathcal{T}_t} \alpha_i (K_{\sigma_i} * p_i)(x) - p_t(x), \end{aligned} \quad (12)$$

where $*$ denotes the convolution, and $p_i(\cdot)$ is the true density of batch i .

Using $V(\cdot)$ to denote the variance operator, the estimator variance can be calculated as

$$V(\hat{h}_t(x)) = \sum_{i \in \mathcal{T}_t} \alpha_i^2 V(\hat{p}_i(x; \sigma_i)), \quad (13)$$

due to the independence of batches, where

$$V(\hat{p}_i(x; \sigma_i)) = \frac{1}{n_i} \left((K_{\sigma_i}^2 * p_i)(x) - (K_{\sigma_i} * p_i)^2(x) \right). \quad (14)$$

The decomposition of the MSE of the "sliding window" estimator \hat{h}_t is as follows

$$\begin{aligned} MSE(\hat{h}_t, p_t) &= \mathbb{E}[(\hat{h}_t(x) - p_t(x))^2] \\ &= V(\hat{h}_t(x)) + B^2(\hat{h}_t(x)). \end{aligned} \quad (15)$$

Integrating above over x , we have

$$MISE(\hat{h}_t, p_t) = \int MSE(\hat{h}_t, p_t) dx. \quad (16)$$

Given the expressions of bias (12) and variance (14), to calculate AMISE, we need to derive the Taylor approximations of the following quantities

$$\begin{aligned} (K_{\sigma_i}^2 * p_i)(x) \\ (K_{\sigma_i} * p_i)(x). \end{aligned} \quad (17)$$

First, we have

$$\begin{aligned} (K_{\sigma_i}^2 * p_i)(x) &= \int K_{\sigma_i}^2(x - y) p_i(y) dy \\ &= \frac{1}{\sigma_i} \int K^2(z) p_i(x - \sigma_i z) dz \\ &= \frac{p_i(x)}{\sigma_i} R(K) + o(1), \end{aligned} \quad (18)$$

where we note that $p_i(x - \sigma_i z) = p_i(x) + o(1)$ holds, because $\sigma_i \rightarrow 0$ as $n_i \rightarrow \infty$. We also have that

$$\begin{aligned} (K_{\sigma_i} * p_i)(x) &= \int K_{\sigma_i}(x - y) p_i(y) dy \\ &= \int K(z) p_i(x - \sigma_i z) dz \\ &= \int K(z) (p_i(x) - \sigma_i z p_i'(x) \\ &\quad + \frac{1}{2} \sigma_i^2 z^2 p_i''(x) + o(\sigma_i^2)) dz \\ &= p_i(x) + \frac{1}{2} \sigma_i^2 p_i''(x) \mu_2(K) + o(\sigma_i^2). \end{aligned} \quad (19)$$

where we used the assumptions that $\int K(z) dz = 1$ and $\int z K(z) dz = 0$. Given the above asymptotic characterization of the quantities, we can rewrite the bias term (12) as

$$B(\hat{h}_t(x)) = \sum_{i \in \mathcal{T}_t} \left(\alpha_i b_i^{(t)}(x) + \frac{1}{2} \alpha_i \sigma_i^2 p_i''(x) \mu_2(K) + o(\sigma_i^2) \right). \quad (20)$$

We can also write the variance (14) as

$$V(\hat{h}_t(x)) = \sum_{i \in \mathcal{T}_t} \left(\frac{\alpha_i^2}{n_i \sigma_i} R(K) p_i(x) + o\left(\frac{1}{n_i \sigma_i}\right) \right). \quad (21)$$

We can now simplify the MSE (15) as

$$\begin{aligned} MSE(\hat{h}_t, p_t) &= \sum_{i \in \mathcal{T}_t} \left(\frac{\alpha_i^2}{n_i \sigma_i} R(K) p_i(x) + o\left(\frac{1}{n_i \sigma_i}\right) \right) \\ &\quad + \left(\sum_{i \in \mathcal{T}_t} \alpha_i b_i^{(t)}(x) + \sum_{i \in \mathcal{T}_t} \frac{1}{2} \sigma_i^2 \alpha_i p_i''(x) \mu_2(K) + \sum_{i \in \mathcal{T}_t} o(\sigma_i^2) \right)^2. \end{aligned} \quad (22)$$

Disregarding the terms that converge to zero and taking integral over x , we can derive an upper bound for AMISE as

$$\begin{aligned} AMISE(\hat{h}_t, p_t) &\leq \sum_{i \in \mathcal{T}_t} \frac{\alpha_i^2}{n_i \sigma_i} R(K) \\ &\quad + (2|\mathcal{T}_t| - 1) \sum_{i \in \mathcal{T}_t} \alpha_i^2 R(b_i^{(t)}) \\ &\quad + \frac{2|\mathcal{T}_t| - 1}{4} \mu_2^2(K) \sum_{i \in \mathcal{T}_t} \alpha_i^2 \sigma_i^4 R(p_i''), \end{aligned} \quad (23)$$

where the last two lines follow from the Cauchy-Schwarz inequality for the $2|\mathcal{T}_t| - 1$ terms in the square. Note that $b_i^{(t)} = 0$ by definition. Observing that $|\mathcal{T}_t| = T_t$ completes the proof of Theorem 1. \square

Let us call the three lines in the right hand side of (9) as term 1, term 2, and term 3, respectively. Term 1 is due to the variance of the estimator, and terms 2 and 3 are the bias terms. Terms 1 and 3 are asymptotically vanishing in the sense that when $n_i \rightarrow \infty$, they both go to zero per condition (2). We can make several observations about the upper bound expression (9). First, the dynamic density estimation with "sliding window" kernel density estimators will have a non-vanishing error term 2, induced by keeping densities of various time stamps in the memory. We will later see in Corollary 3 that under optimal weight design, this term can also go to zero when $n_i \rightarrow \infty$. Second, when the distribution evolution is mild (i.e., $R(b_i^{(t)})$ is small), there can be a theoretical advantage in including previous samples in the memory to reduce the variance term 1. Later simulations will show this advantage can be significant in practice. Third, when the previous distributions are very different from the current distribution, it is desirable to only keep one batch (the current batch) in the memory, i.e., $\mathcal{T}_t = \{t\}$ and $T_t = 1$. In this case, $R(b_i^{(t)}) = 0$ by definition (10) and the upper bound (9) exactly recovers the AMISE for the traditional kernel density estimator in (7).

3.2 Window Generator

In the existing literature, kernel density estimators are modified using arbitrary "sliding windows" to adapt to the

dynamic estimation. This approach performs better than the traditional kernel density estimator, as a static kernel density estimator works poorly for dynamic density estimation [2]. However, this heuristic approach lacks a theoretical justification. In fact, based on the theoretical upper bound of AMISE (9), it is intuitive that the window size should depend on the density evolution to keep the AMISE small. For example, when the true density changes drastically, it is ideal to decrease the window size to adapt to the fast density change. Therefore, we propose a histogram-based window size generator that will allow the kernel density estimator to be adaptive to dynamic changes.

We observe in (9) that compared to the static AMISE, the worst-case AMISE for dynamic density estimation depends on one more quantity, namely the difference function $b_i^{(t)}$. In principle, we can use this quantity as an indicator to adapt the dynamic kernel density estimator to the changes in the underlying density function.

We define a cutoff threshold to determine the number of batches (sliding window size) to be kept in the memory of the dynamic kernel density estimator. In doing so, we first define the temporal adaptive (TA) distance between two density functions. Here, we use histograms to approximate the density functions as true density functions are unavailable. We denote the number of bins in the histograms by m , set using the Sturges' rule [19]

$$m = 1 + 3.322 \log n, \quad (24)$$

where n is the smallest batch size among all batches in the current memory. Sturges' rule is a widely adopted, simple binning algorithm in the literature. It is derived for normally distributed data. The user can choose other binning rules, such as Doane's rule [20], Scott's rule [21], or Freedman and Diaconis's rule [22] as appropriate. However, we note that all existing binning guidelines provide bins similar to Sturges' rule under low data volume (less than 200) [4].

The temporal adaptive distance between two histograms $hist_i$ and $hist_j$ is expressed as

$$\|hist_i, hist_j\|_{TA} \triangleq \|\mathbf{y}_i - \mathbf{y}_j\|_2, \quad (25)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm and \mathbf{y}_i is the probability mass vector on bins in batch i , i.e., $\|\mathbf{y}_i\|_1 = 1$. This TA distance serves as a measure proportional to $\hat{R}(b_i^{(t)})$, the approximation of $R(b_i^{(t)})$ in (9), i.e.,

$$\hat{R}(b_i^{(t)}) \propto \|hist_i, hist_t\|_{TA}. \quad (26)$$

To control the bias, one can set a cutoff threshold s for the TA distance. Upon receiving batch t , the number of batches to be kept in the memory can be set as T_t that satisfies the following two inequalities

$$\sum_{j=t-T_t}^{t-1} \|hist_j, hist_t\|_{TA} > s, \quad \sum_{j=t-T_t+1}^{t-1} \|hist_j, hist_t\|_{TA} \leq s. \quad (27)$$

Note that from a practical standpoint, the cutoff threshold s should not be the only criterion for window selection, because when the true density goes through a long static period, it is possible that (27) will induce a large memory window that exceeds the computational limit for real-time density estimation. Therefore, there should exist a hard

cap w to account for computational limits. Combining both considerations, the actual number of batches in the memory should be set as $\min(T_t, w)$.

Remark 1. Note that the main purpose of cutoff value s is to reduce the window size (and computation cost) when dealing with rapidly changing densities. The bias-variance decomposition suggests that including more batches in TAKDE can induce a lower variance (first term in equation (21)) at the cost of increasing the bias (first term in equation (20)). Moreover, we will show in Corollary 3 that TAKDE is consistent regardless of window size T_t . Later, synthetic data simulation also suggests the empirical performance difference is not too sensitive to the cutoff value, so one can heuristically choose it in favor of fast processing rather than through intensive cross-validation.

3.3 Bandwidth and Weight Generator

The dynamic nature of the underlying true density makes it practically impossible to understand the actual difference functions and the second derivative of the true densities. However, using the AMISE upper bound in Theorem 1, we can find theoretically optimal sequences for kernel bandwidths and weights, which in turn helps in the algorithm design. In view of Theorem 1, we present the following corollary.

Corollary 2. The optimal sequences of weights and bandwidths that minimize the AMISE upper bound of the dynamic kernel density estimator are as follows

$$\begin{aligned} \sigma_i &= \left[\frac{R(K)}{n_i \mu_2^2(K) R(p'_i) (2T_t - 1)} \right]^{\frac{1}{5}}, \\ \alpha_i &= \frac{1/S_i}{\sum_{j \in \mathcal{T}_t} 1/S_j}, \end{aligned} \quad (28)$$

where the sequence S_i (with superscript (t) omitted) is such that

$$S_i = \frac{5R(K)}{4n_i \sigma_i} + (2T_t - 1)R(b_i^{(t)}). \quad (29)$$

Proof. Equation (23) shows that the upper bound on AMISE depends on the weight sequence α_i and the bandwidth sequence σ_i . Therefore, we can minimize the upper bound with respect to both of these parameters.

Differentiating with respect to σ_i yields the following (optimal) sequence

$$\sigma_i = \left[\frac{R(K)}{n_i \mu_2^2(K) R(p'_i) (2T_t - 1)} \right]^{\frac{1}{5}}. \quad (30)$$

We can find the optimal sequence of weights by simply solving the minimization of Lagrangian of (23) with the constraint $\sum \alpha_i = 1$ and incorporating (30). This will result in the following expression for the sequence α_i

$$\begin{aligned} S_i &= \frac{5R(K)}{4n_i \sigma_i} + (2T_t - 1)R(b_i^{(t)}) \\ \alpha_i &= \frac{1/S_i}{\sum_{j \in \mathcal{T}_t} 1/S_j}, \end{aligned} \quad (31)$$

which completes the proof of Corollary 2. \square

Remark 2. Corollary 2 provides some insights concerning the bandwidth and weight choices.

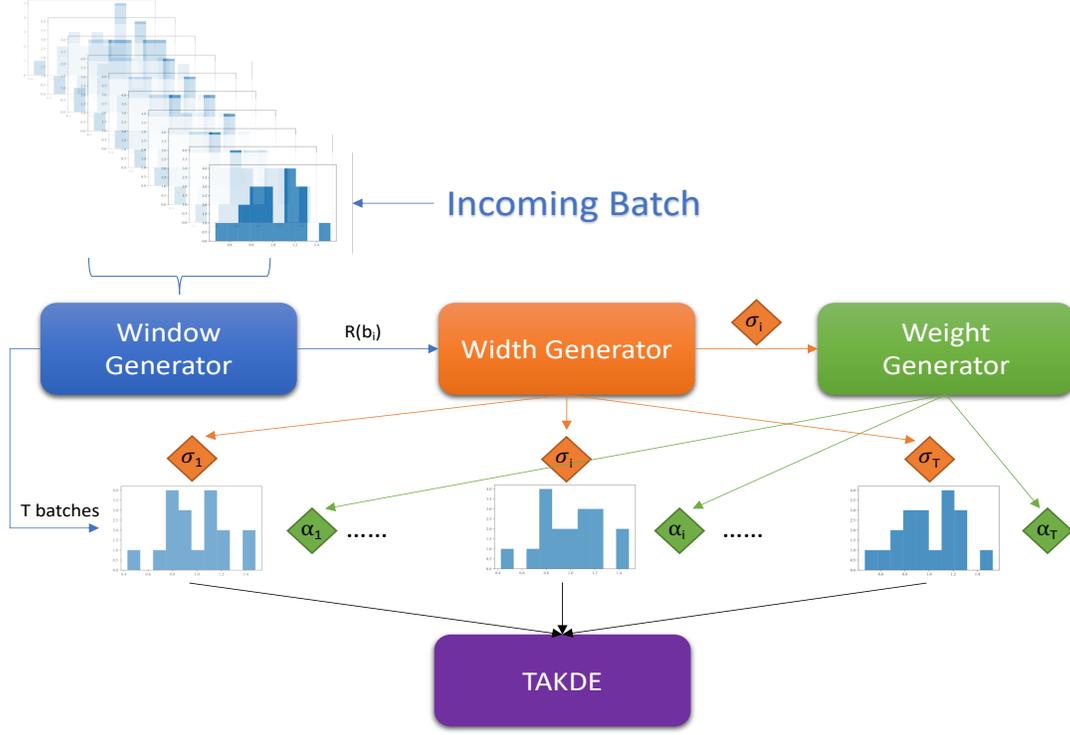


Fig. 1. TAKDE framework.

- 1) The bandwidth sequence suggests that we should make the kernel more flexible as more batches of data points are included in the estimation. This aligns with the intuition from the traditional kernel density estimator, where the estimator can be more flexible with more sample points.
- 2) The weight sequence provides the following insights. First, when the number of data points at a particular batch is considerably large, we should assign more weight to that batch with the hope of extracting more information to infer the current density. Second, the $R(b_i^{(t)})$ quantity provides a countermeasure to prevent us from assigning a large weight to data points coming from a very different distribution compared to the current batch. Third, we should assign more weights to the batches with larger kernel bandwidths, which means we are favoring smoother estimators in principle.

Corollary 3. Under Assumptions 1-2, the optimal weight sequence and kernel bandwidth sequence in Corollary 2 will ensure that for any $\epsilon > 0$,

$$\Pr(|\hat{h}_t - p_t|^2 > \epsilon) \rightarrow 0, \quad (32)$$

as $n_i \rightarrow \infty$.

Proof. First, notice that following Corollary 2, we have $\sigma_i \rightarrow 0$ and $\alpha_i \rightarrow 0$ for every batch except the last batch where $\alpha_t \rightarrow 1$ (since $R(b_i^{(t)}) = 0$) as $n_i \rightarrow \infty$. It is easy to verify that $\mathbb{E}[|\hat{h}_t - p_t|^2] \rightarrow 0$ under this bandwidth and weight sequence, based on the expression of the mean squared error in (22). Then, by Markov inequality, we have

$$\Pr(|\hat{h}_t - p_t|^2 > \epsilon) \leq \frac{\mathbb{E}[|\hat{h}_t - p_t|^2]}{\epsilon} \rightarrow 0. \quad (33)$$

The proof is complete. \square

Corollary 3 shows that TAKDE is weakly consistent as $n_i \rightarrow \infty$ regardless of T_t . This is rather intuitive as TAKDE can precisely recover the traditional KDE in this extreme case.

3.4 Kernel Bandwidth Selector

The bandwidth sequence in Corollary 2 presents a principle for choosing the kernel bandwidth. However, the quantity $R(p_i'')$ is unknown in practice, and we still need to find a kernel bandwidth selector to calculate the actual kernel bandwidth values. There exist extensive studies for the choice of bandwidth in traditional kernel density estimation. One popular choice is the cross-validation approach [23], [24], [25], [26]. However, the computational cost of cross-validation prohibits its application in high-frequency density estimation as every new batch of data points needs to be cross-validated for a new kernel bandwidth.

Minimizing AMISE in (7) reveals a simple expression for the optimal kernel bandwidth. [14] characterized the optimal kernel bandwidth based on (7) as follows

$$\sigma_{AMISE} = \left[\frac{R(K)}{n\mu_2^2(K)R(p'')} \right]^{\frac{1}{5}}. \quad (34)$$

We notice that (34) coincides with the optimal kernel bandwidth sequence we derived in Corollary 2 except for a factor of $(2T_t - 1)^{1/5}$. This relationship allows us to directly adopt existing kernel bandwidth selection methods for optimal AMISE.

Expression (34) is still dependent on the unknown $R(p'')$, but there exist a number of studies that explore different

methods for estimating $R(p'')$. For example, [27] approximates the AMISE objective function assuming the density is Poisson and then proceeds to find the minimizer as the optimal kernel bandwidth. However, this method is not applicable in real-time dynamic density estimation as the optimization process is expensive. [9] provides an iterative update framework by estimating $R(p'')$ through $R(\hat{p}'')$, which is the numerical square integration of the second derivative of the density estimator. This approach does not impose any strict assumption on the underlying distribution, which offers a robust estimation of $R(p'')$. However, the iterative algorithm still requires numerical operations like numerical derivatives and numerical integration, which may not be efficient enough for real-time density estimation.

In TAKDE, we adopt the normal rule introduced in [28]. Assuming the true density is Gaussian, the optimal kernel bandwidth can be approximated as follows

$$\sigma_{AMISE} \approx c\hat{\sigma}n^{-\frac{1}{5}}, \quad (35)$$

where c is the smoothness parameter depending on the kernel function and the underlying true density, and $\hat{\sigma}$ is the sample standard deviation of the data points. The normal rule is particularly appealing for the design of TAKDE due to its simple structure, which allows a direct plug-in of smoothness parameter c and enables fast real-time processing.

There are two commonly used recommendations for the smoothness parameter c in (35). The first choice given in [14] is as follows

$$\sigma_{AMISE} \approx \left[\frac{8\pi^{1/2}R(K)}{3\mu_2^2(K)n} \right]^{\frac{1}{5}} \hat{\sigma}, \quad (36)$$

where $\hat{\sigma}$ is the estimated standard deviation assuming the true density is normal. The smoothness parameter c of Gaussian Kernel in this setting is $(32/3)^{1/5}$.

The second recommendation [29] comes from the upper bound of the AMISE-optimal kernel bandwidth using $beta(4, 4)$ or triweight density function, that is,

$$\sigma_{AMISE} \leq \left[\frac{243R(K)}{35\mu_2^2(K)n} \right]^{\frac{1}{5}} \hat{\sigma}. \quad (44)$$

This bandwidth provides an oversmoothed density estimator that might not perform well with respect to metrics like log-likelihood or MSE. However, an oversmoothed density estimator is often preferred for real-world applications, because the results are visually plausible. In this case, the smoothness parameter c of Gaussian Kernel is $(972/35\sqrt{\pi})^{1/5}$.

Remark 3. *The only reason for adopting the normal rule in TAKDE is its computation simplicity. We must note that the weight sequence given in Corollary 2 is compatible with any existing $R(p'')$ approximation method.*

3.5 Algorithm Design

In this subsection, we present the final form of TAKDE. The algorithm requires as input a cutoff value s , a hard cap w , a smoothness parameter c , and a kernel function K . Upon receiving the batch of data points at time t , the window generator decides the set of batches \mathcal{T}_t to be used for the density estimation. The window generator will also return the sequence of approximated $\hat{R}(b_j^{(t)})$ as in (26) for all

Algorithm 1 Temporal Adaptive Kernel Density Estimator (TAKDE)

Input: Kernel function $K(\cdot)$, cutoff value s , hard cap w , smoothness parameter c .

For $t = 1, 2, \dots$

1: Receive new batch of data $\mathbf{x}^{(t)}$ at time t .

2: **Window Generator:** Generate and record $hist_t$ and forget $hist_{t-w}$. Set $Distance = 0, T_t = 0, \mathcal{T}_t = \emptyset$.

While $T_t < w$:

$$Distance = Distance + \|hist_t, hist_{t-T_t}\|_{TA} \quad (37)$$

Break If:

$$Distance > s, \quad (38)$$

Else:

$$\mathcal{T}_t = \mathcal{T}_t \cup \mathbf{x}^{(t-T_t)} \quad T_t = T_t + 1. \quad (39)$$

Return: \mathcal{T}_t and T_t and the sequence $\{\hat{R}(b_j^{(t)})\}_{j \in \mathcal{T}_t}$ where

$$\hat{R}(b_j^{(t)}) = m \|hist_j, hist_t\|_{TA}. \quad (40)$$

3: **Bandwidth Generator:** Receive the batch set \mathcal{T}_t .

For $j \in \{t - T_t + 1, \dots, t\}$:

$$\sigma_j = \frac{c\hat{\sigma}_j}{((2T_t - 1)n_j)^{1/5}}, \quad (41)$$

where c is defined by the kernel bandwidth selector, $n_j = |\mathbf{x}^{(j)}|$, and $\hat{\sigma}_j$ is the sample standard deviation of data in batch j .

Return: Bandwidth sequence σ_j .

4: **Weight Generator:** Receive bandwidth sequence σ_j and the approximated $\hat{R}(b_j^{(t)})$ sequence. Let

$$\alpha_j = \frac{1/S_j}{\sum_{i \in \mathcal{T}_t} 1/S_i}, \quad (42)$$

$$S_j = \frac{5R(K)}{4n_j\sigma_j} + (2T_t - 1)\hat{R}(b_j^{(t)}).$$

Return: Weight sequence α_j .

Output: The Temporal Adaptive Kernel Density Estimator given as

$$\hat{h}_t(x) = \sum_{j \in \mathcal{T}_t} \alpha_j \hat{p}_j(x; \sigma_j), \quad (43)$$

$$\hat{p}_j(x; \sigma_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} K_{\sigma_j}(x - x_i^{(j)}).$$

batches in the memory. Then, all batches within the memory will be fed into the bandwidth generator to generate the sequence of kernel bandwidths σ_j as in Corollary 2. Then, the approximated $\hat{R}(b_j^{(t)})$ and bandwidth sequence σ_j will be fed into the weight generator to generate the sequence α_j as in Corollary 2. Finally, all parameters will be put together to generate a proper kernel density estimator for estimating the density at time t . Fig. 1 illustrates the workflow of TAKDE. The algorithmic presentation of TAKDE is outlined in Algorithm 1.

4 EXPERIMENT

We now present numerical experiments to verify the efficiency of TAKDE both on synthetic data and real-world data. All experimental results established in this section are based on Gaussian kernel function.

4.1 Algorithm Design Evaluation

Before we compare TAKDE with other established benchmark algorithms, we evaluate the design of TAKDE on synthetic data. The specific question that we aim to address is that whether our proposed weighting scheme, derived from the AMISE upper bound, outperforms other heuristic weight sequences such as uniform (or average) weighting and exponentially decaying weighting.

4.1.1 Synthetic Dataset Design

We create a synthetic dataset to test the performance of TAKDE in dynamic density estimation. We design the synthetic dataset following some general principles.

- 1) The true densities involved in the generation of the dataset need to have analytical forms and have already been established in the literature.
- 2) Each batch of data points has a size in the range of [5, 20], so that the batches do not differ too drastically in terms of the data amount.
- 3) The number of testing points for all batches should be the same for comparison purposes.
- 4) The dynamics of the underlying densities varies for different batches.

Following the above principles, we adopt the 15 Gaussian mixture densities, recommended by [30], as the baseline densities for our synthetic dataset design. The 15 densities are shown in Fig. 2.

To design the true density, we first consider 14 sections, where each section consists of multiple batches. Let us denote the 15 Gaussian mixtures with $g_1(x), \dots, g_{15}(x)$ and represent the 14 sections with $\mathcal{S}_1, \dots, \mathcal{S}_{14}$, where $|\mathcal{S}_1| + \dots + |\mathcal{S}_{14}|$ equals to the total number of batches in the dataset. To be specific, section \mathcal{S}_i has $|\mathcal{S}_i|$ consecutive batches of data points in it, and the first batch of data in section \mathcal{S}_{i+1} will start after the last batch in section \mathcal{S}_i . For batch i in section j , where $1 \leq i \leq |\mathcal{S}_j|$, the density function $h_i^{(j)}(x)$ is defined as follows

$$h_i^{(j)}(x) = \frac{|\mathcal{S}_j| - i + 1}{|\mathcal{S}_j|} g_j(x) + \frac{i - 1}{|\mathcal{S}_j|} g_{j+1}(x). \quad (45)$$

To be consistent with our previous notation, $h_i^{(j)}(x) = p_{t_{ij}}(x)$ for $t_{ij} = |\mathcal{S}_1| + \dots + |\mathcal{S}_{j-1}| + i$. Notice that in section j , the j -th Gaussian mixture linearly transforms to the $j+1$ -th Gaussian mixture. After we move on to section $j+1$, none of previous Gaussian mixtures $g_1(x), \dots, g_j(x)$ will appear in the section. Given the density of batch i in section j , we sample a random number between 5 to 20 as the number of training points and 500 for testing points to perform the comparison. To account for the randomness in partitioning the batches into 14 sections and the randomness in samples, we generate 300 synthetic datasets for Monte-Carlo simulations.

4.1.2 TAKDE Evaluation

We now compare the weight generator in TAKDE with two heuristic approaches in the literature. One approach is to assign uniform weights to the batches, assuming older data points are of the same importance as the new data points, and the other one is to assign exponentially decaying weights, assuming the new points are much more important [8], [9]. To ensure a fair comparison, we only change the weight generator of TAKDE to uniform and exponential weighting, and we keep the other components of the algorithm unchanged. The uniform weight sequence is set as follows

$$\alpha_j = \frac{1}{T_t}, \forall j \in \{t - T_t + 1, \dots, t\}. \quad (46)$$

The exponential weight sequence is set as follows

$$\alpha_j = (1 - e)e^{t-j}, \forall j \in \{t - T_t + 2, \dots, t\}, \quad (47)$$

and $\alpha_{t-T_t+1} = e^{T_t-1}$, where e is the decay ratio. We compare the above to α_j corresponding to the expression in (42). In our simulation, $e = 0.9$ in general yields the best result under different settings; therefore, the decay ratio for exponential weight sequence is set to $e = 0.9$.

Our comparison is performed under several kernel bandwidth selectors, including the normal selector and oversmooth selector mentioned in Section 3.4 and under various cutoff values.

First, we consider normal bandwidth selector (36) and oversmooth bandwidth selector (44). For each bandwidth selector, we conduct the comparison with datasets having from 100 to 500 batches of data to reflect different underlying dynamics. Notice that for the data with 100 batches, the dynamic change is more drastic than that of the data with 500 batches.

The simulation result is shown in Fig. 3. We can observe that TAKDE with AMISE-based weight sequence dominates the uniform and exponential weight sequences in terms of the test log-likelihood. We also see that when using the heuristic weight sequences, increasing the memory (i.e., larger cutoff value) mostly exacerbates the density estimation performance. The results show that the performance difference between TAKDE and other two methods is larger when the total number of batches is smaller. This suggests that TAKDE with AMISE-based weight sequence is better at adapting to more drastic dynamic changes. The smaller differences in 500-batch simulations are consistent with our theoretical results, where the weighting sequence in Corollary 2 gets closer to uniform weighting as $R(b_i^{(t)})$ converges to 0, equivalent to a static density estimation. We observe that changes in the cutoff value do not have a significant effect on TAKDE performance compared to others. This verifies our discussion in Remark 1.

Second, we conduct the comparison using a synthetic dataset with 100 batches of data for different bandwidth selectors, i.e., varying the smoothness parameter c in (35). The simulation results are shown in Fig. 4. Again, we observe the same performance trend for the algorithms. These simulations empirically verify that the performance advantage of our proposed weight sequence against the heuristic weight sequences is robust to different kernel bandwidths and different window sizes.

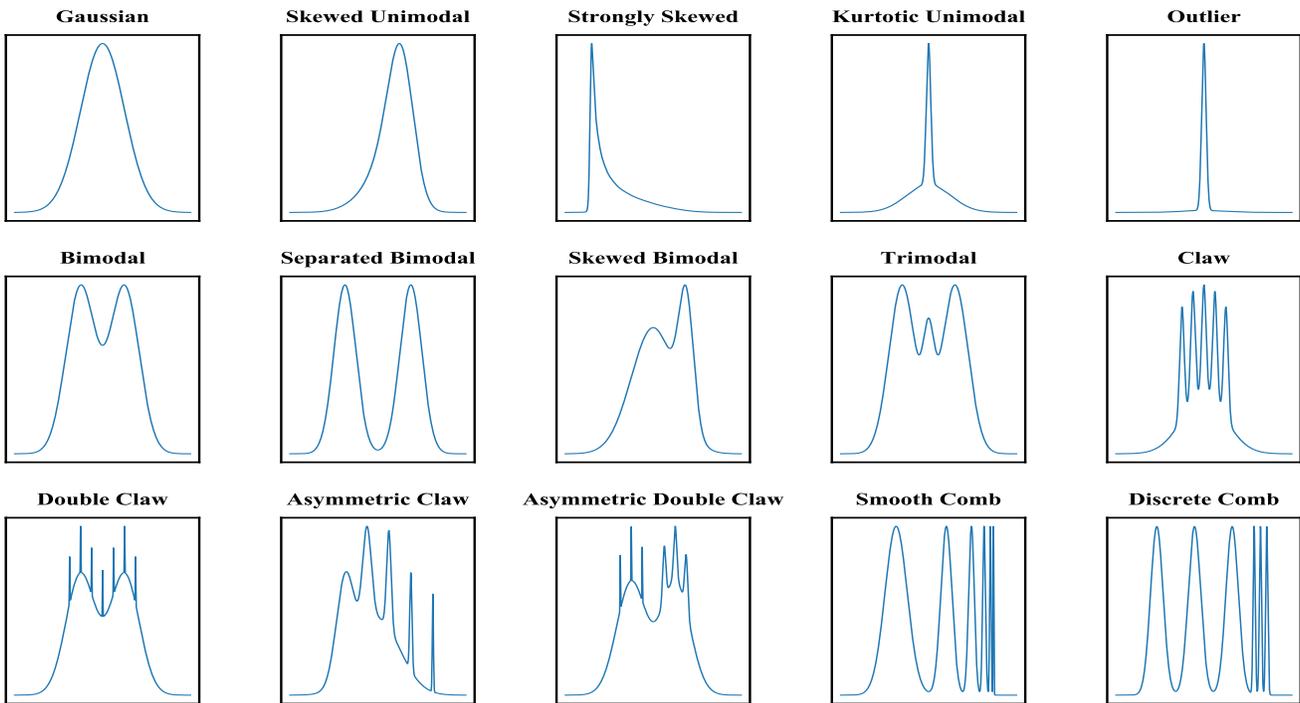


Fig. 2. The 15 Gaussian mixture densities used in the synthetic dataset design.

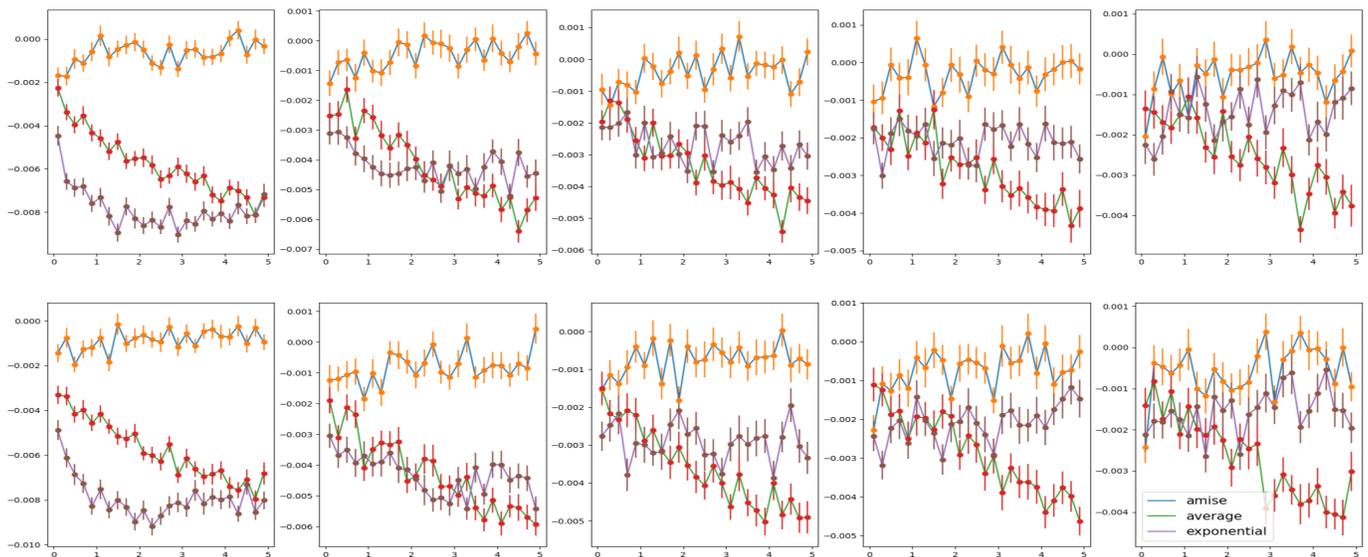


Fig. 3. The test log-likelihood comparison between TAKDE vs. the heuristic approaches. The x-axis represents the cutoff value and the y-axis represents the test log-likelihood. The first row shows the result under normal bandwidth selector and the second row shows the result under oversmooth bandwidth selector. In each row, the plots from left to right represent the simulation results using synthetic datasets with 100, 200, 300, 400, 500 batches of data.

4.2 Comparison with Benchmark Algorithms

4.2.1 Benchmark Algorithms

Next, we compare TAKDE with three density estimation methods on real-world datasets. We consider both the mean test log-likelihood and the run-time to show the advantages of TAKDE.

- 1) **Kernel Density Estimator (KDE):** The first benchmark algorithm is the traditional kernel density estimator. The main reason to include kernel density estimator in the comparison is to show why a traditional density estimator is not ideal for dynamic

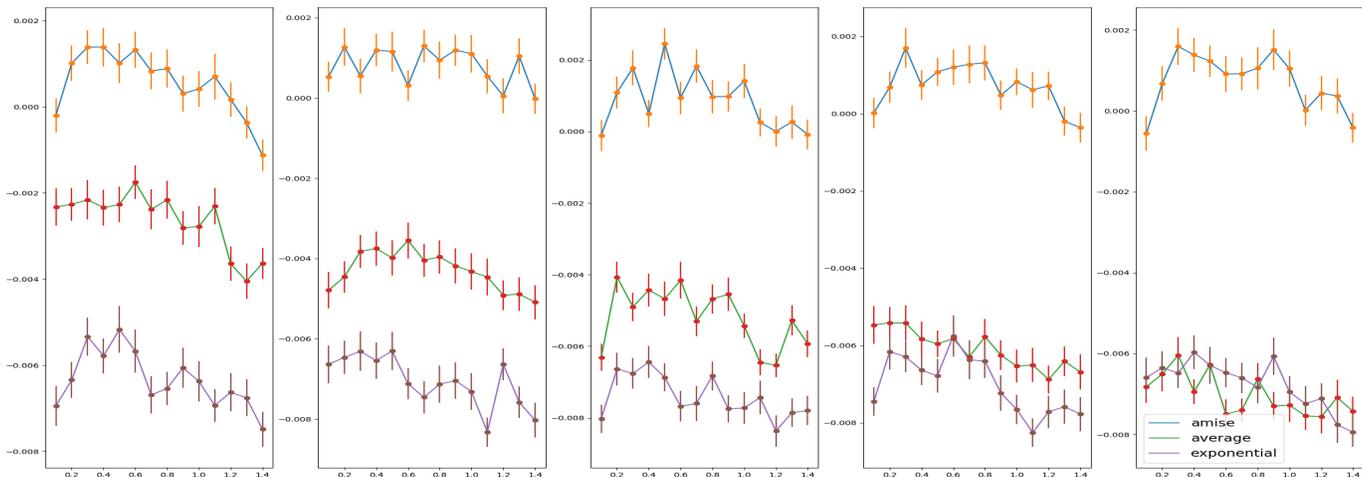


Fig. 4. The test log-likelihood comparison between TAKDE vs. the heuristic approaches over different bandwidth selectors. The x-axis represents the value of the smoothness parameter c . The y-axis represents the test log-likelihood. Each plot from left to right represents the simulation conducted with cutoff values from 1 to 5.

TABLE 1
The best experimental settings for different benchmark algorithms in different datasets.

Algorithm	Noise Parameter α_1	Noise Parameter α_2	Smoothness Parameter c	Cutoff Value (Window Size)
TEM				
KDE	-	-	1.34	-
B-spline	0.66	0.04	-	-
KDEtrack	-	-	0.45	1(16)
TAKDE	-	-	0.15	1(16)
ECG				
KDE	-	-	0.98	-
B-spline	0.82	0.05	-	-
KDEtrack	-	-	0.1	1(60)
TAKDE	-	-	0.7	1(60)
Wafer				
KDE	-	-	0.22	-
B-spline	0.96	0.06	-	-
KDEtrack	-	-	1.05	1(20)
TAKDE	-	-	0.15	1(20)
Earth				
KDE	-	-	0.4	-
B-spline	0.81	0.05	-	-
KDEtrack	-	-	0.8	0.2(15)
TAKDE	-	-	0.05	1.4(90)
Star				
KDE	-	-	0.9	-
B-spline	0.30	0.02	-	-
KDEtrack	-	-	1	0.3(20)
TAKDE	-	-	0.35	1.8(38)

density estimation. The kernel density estimator is formulated as (1). The bandwidth selector is

$$\sigma = c\hat{\sigma}n^{-\frac{1}{5}}, \quad (48)$$

where we use cross-validation to choose c (rather than the actual bandwidth) for easy comparison with TAKDE.

- 2) **B-spline Kalman Filter (BKF) [2]:** B-spline Kalman filter models the underlying density function as a count measure defined on the partitions of the density support. The density estimator is defined as

$$\hat{p}(x) = \frac{1}{C} \exp \sum_{i=1}^m \beta_i B_i(x), \quad (49)$$

where C is the normalization constant calculated with numerical integration, m is the number of partitions, and $B_i(x)$ are the B-spline bases. The algorithm updates its states β_i using a B-spline matrix evaluated on the centers of the density support partitions and the count vector at each batch.

- 3) **KDEtrack [9]:** KDEtrack partitions the support of the density using a collection of grid points. The set of grid points and the density values at the grid points are updated after each new batch of data points is received and evaluated. The density evaluation at a test point will be the linear interpolation at the test point using the closest grid points.

Remark 4. We do not include the M-kernel and LRKDE methods

TABLE 2
Mean test log-likelihood on five real datasets.

Algorithm	TEM	ECG	Wafer	Earth	Star
KDE	-0.026 ± 0.0001	0.060 ± 0.00002	0.0229 ± 0.0007	0.048 ± 0.0002	0.0078 ± 0.00002
B-spline Kalman Filter	0.171 ± 0.0062	1.580 ± 0.0011	1.204 ± 0.0034	1.324 ± 0.0051	0.685 ± 0.0024
KDEtrack	0.245 ± 0.0057	1.095 ± 0.0009	0.866 ± 0.0018	0.915 ± 0.0012	0.640 ± 0.0007
TAKDE(normal)	0.130 ± 0.0016	1.639 ± 0.0004	1.530 ± 0.0015	1.247 ± 0.0017	0.696 ± 0.0007
TAKDE(cor)	0.246 ± 0.0022	1.625 ± 0.0010	1.627 ± 0.0017	1.331 ± 0.0026	0.705 ± 0.0008
TAKDE	0.362 ± 0.0036	1.648 ± 0.0009	1.848 ± 0.0025	1.504 ± 0.0026	0.710 ± 0.0012

TABLE 3
Run-time comparison (seconds) on five real datasets.

Algorithm	TEM	ECG	Wafer	Earth	Star
B-spline Kalman Filter	7.08	4.099	0.379	0.907	1.752
KDEtrack	5.461	4.712	1.542	1.569	14.85
TAKDE	0.378	0.557	0.114	0.704	0.851

since [9] has showed that KDEtrack is superior to these two methods.

4.2.2 Datasets

- **In situ TEM video data:** The first dataset we use is in situ TEM dataset introduced in Section 1. It is the 76.6 second in situ TEM video published in [1]. It has a total of 1149 frames of images and 5 – 20 particle counts in each frame.
- **CinCECGTorso (ECG) data:** CinCECGTorso dataset is an ECG dataset taken from multiple torso surface sites of four patients from the Computers in Cardiology challenges. This dataset is available on UCR time-series data archive [15].

The dataset consists of ECG measurements of four patients. We use the ECG signal sequence of one person to highlight the density dynamics over time. Note that simulations on all four patients yield similar results. There are 342 ECG signals (data points) available at each batch, and there are a total of 1639 batches of data points over time. The batches are collected at 2-kHz frequency, which requires the density estimator to be updated 2000 times per second. For each batch of data points at a certain time stamp, we randomly sample 5 to 20 data points to train and use the rest of the data points to evaluate the algorithms. The number of training data points at each batch is determined only once throughout all the Monte-Carlo simulations. However, the set of training points are sampled randomly in each Monte-Carlo simulation.

- **Wafer data:** Wafer dataset is a collection of sensor readings in a semiconductor wafer manufacturing process over time, available on UCR time-series data archive [15]. Unlike the previous two datasets, a wafer manufacturing process is a rather slow process that could span over 10 weeks. However, this dataset is still illustrative for evaluating the accuracy of TAKDE. We use the readings in the normal state wafer manufacturing process to conduct our analysis. There are 600 readings (data points) available at each batch, and there are a total of 152 batches of data points over time. Again, we adopt the same train-test

split approach as in the ECG dataset.

- **Earthquakes (Earth) data:** The earthquake dataset is a sensor reading dataset from Northern California Earthquake Data Center available on UCR time-series data archive [15]. It consists of 461 readings at each batch with a total of 512 batches.
- **StarLight Curves (Star) data:** The starlight curves dataset consists of time-series sensor readings on the brightness of a collection of celestial objects. It is also available on UCR time-series data archive [15]. This dataset includes the readings of 1000 celestial objects at each batch with a total of 1024 batches.

4.2.3 Experimental Settings

In comparing across different density estimators, we only present the best performance of B-spline Kalman filter, where the noise prior parameters are cross-validated using a grid search with an interval size of 0.01. For the traditional kernel density estimator, we report its best performance, but even that is significantly inferior to other density estimators. For KDEtrack and TAKDE, we report the best settings performances (in terms of smoothness parameter c and cutoff value s). Notice we do not adopt the iterative bandwidth update in KDEtrack for the computation reason explained in Section 3.4, but instead we use the same bandwidth generator as in TAKDE. All the simulations are conducted over 100 Monte-Carlo simulations for random training-testing splits to generate the standard errors of the performance. The performance metric is the mean test log-likelihood of the test points.

4.2.4 Performance

The parameter settings leading to respective best performance for all benchmark algorithms are shown in Table 1. These settings are cross-validated using the first 10% batches of each dataset (20% for Wafer and Earth dataset).

The results are tabulated in Table 2. TAKDE tagged with "(normal)" represents the performance achieved with smoothness parameter recommended in equation (36) (normal bandwidth selector) and the optimal cutoff in Table 1. TAKDE tagged with "(cor)" represents the performance achieved by TAKDE under KDEtrack best settings in terms of

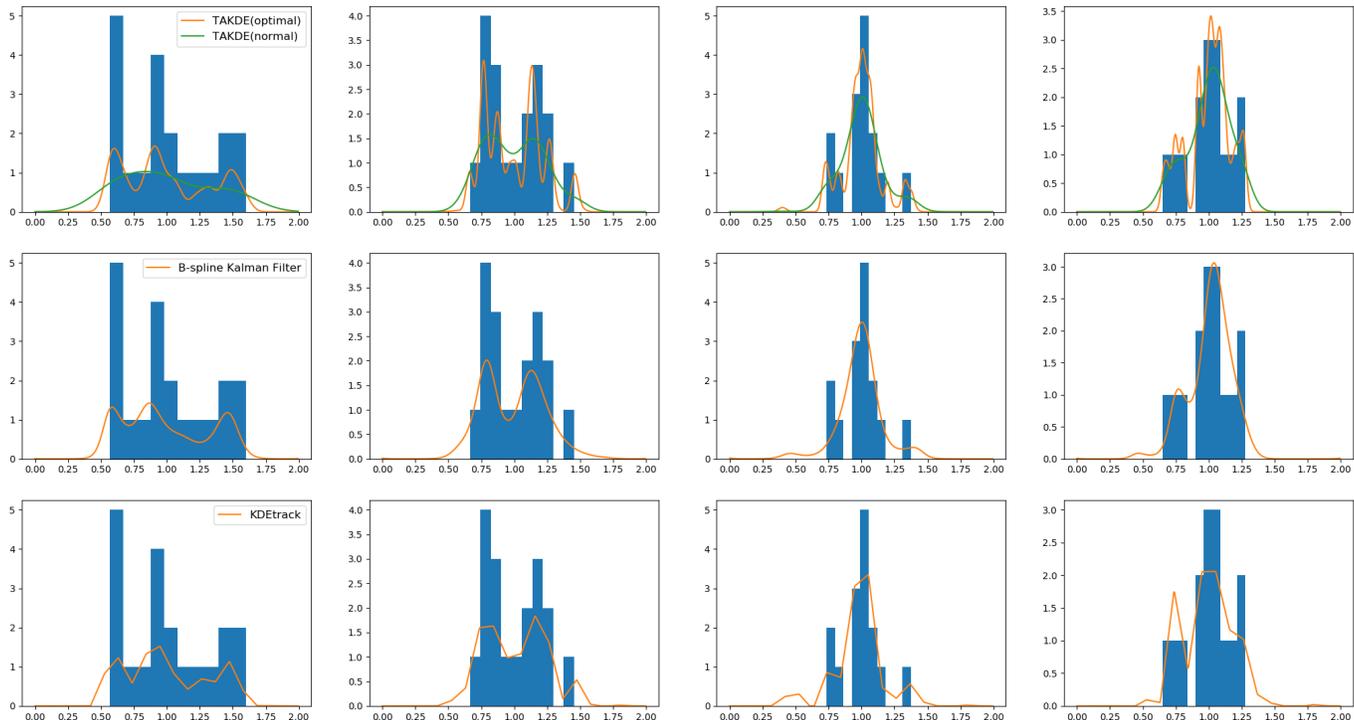


Fig. 5. Visualization of the density estimators on the TEM dataset. The first row shows TAKDE at its normal setting and optimal setting. The second row shows B-spline Kalman Filter at its optimal setting. The third row shows KDEtrack at its optimal setting. Figures from left to right represent the estimation at time stamps 225, 450, 675, and 900, respectively.

cutoff value and smoothness parameter. As we can observe, TAKDE dominates all other benchmark algorithms in terms of test log-likelihood by a large margin. TAKDE is also robust with respect to different cutoff values and different smoothness parameters, as it dominates all other benchmark algorithms even under the best settings for KDEtrack. The only exception is TAKDE with normal bandwidth selector on the TEM dataset. The underlying reason is that the low data volume available at different batches (training and testing combined) forces the "true" density distribution at each time stamp to an average of Dirac measures, which is far from the normal assumption of the normal bandwidth selector.

The run-time comparisons are shown in Table 3. The values represent the time used for executing the density estimation for all test data points in all batches. We can observe that in addition to being more accurate than the benchmark algorithms, TAKDE is also much faster in speed as it requires negligible calculations in addition to kernel density evaluation. The computation advantage makes a huge difference for the ECG dataset in particular, as the other two benchmark algorithms do not run nearly fast enough to catch up with the 2kHz data collection rate.

4.3 Visual Examination

In this subsection, we visualize the previously compared density estimators. We pick the time stamps $\{225, 450, 675, 900\}$ in 1150 batches of data in the TEM dataset for visualization. The results are shown in Fig. 5. As we can observe, TAKDE at its optimal setting (for test log-likelihood) yields a more flexible model compared to other algorithms. TAKDE with

normal smoothness parameter yields the smoothest model among all. Our results in Table 2 also show that the normal smoothness parameter can achieve estimation performance close to the optimal setting while yielding smooth density functions that facilitate easy interpretation. For this reason, in most real-world applications that do not place estimation accuracy as their first priority, we do recommend using the normal smoothness parameter (36) to avoid cross-validation.

5 CONCLUSION

In this paper, we established a theoretical AMISE upper bound expression for the "sliding window" kernel density estimator in dynamic density estimation. We proposed the temporal adaptive kernel density estimator that maintains the fast processing advantage of the "sliding window" kernel density estimator, while being theoretically optimal under the worst-case AMISE. We provided extensive numerical simulations to verify that TAKDE is superior to state-of-the-art real-time dynamic density estimators in terms of the mean test log-likelihood. TAKDE also dominated these algorithms in terms of achieving smaller run-times.

The proposed weight sequence is reminiscent of the attention mechanism in a transformer neural network for sequence re-weighting [31]. Considering the massive success of transformers in different fields, one of the future research directions is to see whether learning the weight sequence through the attention mechanism can result in a better performance.

Note that TAKDE in its current state only works for univariate density estimation. Thus, another future direction is to extend it to multivariate density cases.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NSF Award #2038625 as part of the NSF/DHS/DOT/NIH/USDA-NIFA Cyber-Physical Systems Program.

REFERENCES

- [1] H. Zheng, R. K. Smith, Y. Jun, C. Kisielowski, U. Dahmen, and A. P. Alivisatos, "Observation of single colloidal platinum nanocrystal growth trajectories," *Science*, vol. 324, no. 5932, pp. 1309–1312, 2009.
- [2] Y. Qian, J. Z. Huang, C. Park, and Y. Ding, "Fast dynamic nonparametric distribution tracking in electron microscopic data," *The Annals of Applied Statistics*, vol. 13, no. 3, pp. 1537–1563, 2019.
- [3] T. J. Woehl, C. Park, J. E. Evans, I. Arslan, W. D. Ristenpart, and N. D. Browning, "Direct observation of aggregative nanoparticle growth: Kinetic modeling of the size distribution and growth rate," *Nano letters*, vol. 14, no. 1, pp. 373–378, 2014.
- [4] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.
- [5] M. Kristan, D. Skočaj, and A. Leonardis, "Online kernel density estimation for interactive learning," *Image and Vision Computing*, vol. 28, no. 7, pp. 1106–1116, 2010.
- [6] H. Hang, I. Steinwart, Y. Feng, and J. A. Suykens, "Kernel density estimation for dynamical systems," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1260–1308, 2018.
- [7] A. Zhou, Z. Cai, L. Wei, and W. Qian, "M-kernel merging: Towards density estimation over data streams," in *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications, 2003.(DASFAA 2003)*, 2003, pp. 285–292.
- [8] C. Heinz and B. Seeger, "Cluster kernels: Resource-aware kernel density estimators over streaming data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 7, pp. 880–893, 2008.
- [9] A. Qahtan, S. Wang, and X. Zhang, "KDE-track: An efficient dynamic density estimator for data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 3, pp. 642–655, 2016.
- [10] A. P. Boedihardjo, C.-T. Lu, and F. Chen, "A framework for estimating complex probability density structures in data streams," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008, pp. 619–628.
- [11] F. Caron, M. Davy, and A. Doucet, "Generalized pólya urn for time-varying dirichlet process mixtures," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, 2007, pp. 33–40.
- [12] A. Rodriguez and E. Ter Horst, "Bayesian dynamic density estimation," *Bayesian Analysis*, vol. 3, no. 2, pp. 339–365, 2008.
- [13] R. H. Mena and M. Ruggiero, "Dynamic density estimation with diffusive dirichlet mixtures," *Bernoulli*, vol. 22, no. 2, pp. 901–926, 2016.
- [14] M. P. Wand and M. C. Jones, *Kernel Smoothing*. CRC press, 1994.
- [15] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Y. Chen, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, and Hexagon-ML, "The UCR time series classification archive," October 2018, https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
- [16] D. W. Scott and S. J. Sheather, "Kernel density estimation with binned data," *Communications in Statistics-Theory and Methods*, vol. 14, no. 6, pp. 1353–1359, 1985.
- [17] P. Hall and P. Patil, "Formulae for mean integrated squared error of nonlinear wavelet-based density estimators," *The Annals of Statistics*, pp. 905–928, 1995.
- [18] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *The annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.
- [19] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.
- [20] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.
- [21] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.
- [22] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.
- [23] A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, no. 2, pp. 353–360, 1984.
- [24] P. Hall, S. J. Sheather, M. Jones, and J. S. Marron, "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, vol. 78, no. 2, pp. 263–269, 1991.
- [25] P. Robert, "On the choice of smoothing parameters for parzen estimators of probability density functions," *IEEE Transactions on Computers*, vol. 25, no. 11, pp. 1175–1179, 1976.
- [26] M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, vol. 9, pp. 65–78, 1982.
- [27] H. Shimazaki and S. Shinomoto, "Kernel bandwidth optimization in spike rate estimation," *Journal of Computational Neuroscience*, vol. 29, no. 1, pp. 171–182, 2010.
- [28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.
- [29] G. R. Terrell, "The maximal smoothing principle in density estimation," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 470–477, 1990.
- [30] J. S. Marron and M. P. Wand, "Exact mean integrated squared error," *The Annals of Statistics*, vol. 20, no. 2, pp. 712–736, 1992.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.



Yinsong Wang received his B.S. degree in Mechanical Engineering from Shandong University, China, in 2017, and his M.S. degree in Manufacturing System Engineering and Management from The Hong Kong Polytechnic University, Hong Kong, in 2018. He is currently working toward a Ph.D. degree in Industrial Engineering at Northeastern University. His research interests include machine learning, data science, and kernel methods.



Yu Ding (M'01, SM'11) received B.S. from University of Science & Technology of China (1993); M.S. from Tsinghua University, China (1996); M.S. from Penn State University (1998); received Ph.D. in Mechanical Engineering from University of Michigan (2001). He is currently the Mike and Sugar Barnes Professor of Industrial & Systems Engineering and a Professor of Electrical & Computer Engineering at Texas A&M University. His research interests are in data and quality science. Dr. Ding is the Editor-in-Chief of *IJSE Transactions* for the term of 2021–2024. Dr. Ding is a fellow of IIE, a fellow of ASME, a senior member of IEEE, and a member of INFORMS.



Shahin Shahrapour received the Ph.D. degree in Electrical and Systems Engineering, the M.A. degree in Statistics (The Wharton School), and the M.S.E. degree in Electrical Engineering, all from the University of Pennsylvania, in 2015, 2014, and 2012, respectively. He is currently an Assistant Professor in the Department of Mechanical and Industrial Engineering at Northeastern University. His research interests include machine learning, optimization, sequential decision-making, and distributed learning, with a focus on

developing computationally efficient methods for data analytics. He is a Senior Member of the IEEE.