UNIVERSAL ESTIMATION OF TIME-VARYING DISTRIBUTIONS

Kaan Gokcesu and Suleyman S. Kozat, Senior Member, IEEE

Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey,
Email: {gokcesu,kozat}@ee.bilkent.edu.tr

ABSTRACT

We investigate the estimation of distributions with time-varying parameters. We introduce an algorithm that achieves the optimal negative likelihood performance against the true probability distribution. We achieve this optimum regret performance without any knowledge about the total change of the parameters of true distribution. Our results are guaranteed to hold in an individual sequence manner such that we have no assumptions on the underlying sequences. Our log-loss performance with respect to the true probability density has regret bounds of $O(\sqrt{CT})$, where C is the total change (drift) in the natural parameters of the underlying distribution. We achieve this square-root regret with computational complexity only logarithmic in the time horizon, thus our algorithm is suitable for big data. Apart from the regret bounds, through synthetic and real life experiments, we demonstrate substantial performance gains with respect to the state-of-the-art probability density estimation algorithms in the literature.

Index Terms— Sequential density estimation, exponential family, nonstationary source, individual sequence manner.

I. INTRODUCTION

In this paper, we investigate sequential probability density estimation, which arise in several different machine learning applications [1]–[6], where we sequentially observe $\{x_1, x_2, \ldots\}$ and learn a probability distribution at each time t based on the past observations. We assume $\{x_t\}_{t\geq 1}$ are generated from a possibly nonstationary memoryless source since in most applications in engineering systems, statistics of data may change over time (especially, big data) [7].

We approach this problem from a competitive algorithm perspective and compete against the true density function. At each time t, we observe a sample feature vector x_t distributed according to some unknown $f_t(x_t)$, and based on our past observations $\{x_\tau\}_{\tau\geq 1}^{t-1}$, we produce an estimate $\hat{f}_t(x_t)$. We use negative likelihood (log-loss) as the loss, i.e., $-\log(\hat{f}_t(x_t))$, since it is the most widely used loss function for probability distributions [8]. To provide strong results in an individual sequence manner [9], we use the notion of ”regret” on the log-losses to define our performance such that the regret at time t is $r_t = -\log(\hat{f}_t(x_t)) + \log(f_t(x_t))$. and the cumulative regret up to time T is $R_T = \sum_{t=1}^{T}\Big(-\log(\hat{f}_t(x_t)) + \log(f_t(x_t))\Big)$.

We seek to achieve the performance of the best non-stationary distribution from an exponential family. In this sense, we assume that there exists a density function $f_t(x_t)$ that exactly or most closely represents the true distribution and $f_t(x_t)$ belongs to an exponential family [10] with a possibly changing natural parameter α_t. We specifically investigate the exponential family of distributions since they cover a wide range of parametric models [6] and accurately approximates many nonparametric classes of densities [11].

We denote the drift of α_t in T rounds by C_α such that

$$C_\alpha \triangleq \sum_{t=2}^{T} \|\alpha_t - \alpha_{t-1}\|, \tag{1}$$

where $\|\cdot\|$ is the L^2-norm. As an example, for stationary sources (unchanging natural parameter), $C_\alpha = 0$. Following [6] and [12], one can show that a regret bound of $O(\log(T))$ can be achieved for a stationary source with fixed computational complexity. However, for nonstationary sources, logarithmic regret bound is infeasible under low computational complexity [6]. The results of [13] imply fixed complexity learning algorithms that achieve a regret bound of $O(\sqrt{C_\alpha T})$ when the time horizon T and the total drift in parameter vector C_α is known a priori. However, if no prior knowledge about C_α is given, an algorithm with fixed complexity that achieves only the regret bound $O(C_\alpha\sqrt{T})$ is proposed in [6]. Hence, achieving $O(\sqrt{C_\alpha T})$ is not possible with the state-of-the-art methods if no prior information is known about C_α (drift) of a nonstationary source.

As the first time in literature, we introduce an algorithm that achieves the optimal regret $O(\sqrt{C_\alpha T})$ for nonstationary sources without a priori knowledge. Our results are guaranteed to hold in a strong deterministic sense for all possible observation sequences. Our algorithm is truly sequential such that neither the time horizon T nor the total drift C_α are known. We achieve this performance with a computational complexity of only log-linear in time length T.

In Section II, we first introduce the basic density estimators. Then, in Section III, we introduce the universal density estimator that merges the beliefs of the basic density estimators. In Section IV, experiments, we illustrate significant

This work was supported in part by the Turkish Academy of Sciences Outstanding Researcher Programme and the Scientific and Technological Research Council of Turkey under Contract 113E517.

performance gains with respect to the state-of-the-art and finish with final remarks in Section V.

II. Basic Density Estimator

We first construct basic density estimators that can only achieve the optimal regret bound with a priori information on the underlying sequence. This basic estimators are subsequently used in Section III to construct the final algorithm achieving the optimal regret without any a priori information. Here, at each time t, we observe $x_t \in \mathbb{R}^{d_x}$ distributed according to a memoryless exponential family distribution $f_t(x_t) = \exp\left(-\langle \alpha_t, z_t\rangle - A(\alpha_t)\right)$, where $\alpha_t \in \mathbb{R}^d$ is the natural parameter of the exponential-family distribution belonging to a bounded convex feasible set S such that $D = \max_{\alpha \in S}\|\alpha\|$, $A(\cdot)$ is a function of the parameter α_t (normalization factor), $\langle \cdot, \cdot \rangle$ is the inner product and z_t is the d-dimensional sufficient statistic of x_t [10], i.e., $z_t = \mathcal{T}(x_t)$.

Instead of directly estimating the distribution $f_t(x)$, we estimate the natural parameter α_t at each time t according to our observations $\{x_\tau\}_{\tau=1}^{t-1}$ and show that we achieve Hannan consistent [14] regret bounds. The estimate of the underlying true distribution is given by $\hat{f}_t(x_t) = \exp(-\langle \hat{\alpha}_t, z_t\rangle - A(\hat{\alpha}_t))$. We use online gradient descent [13] to sequentially produce our estimation $\hat{\alpha}_t$, where we start from an initial estimate $\hat{\alpha}_1$, and update $\hat{\alpha}_t$ based on observation x_t. To update $\hat{\alpha}_t$, we first observe x_t and incur the loss $l(\hat{\alpha}_t, x_t)$, which is the log-loss as

$$l(\hat{\alpha}_t, x_t) = -\log(\hat{f}_t(x_t)) = \langle \hat{\alpha}_t, z_t\rangle + A(\hat{\alpha}_t). \qquad (2)$$

Hence the gradient of the loss with respect to $\hat{\alpha}_t$,

$$\nabla_\alpha l(\hat{\alpha}_t, x_t) = z_t + \nabla_\alpha A(\hat{\alpha}_t) = z_t - \mu_{\hat{\alpha}_t}, \qquad (3)$$

where $\mu_{\hat{\alpha}_t}$ is the mean of z_t if x_t were distributed according to $\hat{f}_t(x_t)$. We update the parameter $\hat{\alpha}_t$ such that

$$\hat{\alpha}_{t+1} = P_S(\hat{\alpha}_t - \eta(z_t - \mu_{\hat{\alpha}_t})), \qquad (4)$$

where $P_S(\cdot)$ is the projection onto S and is defined as

$$P_S(x) = \underset{y \in S}{\arg\min}\|x - y\| \qquad (5)$$

The complete algorithm is provided in Alg. 1.

Next, we provide performance bounds of Alg. 1. Theorem 1 shows that using Alg. 1 with fixed learning rate, we can achieve optimal regret $O(\sqrt{C_\alpha T})$ if C_α is known.

Theorem 1. *Running Alg. 1 with parameter η to estimate the distribution $f_t(x_t)$ has the regret bound*

$$R_T \leq \frac{1}{\eta} DC + \eta T G, \qquad (6)$$

where $D = \max_{\alpha \in S}\|\alpha\|$, $C = 2.5D + C_\alpha$ such that C_α is as in (1)*, and $G = (\phi_2 + 2\phi_1 M + M^2)/2$ such that $M = \max_{\alpha \in S}\mu_\alpha$, $\phi_1 = \sum_{t=1}^T \|z_t\|/T$, $\phi_2 = \sum_{t=1}^T \|z_t\|^2/T$.*

Proof of Theorem 1. The regret at time t is defined as $r_t = l(\hat{\alpha}_t, x_t) - l(\alpha_t, x_t)$, where $l(\alpha, x)$ is as in (2). Since the loss function is convex

$$r_t \leq \langle \nabla_\alpha l(\hat{\alpha}_t, x_t), (\hat{\alpha}_t - \alpha_t)\rangle. \qquad (7)$$

Algorithm 1 Basic Density Estimator

```
Initialize constant $\eta \in \mathbb{R}^+$
Select initial parameter $\hat{\alpha}_1$
Calculate the mean $\mu_{\hat{\alpha}_1}$
for $t = 1$ to $T$ do
    Declare estimation $\hat{\alpha}_t$
    Observe $x_t$
    Calculate $z_t = \mathcal{T}(x_t)$
    Update parameter: $\tilde{\alpha}_{t+1} = \hat{\alpha}_t - \eta(z_t - \mu_{\hat{\alpha}_t})$
    Project onto convex set: $\hat{\alpha}_{t+1} = P_S(\tilde{\alpha}_{t+1})$
    Calculate the mean $\mu_{\hat{\alpha}_{t+1}}$
end for
```

We bound the right hand side of (7) using the update rule (4). By (4) and (5), we have

$$\langle \nabla_\alpha l(\hat{\alpha}_t, x_t), (\hat{\alpha}_t - \alpha_t)\rangle$$
$$\leq \frac{1}{2\eta}(\|\hat{\alpha}_t\|^2 - \|\hat{\alpha}_{t+1}\|^2 - 2\langle \hat{\alpha}_t - \hat{\alpha}_{t+1}, \alpha_t\rangle) + \frac{\eta}{2}\|\nabla_\alpha l(\hat{\alpha}_t, x_t)\|^2_;$$

since $\eta > 0$. Using (7) and (3) yields

$$r_t \leq \frac{1}{2\eta}(\|\hat{\alpha}_t\|^2 - \|\hat{\alpha}_{t+1}\|^2) - \frac{1}{\eta}\langle \hat{\alpha}_t - \hat{\alpha}_{t+1}, \alpha_t\rangle + \frac{\eta}{2}\|z_t - \mu_{\hat{\alpha}_t}\|^2_.$$

Thus, the cumulative regret up to time T is given by

$$\begin{aligned} R_T \leq & \frac{1}{2\eta}(\|\hat{\alpha}_1\|^2 - \|\hat{\alpha}_{T+1}\|^2) + \frac{\eta}{2}\sum_{t=1}^T \|z_t - \mu_{\hat{\alpha}_t}\|^2 \\ & - \frac{1}{\eta}(\langle \hat{\alpha}_1, \alpha_1\rangle + \sum_{t=2}^T \langle \hat{\alpha}_t, \alpha_t - \alpha_{t-1}\rangle - \langle \hat{\alpha}_{T+1}, \alpha_T\rangle), \\ \leq & \frac{1}{\eta}(2.5D^2 + DC_\alpha) + \frac{\eta T}{2}\left(\phi_2 + 2\phi_1 M + M^2\right), \end{aligned}$$

where C_α is as in (1), $D = \max_{\alpha \in S}\|\alpha\|$, $M = \max_{\alpha \in S}\mu_\alpha$, $\phi_1 = \sum_{t=1}^T \|z_t\|/T$, $\phi_2 = \sum_{t=1}^T \|z_t\|^2/T$. We denote $G = (\phi_2 + 2\phi_1 M + M^2)/2$, which is related to the gradient of the log-loss and $C = C_\alpha + 2.5D$, which is the effective change parameter. Hence, we get (6).

□

The result in Theorem 1 is for an estimator that uses fixed learning rate, which will be used to prove the performance bound of the universal estimator in the next section.

III. Universal Online Density Estimation

In Section II, we constructed basic estimators achieve the optimal regret with a priori information. In this section, we construct a universal algorithm that achieves the optimal regret with no a priori information by mixing the beliefs of basic estimators with carefully constructed learning rates.

Alg. 1, when used with η, achieves the regret

$$R_T \leq \sqrt{DCGT}\left(\frac{\eta_*}{\eta} + \frac{\eta}{\eta_*}\right), \qquad (8)$$

where $\eta_* \triangleq \sqrt{(DC)/(GT)}$. To achieve the optimal regret with Alg. 1, one must have some knowledge of η_*. However, with no prior information, it is not possible to achieve the optimal regret using Alg. 1. Therefore, instead of just using

Algorithm 2 Universal Density Estimator

```
Initialize constants $\eta_r$, for $r \in \{1, 2, \dots, N\}$
Create $N$ nodes each running Alg. 1 with parameters $\eta_r$
Initialize weights $w_1^r = 1/N$
**for** $t = 1$ **to** $T$ **do**
    Declare estimation $\hat{f}_t^u(x) = \sum_{r=1}^{N} w_t^r \hat{f}_t^r(x)$
    Observe $x_t$
    Calculate $z_t = \mathcal{T}(x_t)$
    **for** $r = 1$ **to** $N$ **do**
        Update parameters $\hat{\alpha}_t^r$ according to Alg. 1
        $w_{t+1}^r = w_t^r \hat{f}_t^r(x_t) / \hat{f}_t^u(x_t)$
    **end for**
**end for**
```

Alg. 1 with a fixed learning rate, we combine Alg. 1's with different learning rates, which will approximate η_* to a sufficient degree to achieve the optimal regret.

To this end, we first construct a parameter vector $\boldsymbol{\eta}$ of size N such that $\boldsymbol{\eta}[r] = \eta_r$, for $r \in \{1, 2, \dots, N\}$. We construct N experts each of which runs Alg. 1 with parameter η_r, i.e., r^{th} element of the parameter vector $\boldsymbol{\eta}$. Each one of the N experts takes the input x_t and outputs a belief $\hat{f}_t^r(x_t)$ at each round t (prediction stage). Then, we mix the outputs of all the beliefs in a weighted combination such that

$$\hat{f}_t^u(x_t) = \sum_{r=1}^{N} w_t^r \hat{f}_t^r(x_t), \tag{9}$$

where w_t^r is the combination weight of the belief of the r^{th} expert at time t (mixture stage). Initially we assign uniform weights to all expert outputs such that their combination weights are given by $w_1^r = 1/N$. Then, at each time t, we update their weights according to the rule

$$w_{t+1}^r = w_t^r \hat{f}_t^r(x_t) / \hat{f}_t^u(x_t), \tag{10}$$

where $\hat{f}_t^u(x_t)$ acts as a normalizer. We have provided a complete description of the universal algorithm in Alg. 2.

Next, we provide the performance bounds of the universal density estimator, i.e., Alg. 2. The results of Theorem 2 and Corollary 1 show that the optimal regret bound $O(\sqrt{CT})$ is achieved without any prior information on C.

Theorem 2. *Alg. 2 has the regret bound*

$$R_T \le \log(N) + \sqrt{DCGT}\left[\min_{i \in \{1,2,\dots,N\}} \left(\frac{\eta_*}{\eta_i} + \frac{\eta_i}{\eta_*}\right)\right],$$

where $D = \max_{\alpha \in S} \|\alpha\|$, $C = 2.5D + C_\alpha$ such that C_α is defined as in (1)*, $G = (\phi_2 + 2\phi_1 M + M^2)/2$ such that $M = \max_{\alpha \in S} \mu_\alpha$, $\phi_1 = \sum_{t=1}^{T} \|z_t\|/T$, $\phi_2 = \sum_{t=1}^{T} \|z_t\|^2/T$, $\eta_* = \sqrt{DCGT}$ and η_i for $i \in \{1, 2, \dots, N\}$ are the parameters used by the mixed experts.*

Proof of Theorem 2. The regret at time t is given by

$$r_t = -\log(\hat{f}_t^u(x_t)) + \log(f_t(x_t)) \tag{11}$$

Summing (11) from $t = 1$ to T gives

$$R_T = -\log\Big(\prod_{t=1}^{T}\Big(\sum_{r=1}^{N} w_t^r \hat{f}_t^r(x_t)\Big)\Big) + \sum_{t=1}^{T} \log(f_t(x_t)), \tag{12}$$

where we used (9). From (10), we can infer

$$w_t^r = \frac{\prod_{\tau=1}^{t-1} \hat{f}_\tau^r(x_\tau)}{\sum_{r=1}^{N} \prod_{\tau=1}^{t-1} \hat{f}_\tau^r(x_\tau)}. \tag{13}$$

Hence, putting (13) in (12) produces,

$$\begin{aligned} R_T &= -\log\Big(\sum_{r=1}^{N} \prod_{\tau=1}^{T} \hat{f}_t^r(x_t)\Big) + \log(N) + \sum_{t=1}^{T} \log(f_t(x_t)) \\ &\le \log(N) - \max_r\Big(\sum_{t=1}^{T} \log(\hat{f}_t^r(x_t))\Big) + \sum_{t=1}^{T} \log(f_t(x_t)) \quad (14) \\ &\le \log(N) + \sqrt{DCGT}\left[\min_{i \in \{1,2,\dots,N\}} \left(\frac{\eta_*}{\eta_i} + \frac{\eta_i}{\eta_*}\right)\right], \quad (15) \end{aligned}$$

and concludes the proof.

□

The result of Theorem 2 shows that the bound is dependent on the set of learning rates used in the algorithm. In Corollary 1, we exploit that result and show that we can achieve the optimal regret bound with log-linear complexity.

Corollary 1. *Suppose we choose to run the experts with parameters between η' and η''. We denote $K = \eta''/\eta'$ and $N = \lceil \log_2 K \rceil + 1$. Then running Alg. 2 with parameter vector $\eta_i = 2^{i-1}\eta'$ for $i \in \{1, 2, \dots, N\}$ gives the following regret bounds for different values of η_*.*

1) *If $\eta' \le \eta_* \le \eta''$:*

$$R_T \le \log(\lceil \log_2 \eta''/\eta' \rceil + 1) + \frac{3\sqrt{2}}{2}\sqrt{DCGT},$$

since $(\eta_*/\eta_i + \eta_i/\eta_*)$ is the maximum if η_* is of the form $\eta_* = 2^a\sqrt{2}$ for some a.

2) *If $\eta_* \ge \eta''$*

$$R_T \le \log(\lceil \log_2 \eta''/\eta' \rceil + 1) + (1 + \frac{\eta_*}{\eta''})\sqrt{DCGT}.$$

Since $\eta_* \le \sqrt{(4 + 1/T)D^2M^{-2}}$, setting $\eta'' \ge \sqrt{(4 + T^{-1})D^2M^{-2}}$ yields this case invalid.

3) *If $\eta_* \le \eta'$*

$$R_T \le \log(\lceil \log_2 \eta''/\eta' \rceil + 1) + (1 + \frac{\eta'}{\eta_*})\sqrt{DCGT}.$$

Since $\eta_* \ge \sqrt{2.5D^2(TG)^{-1}}$, $\eta' \le \sqrt{2.5D^2T^{-1}}$ gives

$$R_T \le \log(\lceil \log_2 \eta''/\eta' \rceil + 1) + (1 + \sqrt{G})\sqrt{DCGT}.$$

Hence, by running the Alg. 2 with an appropriate parameter vector, we achieve $O(\sqrt{CT})$ regret with $O(\log T)$ computational complexity, since the separation between η' and η'' is mainly dependent on C, which is bounded as $2.5D \le C \le (2T + 0.5)D$. For unknown time horizon T, we use the doubling trick and still achieve $O(\sqrt{CT})$ regret.

IV. Experiments

In this section, we demonstrate the performance of our algorithm both on real and synthesized data, where we show how it performs individually and in comparison to the-state-of-art [6], [13]. We denote the technique used in [13] by OCP.static, since the algorithm uses online convex programming with fixed step size. We denote the algorithm of [6]

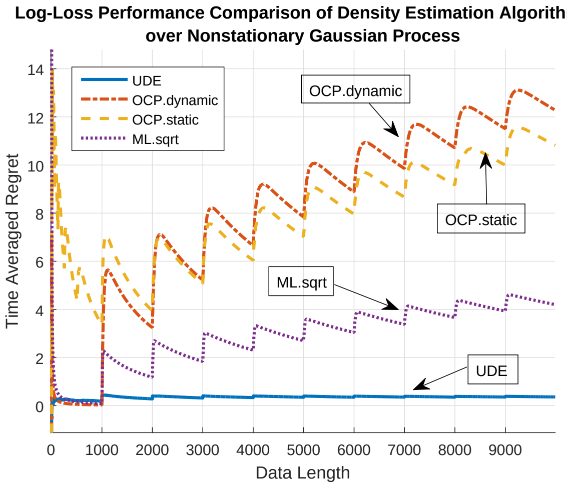

Fig. 1. Time averaged regret performance of the density estimation algorithms over nonstationary Gaussian process.

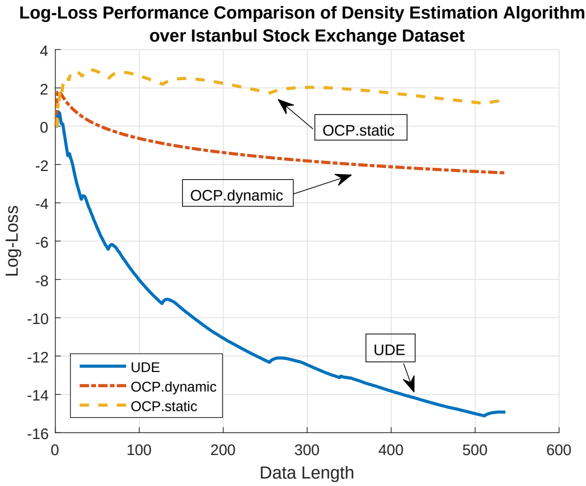

Fig. 2. Time averaged log-loss performance of the density estimation algorithms over the Istanbul Stock Exchange dataset.

as OCP.dynamic, since the algorithm uses a dynamically changing step size in each round. We also compare our algorithm to the online version of the widely used Maximum Likelihood (ML) estimation [15]. We run our algorithm with parameters in the range $1/T \leq \eta \leq T$ for a T length run. All the algorithms are constructed as in their publications.

We first synthesize a dataset of size 10000 from a univariate Gaussian process with unit standard deviation, i.e., $\sigma = 1$ and mean value alternating between 10 and -10 in every 1000 samples. Since the dataset is nonstationary, the ML estimator performs very poorly. Therefore, for a fair performance comparison, we created a new ML algorithm called ML.sqrt, which uses only the last $\sqrt{t}$ samples in each round t. In Fig. 1, we illustrate the regret performance of these four algorithms. While OCP.dynamic converges extremely fast in the first 1000 samples, OCP.static is not able to converge that rapidly in the first segment since it resets its algorithm in each power of 2 (courtesy of the doubling trick) and the learning rate is not sufficiently large. However, after approximately round 2000, OCP.static beats OCP.dynamic, and performs continuously better in the remaining rounds. Nevertheless, both of them are inferior to ML.sqrt, which has better performance in all rounds. However, all of these algorithms are still very sensitive to the changes in the data statistics and the average regret jumps upward in each change, i.e., in every 1000 rounds. Moreover, the average regrets of all three algorithms has a quasi-linear increase with respect to the data length. However, our algorithm Universal Density Estimator (UDE) has no such problem. UDE uses a mixture of carefully constructed density estimators and is robust due to its universality over learning rates. Hence, the changes in the statistics has little to no effect on the regret of UDE. UDE uniformly and significantly outperforms all three algorithms OCP.dynamic, OCP.static and ML.sqrt.

We use the Istanbul Stock Exchange (ISE) [16] dataset for real data benchmark purposes, which is readily accessible online. We have assumed a possibly nonstationary multivariate Gaussian process for ISE dataset and run the algorithms to estimate this distribution. Since the true distribution is not known, we have compared the performances of the algorithms directly with their log-losses instead of their regrets. In Fig. 2, we have illustrated the log-loss performance of UDE, OCP.static and OCP.dynamic. We needed to discard the ML and ML.sqrt algorithms since both of these algorithms performed very poorly. Since the dataset is small (with only 536 time instances), OCP.static performs poorly because of its resetting behavior. Log-loss of OCP.static cannot convergence fast enough to keep up with the nonstationarity of data, and thus, is unable to produce a successful density estimation. On the other hand, OCP.dynamic has a nice convergence through the data sequence, however, starts to floor after the 300^{th} sample. Nonetheless, UDE greatly outperforms them continuously, since it uses a mixture of carefully constructed learning rates. Therefore, UDE has a fast convergence.

V. Concluding Remarks

We have introduced a truly sequential algorithm, which estimates the density of a nonstationary exponential family source with optimum regret $\sqrt{CT}$ without knowing C (the drift of statistics) beforehand. Our results are guaranteed to hold in a strong deterministic sense for all possible observation sequences. By carefully designing different density estimators and universally combining them, we achieve optimal performance with a computational complexity only log-linear in the time horizon, thus, our algorithm can be effectively used in applications involving big data.

VI. References

[1] Y. Nakamura and O. Hasegawa, "Nonparametric density estimation based on self-organizing incremental neural network for large noisy data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–10, 2016.

[2] A. Penalver and F. Escolano, "Entropy-based incremental variational bayes learning of gaussian mixtures," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 534–540, March 2012.

[3] X. Ding, Y. Li, A. Belatreche, and L. P. Maguire, "Novelty detection using level set methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 576–588, March 2015.

[4] E. Mller, I. Assent, R. Krieger, S. Gnnemann, and T. Seidl, "Densest: Density estimation for data mining in high dimensional spaces," in *Proc. SIAM International Conference on Data Mining (SDM 2009), Sparks, Nevada, USA.* SIAM, 2009, pp. 173–184.

[5] Y. Cao, H. He, and H. Man, "Somke: Kernel density estimation over data streams by sequences of self-organizing maps," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1254–1268, Aug 2012.

[6] M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia, "Sequential anomaly detection in the presence of noise and limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5544–5562, Aug 2012.

[7] K. B. Dyer, R. Capo, and R. Polikar, "Compose: A semisupervised learning framework for initially labeled nonstationary streaming data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 12–26, Jan 2014.

[8] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[9] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *J. ACM*, vol. 44, no. 3, pp. 427–485, May 1997.

[10] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical Society*, vol. 39, no. 3, pp. 399–409, 1936.

[11] A. R. Barron and C.-H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, no. 3, pp. 1347–1369, 09 1991.

[12] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2, pp. 169–192, 2007.

[13] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent." in *ICML*, T. Fawcett and N. Mishra, Eds. AAAI Press, 2003, pp. 928–936.

[14] S. Hart and A. Mas-Colell, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, no. 1, pp. 26 – 54, 2001.

[15] I. J. Myung, "Tutorial on maximum likelihood estimation," *J. Math. Psychol.*, vol. 47, no. 1, pp. 90–100, Feb. 2003.

[16] O. Akbilgic, H. Bozdogan, and M. E. Balaban, "A novel hybrid rbf neural networks model as a forecaster," *Statistics and Computing*, vol. 24, no. 3, pp. 365–375, 2014.