Sequential Anomaly Detection in the Presence of Noise and Limited Feedback

Maxim Raginsky, *Member, IEEE*, Rebecca M. Willett, *Senior Member, IEEE*, Corinne Horn, Jorge Silva, *Member, IEEE*, and Roummel F. Marcia, *Member, IEEE*

***Abstract*—This paper describes a methodology for detecting anomalies from sequentially observed and potentially noisy data. The proposed approach consists of two main elements: 1) *filtering*, or assigning a belief or likelihood to each successive measurement based upon our ability to predict it from previous noisy observations and 2) *hedging*, or flagging potential anomalies by comparing the current belief against a time-varying and data-adaptive threshold. The threshold is adjusted based on the available feedback from an end user. Our algorithms, which combine universal prediction with recent work on online convex programming, do not require computing posterior distributions given all current observations and involve simple primal-dual parameter updates. At the heart of the proposed approach lie exponential-family models which can be used in a wide variety of contexts and applications, and which yield methods that achieve sublinear per-round regret against both static and slowly varying product distributions with marginals drawn from the same exponential family. Moreover, the regret against static distributions coincides with the minimax value of the corresponding online strongly convex game. We also prove bounds on the number of mistakes made during the hedging step relative to the best offline choice of the threshold with access to all estimated beliefs and feedback signals. We validate the theory on synthetic data drawn from a time-varying distribution over binary vectors of high dimensionality, as well as on the Enron email dataset.**

***Index Terms*—Anomaly detection, exponential families, filtering, individual sequences, label-efficient prediction, minimax regret, online convex programming (OCP), prediction with limited feedback, sequential probability assignment, universal prediction.**

Manuscript received July 14, 2010; revised February 05, 2012; accepted February 23, 2012. Date of publication May 30, 2012; date of current version July 10, 2012. This work was supported in part by the NSF CAREER Award CCF-06-43947, in part by the NSF Award DMS-08-11062, in part by the DARPA Grant HR0011-07-1-003, and in part by the ARO Grant W911NF-09-1-0262. The material in this paper was presented in part at the 2009 IEEE International Symposium on Information Theory.

M. Raginsky was with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA. He is now with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign IL 61801 USA (e-mail: maxim@illinois.edu).

R. M. Willett and J. Silva are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: willett@duke.edu; jg.silva@duke.edu).

C. Horn was with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA. She is now with the Department of Electrical and Computer Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: cehorn@stanford.edu).

R. F. Marcia is with the School of Natural Sciences, University of California, Merced, CA 95343 USA (e-mail: rmarcia@ucmerced.edu).

Communicated by S. Tatikonda, Associate Editor for Communications.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2201375

I. Introduction

In this paper, we explore the performance of online anomaly detection methods built on sequential probability assignment and dynamic thresholding based on limited feedback. We assume that we sequentially monitor the state of some system of interest. At each time step, we observe a possibly *noise-corrupted* version z_t of the current state x_t, and need to infer whether x_t is *anomalous* relative to the actual sequence $x^{t-1} = (x_1, \ldots, x_{t-1})$ of the past states. This inference is encapsulated in a binary decision $\widehat{y}_t$, which can be either -1 (nonanomalous or nominal behavior) or $+1$ (anomalous behavior). After announcing our decision, we may occasionally receive *feedback* on the "true" state of affairs and use it to adjust the future behavior of the decision-making mechanism.

Our inference engine should make good use of this feedback, whenever it is available, to improve its future performance. One reasonable way to do it is as follows. Having observed z^{t-1} (but not z_t), we can use this observation to assign "beliefs" or "likelihoods" to the clean state x_t. Let us denote this likelihood assignment as $p_t(x_t|z^{t-1})$. Then, if we actually had access to the clean observation x_t, we could evaluate $p_t = p_t(x_t|z^{t-1})$ and declare an anomaly ($\widehat{y}_t = +1$) if $p_t < \tau_t$, where τ_t is some positive threshold; otherwise, we would set $\widehat{y}_t = -1$ (no anomaly at time t). This approach is based on the intuitive idea that a new observation x_t should be declared anomalous if it is very unlikely based on our past knowledge (namely, z^{t-1}). In other words, observations are considered anomalous if they are in a portion of the observation domain which has very low likelihood according to the best probability model that can be assigned to them on the basis of previously seen observations. (In fact, anomaly detection algorithms based on density level sets revolve around precisely this kind of reasoning.) The complication here, however, is that we do not actually observe x_t, but rather its noise-corrupted version z_t. Thus, we settle instead for an *estimate* $\widehat{p}_t$ of p_t based on z_t and compare this estimate against τ_t. If we receive feedback y_t at time t and it differs from our label $\widehat{y}_t$, then we adjust the threshold appropriately.

A. Contributions

There are several challenging aspects inherent in the problem of sequential anomaly detection.

1) The observations cannot be assumed to be independent, identically distributed, or even come from a realization of a stochastic process. In particular, an adversary may be injecting false data into the sequence of observations to cripple our anomaly detection system.

2) Observations may be contaminated by noise or be observed through an imperfect communication channel.
3) Declaring observations anomalous if their likelihoods fall below some threshold is a popular and effective strategy for anomaly detection, but setting this threshold is a notoriously difficult problem.
4) Obtaining feedback on the quality of automated anomaly detection is costly as it generally involves considerable effort by a human expert or analyst. Thus, if we have an option to request such feedback at any time step, we should exercise this option sparingly and keep the number of requests to a minimum. Alternatively, the times when we receive feedback may be completely arbitrary and not under our control at all—for instance, we may receive feedback only when we declare false positives or miss true anomalies.

In this paper, we propose a general methodology for addressing these challenges. With apologies to H. P. Lovecraft [1], we will call our proposed framework FHTAGN, or *F*iltering and *H*edging for *T*ime-varying *A*nomaly reco*GN*ition. More specifically, the two components that make up FHTAGN are as follows.

1) *Filtering*—the sequential process of updating *beliefs* on the next state of the system based on the noisy observed past. The term "filtering" comes from statistical signal processing [2] and is intended to signify the fact that the beliefs of interest concern the unobservable *actual* system state, yet can only be computed in a *causal manner* from its noise-corrupted observations.
2) *Hedging*—the sequential process of flagging potential anomalies by comparing the current belief against a time-varying threshold. The rationale for this approach comes from the intuition that a behavior we could not have predicted well based on the past is likely to be anomalous. The term "hedging" is meant to indicate the fact that the threshold is dynamically raised or lowered, depending on the type of the most recent mistake (a false positive or a missed anomaly) made by our inference engine.

Rather than explicitly modeling the evolution of the system state and then designing methods for that model (e.g., using Bayesian updates [2], [3]), we adopt an "individual sequence" (or "universal prediction" [4]) perspective and strive to perform provably well on any individual observation sequence in the sense that our per-round performance approaches that of the best *offline* method with access to the entire data sequence. This approach allows us to sidestep challenging statistical issues associated with dependent observations or dynamic and evolving probability distributions, and is robust to noisy observations. We make the following contributions.

1) We cast both filtering and hedging as instances of online convex programming (or OCP), as defined by Zinkevich [5]. This will permit us to implement both of these ingredients of FHTAGN using a powerful primal-dual method of mirror descent (MR)[6], [7] and quantify their performance in a unified manner via regret bounds relative to the best offline strategy with access to the full observation sequence.
2) We show that the filtering step can be implemented as a sequential assignment of beliefs, or probabilities, to the system state based on the past noisy observations, where the probabilities are computed according to an exponential family model whose natural parameter is dynamically determined based on the past. We present a strategy based on MD for sequentially assigning a time-varying product distribution with exponential-family marginals to the observed noisy sequence of system states and prove regret bounds relative to the best i.i.d. model as well as to the best sufficiently slowly changing model that can be assigned to the observation sequence in hindsight. These regret bounds improve and extend our preliminary results [8]; in addition to tightening the bounds presented in that work, we extend the results to more general settings in which data may be corrupted by noise. The main thing to keep in mind about the individual-sequence setting is that neither the sequence of probability assignments nor the best model that can be chosen offline should be interpreted as estimates of some "true" stochastic process model of the observation sequence. Rather, both should be viewed as *algorithmic strategies* for predicting the next observation given the past (cf., the survey paper by Merhav and Feder [4] for more details on the differences and the similarities between the more familiar probabilistic setting and the deterministic, individual-sequence setting used in this paper).
3) We show that the hedging step can be implemented as a sequential selection of the critical threshold, such that whenever the estimated belief for the current state falls below this threshold, we declare an anomaly. We develop methods to incorporate available feedback and establish regret-type bounds on the number of mistakes relative to the best threshold that can be selected in hindsight with access to the entire sequence of assigned beliefs and feedback.

As described in the useful survey by Chandola *et al.* [9], several methods for anomaly detection have been developed using supervised [10], semisupervised [11], and unsupervised [12] learning methods. In the online, individual-sequence setting we adopt, however, there is no *intrinsic* notion of what constitutes an anomaly. Instead, we focus on *extrinsic* anomalous behavior relative to the best *model* we can guess for the next observation based on what we have seen in the past. We are not aware of any anomaly detection performance bounds in nonstationary or adversarial settings prior to our work.

B. Notation

We will follow the following notational conventions. "Basic" sets will be denoted by sans-serif uppercase letters, e.g., $\mathsf{U}, \mathsf{X}, \mathsf{Z}$, while classes of sets and functions will be denoted by script letters, e.g., $\mathcal{C}, \mathcal{F}, \mathcal{G}$. Given a set X, we will denote by X^k the k-fold Cartesian product of X with itself and by x^k a representative k-tuple from X^k. The set of all (one-sided) infinite sequences over X will be denoted by X^∞, and a representative element will be written in boldface as $\boldsymbol{x} = (x_1, x_2, \ldots)$. The interior of a set U will be denoted by $\mathrm{Int}\,\mathsf{U}$. The standard Euclidean inner product between two vectors $u, v \in \mathbb{R}^m$ will be denoted by $\langle u, v \rangle$.

II. Preliminaries

This section is devoted to setting up the basic terminology and machinery to be used throughout this paper. This includes

background information on OCP (see Section II-A) and on exponential families (see Section II-C).

A. Online Convex Programming

The philosophy advocated in this paper is that the tasks of sequential probability assignment and threshold selection can both be viewed as a *game* between two opponents, the Forecaster and the Environment. The Forecaster is continually predicting changes in a dynamic Environment, where the effect of the Environment is represented by an arbitrarily varying sequence of convex cost functions over a given feasible set, and the goal of the Forecaster is to pick the next feasible point in such a way as to keep the cumulative cost as low as possible. This is broadly formulated as the problem of online convex programming, or OCP [5], [13], [14]. An OCP problem with horizon T is specified by a convex feasible set $\mathsf{U} \subseteq \mathbb{R}^d$ and a family of convex functions $\mathcal{F} = \{f : \mathsf{U} \to \mathbb{R}\}$, and is described as follows.

Algorithm 1 An abstract Online Convex Programming problem

The Forecaster picks an arbitrary initial point $\widehat{u}_1 \in \mathsf{U}$

for $t = 1, 2, \ldots, T$ **do**

The Environment picks a convex function $f_t \in \mathcal{F}$.
The Forecaster observes f_t and incurs the cost $f_t(\widehat{u}_t)$
The Forecaster picks a new point $\widehat{u}_{t+1} \in \mathsf{U}$

end for

The total cost incurred by the Forecaster after T rounds is given by $\sum_{t=1}^T f_t(\widehat{u}_t)$ (here and in the sequel, hats denote quantities selected by the Forecaster on the basis of past observations). At each time t, the Forecaster's move $\widehat{u}_t$ must satisfy a causality constraint in that it may depend only on his past moves $\widehat{u}^{t-1}$ and on the past functions f^{t-1} selected by the Environment. Thus, the behavior of the Forecaster may be described by a sequence of functions

$$\mu_t : \mathsf{U}^{t-1} \times \mathcal{F}^{t-1} \to \mathsf{U}, \qquad t = 1, 2, \ldots, T$$

so that $\widehat{u}_t = \mu_t(\widehat{u}^{t-1}, f^{t-1})$. We will refer to any such sequence $\mu^T = \{\mu_t : \mathsf{U}^{t-1} \times \mathcal{F}^{t-1} \to \mathsf{U}\}_{t=1}^T$ as a T-step *strategy* of the Forecaster. Informally, the goal of the Forecaster is to do almost as well as if he could observe the cost functions $f_1, \ldots, f_T$ all at once. For instance, we might want to minimize the difference between the actual cost incurred after T rounds of the game and the smallest cumulative cost that could be achieved *in hindsight* using a single feasible point. To that end, given a strategy μ^T and a cost function tuple f^T, let us define the *regret* w.r.t. a time-varying tuple $u^T = (u_1, \ldots, u_T) \in \mathsf{U}^T$

$$\begin{aligned} R_T(\mu^T; f^T, u^T) &\stackrel{\triangle}{=} \sum_{t=1}^T f_t(\widehat{u}_t) - \sum_{t=1}^T f_t(u_t) \\ &= \sum_{t=1}^T f_t(\mu_t(\widehat{u}^{t-1}, f^{t-1})) - \sum_{t=1}^T f_t(u_t). \end{aligned}$$

Then the goal would be to select a suitable restricted subset $\mathcal{C}_T \subset \mathsf{U}^T$ and design μ^T to ensure that the worst-case regret

$$\begin{aligned} &\sup_{f^T \in \mathcal{F}^T} \sup_{u^T \in \mathcal{C}_T} R_T(\mu^T; f^T, u^T) \\ &\qquad \equiv \sup_{f^T \in \mathcal{F}^T} \left\{ \sum_{t=1}^T f_t(\widehat{u}_t) - \inf_{u^T \in \mathcal{C}_T} \sum_{t=1}^T f_t(u_t) \right\} \end{aligned}$$

is sublinear in T. (When $u_t = u$ for all t and for some $u \in \mathsf{U}$, we will write the regret as $R_T(\mu^T; f^T, u)$, and the second sup in the worst-case regret would be over all $u \in \mathsf{U}$.)

Note that it is often convenient to think of a comparison tuple $u^T \in \mathcal{C}_T$ as a strategy of the form $\mu_t : \mathsf{U}^{t-1} \times \mathcal{F}^{t-1} \to \{u_t\}$, i.e., u_t does not depend on the previous points or cost functions, but only on the time index t. This allows us to speak of comparison classes of *strategies*; however, to avoid confusion, we will always use the notation μ^T (possibly with subscripts) to distinguish the Forecaster's observation-driven strategy from a comparison strategy u^T, which may be time varying but is always observation-independent. This is similar to the distinction made in control theory between *closed-loop* (or *feedback*) policies and *open-loop* policies [15]: a closed-loop policy is a sequence of *functions* for selecting the next control signal based on the past control signals and past observations, while an open-loop policy is a sequence of control *signals* fixed in advance. From this point of view, the regret pertains to the difference in cumulative costs between a feedback policy (μ^T) and the best open-loop policy in some reference class $\mathcal{C}$.

Remark 1 (Hannan Consistency): More generally, we can consider unbounded-horizon strategies $\mu = \{\mu_t : \mathsf{U}^{t-1} \times \mathcal{F}^{t-1} \to \mathsf{U}\}$. Then, given a *comparison class* $\mathcal{C} \subset \mathsf{U}^\infty$ of open-loop strategies, the design goal is to ensure that

$$R_T(\mu; \mathcal{C}) \stackrel{\triangle}{=} \sup_{\boldsymbol{f} \in \mathcal{F}^\infty} \sup_{\boldsymbol{u} \in \mathcal{C}} R_T(\mu^T; f^T, u^T) = o(T). \qquad (1)$$

Any strategy μ that achieves (1) over a comparison class $\mathcal{C}$ is said to be *Hannan-consistent* w.r.t. $\mathcal{F}$ and $\mathcal{C}$; see the text in [16] for a thorough discussion. One important comparison class is composed of all *static* (or *constant*) sequences, i.e., all $\boldsymbol{u} \in \mathsf{U}^\infty$ such that $u_1 = u_2 = \ldots$. This class, which we will denote by $\mathcal{C}_{\mathrm{stat}}$, is in one-to-one correspondence with the feasible set U, so

$$R_T(\mu; \mathcal{C}_{\mathrm{stat}}) = \sup_{\boldsymbol{f} \in \mathcal{F}^\infty} \sup_{u \in \mathsf{U}} R_T(\mu^T; f^T, u)$$

and we will also denote this worst-case regret by $R_T(\mu; \mathsf{U})$.

B. MD Procedure

A generic procedure for constructing OCP strategies is inspired by the so-called method of MD [6], [7], [17]. In the context of OCP, the rough idea behind MD is as follows. At time t the Forecaster chooses the point

$$\widehat{u}_{t+1} = \arg\min_{u \in \mathsf{U}} \left[\eta_t \langle g_t(\widehat{u}_t), u \rangle + D(u, \widehat{u}_t) \right] \qquad (2)$$

where $g_t(\widehat{u}_t)$ is an arbitrary subgradient[1] of f_t at $\widehat{u}_t$, $D(\cdot,\cdot) \geq 0$ is some measure of proximity between points in U, and $\eta_t > 0$ is a (possibly time-dependent) regularization parameter. The intuition behind (2) is to balance the tendency to stay close to the previous point against the tendency to move in the direction of the greatest local decrease of the cost. The key feature of MD methods is that they can be flexibly adjusted to the geometry of the feasible set U through judicious choice of the proximity measure $D(\cdot,\cdot)$. In particular, when U is the canonical parameter space of an exponential family, a good proximity measure is the Kullback–Leibler divergence. The general measures of proximity used in MD are given by the so-called *Bregman divergences* [19], [20]. Following [16], we introduce them through the notion of a *Legendre function*:

Definition 1: Let $\mathsf{U} \subseteq \mathbb{R}^d$ be a nonempty set with convex interior. A function $F : \mathsf{U} \to \mathbb{R}$ is called *Legendre* if it is:

1) strictly convex and continuously differentiable throughout $\operatorname{Int}\mathsf{U}$;
2) *steep* (or *essentially smooth*) — that is, if $u_1, u_2, \ldots \in \operatorname{Int}\mathsf{U}$ is a sequence of points converging to a point on the boundary of U, then $\|\nabla F(u_i)\| \to \infty$ as $i \to \infty$, where $\|\cdot\|$ denotes any norm.[2]

The *Bregman divergence* induced by F is the nonnegative function $D_F : \mathsf{U} \times \operatorname{Int}\mathsf{U} \to \mathbb{R}$, given by

$$D_F(u,v) \stackrel{\triangle}{=} F(u) - F(v) - \langle \nabla F(v), u - v\rangle, \quad \forall u \in \mathsf{U}, v \in \operatorname{Int}\mathsf{U}.$$

For example, if $\mathsf{U} = \mathbb{R}^d$, then $F(u) = (1/2)\|u\|^2$, where $\|\cdot\|$ is the Euclidean norm, is Legendre, and $D_F(u,v) = (1/2)\|u - v\|^2$. In general, for a fixed $v \in \operatorname{Int}\mathsf{U}$, $D_F(\cdot, v)$ gives the tail of the first-order Taylor expansion of $F(\cdot)$ around v.

We now present the general MD scheme for OCP, where we also allow the possibility of restricting the feasible points to a closed, convex subset S of $\operatorname{Int}\mathsf{U}$.

Algorithm 2 A Generic Mirror Descent Strategy for OCP

Require: A Legendre function $F : \mathsf{U} \to \mathbb{R}$; a decreasing sequence of strictly positive *step sizes* $\{\eta_t\}$

The Forecaster picks an arbitrary initial point $\widehat{u}_1 \in \mathsf{S}$

for $t = 1, 2, \ldots$ **do**

Observe the cost function $f_t \in \mathcal{F}$
Compute a subgradient $g_t(\widehat{u}_t)$ at $\widehat{u}_t$
Output

$$\widehat{u}_{t+1} = \arg\min_{u \in \mathsf{S}} \left[\eta_t \langle g_t(\widehat{u}_t), u\rangle + D_F(u, \widehat{u}_t)\right]$$

end for

[1]A subgradient of a convex function $f : \mathsf{U} \to \mathbb{R}$ at a point $u \in \operatorname{Int}\mathsf{U}$ is any vector $g \in \mathbb{R}^d$, such that

$$f(v) \geq f(u) + \langle g, v - u\rangle$$

holds for all $v \in \operatorname{Int}\mathsf{U}$[18].

[2]Since all norms on finite-dimensional spaces are equivalent, it suffices to establish essential smoothness in a particular norm, say the usual ℓ_2 norm.

In the case when $\mathsf{U} = \mathbb{R}^d$ and $F(\cdot) = (1/2)\|\cdot\|^2$, the aforementioned algorithm reduces to the standard projected subgradient scheme

$$\widetilde{u}_{t+1} = \widehat{u}_t - \eta_t g_t(\widehat{u}_t)$$
$$\widehat{u}_{t+1} = \arg\min_{u \in \mathsf{S}} \|u - \widetilde{u}_{t+1}\|.$$

The name "mirror descent" comes from the following equivalent form of Algorithm 2. Consider the *Legendre–Fenchel dual* of F[18], [21]:

$$F^*(z) \stackrel{\triangle}{=} \sup_{u \in \mathsf{U}} \{\langle u, z\rangle - F(u)\}.$$

Let U^* denote the image of $\operatorname{Int}\mathsf{U}$ under the gradient mapping ∇F: $\mathsf{U}^* = \nabla F(\operatorname{Int}\mathsf{U})$. An important fact is that the gradient mappings ∇F and ∇F^* are inverses of one another [7], [16], [17]:

$$\left.\begin{aligned}\nabla F^*(\nabla F(u)) &= u\\ \nabla F(\nabla F^*(w)) &= w\end{aligned}\right\} \qquad \forall u \in \operatorname{Int}\mathsf{U}, w \in \operatorname{Int}\mathsf{U}^*.$$

Following [16], we may refer to the points in $\operatorname{Int}\mathsf{U}$ as the *primal points* and to their images under ∇F as the *dual points*. Then, for each t, the computation of $\widehat{u}_{t+1}$ in Algorithm 2 can be implemented as follows:

1) compute $\xi_t = \nabla F(\widehat{u}_t)$;
2) perform dual update $\widetilde{\xi}_{t+1} = \xi_t - \eta_t g_t(\widehat{u}_t)$;
3) compute $\widetilde{u}_{t+1} = \nabla F^*(\widetilde{\xi}_{t+1})$;
4) perform projected primal update:

$$\widehat{u}_{t+1} = \arg\min_{u \in \mathsf{S}} D_F(u, \widetilde{u}_{t+1}).$$

The name "mirror descent" reflects the fact that, at each iteration, the current point in the primal space is mapped to its "mirror image" in the dual space; this is followed by a step in the direction of the negative subgradient, and then the new dual point is mapped back to the primal space. In the context of MD schemes, the Legendre function F is referred to as the *potential function*.

The following lemma (see, e.g., [17, Lemma 2.1]) is a key ingredient in bounding the regret of the MD strategy.

Lemma 1: Fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$ and suppose that, on the set S, the Legendre potential F is *strongly convex* w.r.t. $\|\cdot\|$ with parameter $\alpha > 0$, i.e., for any $u, u' \in \operatorname{Int}\mathsf{S}$,

$$F(u') \geq F(u) + \langle \nabla F(u), u' - u\rangle + \frac{\alpha}{2}\|u - u'\|^2. \quad (3)$$

Then for any $u \in \mathsf{S}$ and any t, we have the bound

$$D_F(\widehat{u}_{t+1}, u) \leq D_F(\widehat{u}_t, u) + \eta_t \langle g_t(\widehat{u}_t), u - \widehat{u}_t\rangle + \frac{\eta_t^2}{2\alpha}\|g_t(\widehat{u}_t)\|_*^2 \quad (4)$$

where $\|v\|_* \stackrel{\triangle}{=} \sup\{\langle u, v\rangle : \|u\| \leq 1\}$ is the norm *dual* to $\|\cdot\|$.

Remark 2: The Euclidean norm $\|u\| = (u_1^2 + \cdots + u_d^2)^{1/2}$ is dual to itself, $\|\cdot\|_* = \|\cdot\|$.

The possibility of attaining sublinear regret using MD hinges on the availability of a suitable strongly convex Legendre potential. Typically, the choice of the potential function is influenced

by the geometry of the underlying set U; the reader is invited to consult the papers by Beck and Teboulle [7] and by Nemirovski *et al.* [17] for many examples.

Lemma 1 is central to our proofs of all the regret bounds presented in the sequel. Thus, even though the overall flavor of the proofs is similar to what can be found in the OCP literature [5], [13], [14], [16], we feel that the use of Lemma 1 (instead of the more usual arguments exploiting the primal-dual form of MD and the "three-point formula" for Bregman divergences [16, Lemma 11.1]) leads to simpler and shorter arguments and allows us to seamlessly tie together many different settings (e.g., both static and time-varying comparison classes). The reader may also wish to consult [22], where Lemma 1 is used to analyze adaptive closed-loop control schemes based on MD with data-driven selection of the step size.

C. Background on Exponential Families

This section is devoted to a brief summary of the basics of exponential families; Amari and Nagaoka [23] or Wainwright and Jordan [24] give more details.

We assume that the observation space X is equipped with a σ-algebra $\mathcal{B}$ and a σ-finite measure ν on $(\mathsf{X}, \mathcal{B})$. Given a positive integer d, let $\phi : \mathsf{X} \to \mathbb{R}^d$ be a measurable function, and let ϕ_k, $k = 1, 2, \ldots, d$, denote its components:

$$\phi(x) = \big(\phi_1(x), \ldots, \phi_d(x)\big)^T.$$

Let Θ be the set of all $\theta \in \mathbb{R}^d$ such that

$$\int_{\mathsf{X}} \exp\big\{\langle \theta, \phi(x)\rangle\big\} d\nu(x) < +\infty.$$

We then have the following definition.

Definition 2: The set $\mathcal{P}(\phi)$ of probability distributions on $(\mathsf{X}, \mathcal{B})$ parametrized by $\theta \in \Theta$ such that the probability density function of each $P_\theta \in \mathcal{P}(\phi)$ w.r.t. the measure ν can be expressed as

$$p_\theta(x) = \exp\big\{\langle \theta, \phi(x)\rangle - \Phi(\theta)\big\}$$

where

$$\Phi(\theta) \stackrel{\triangle}{=} \log \int_{\mathsf{X}} \exp\big\{\langle \theta, \phi(x)\rangle\big\} d\nu(x)$$

is called an *exponential family* with *sufficient statistic* ϕ. The parameter $\theta \in \Theta$ is called the *natural parameter* of $\mathcal{P}(\phi)$, and the set Θ is called the natural parameter space. The function Φ is called the log partition function.[3]

We will denote by $\mathbb{E}_\theta[\cdot]$ the expectation w.r.t. P_θ:

$$\begin{aligned}\mathbb{E}_\theta[g(X)] &= \int_{\mathsf{X}} g(x) dP_\theta(x) \\ &= \int_{\mathsf{X}} g(x) \exp\big\{\langle \theta, \phi(x)\rangle - \Phi(\theta)\big\} d\nu(x).\end{aligned}$$

[3]This usage comes from statistical physics.

Example 1: The simplest example is the Bernoulli distribution. In this case, $\mathsf{X} = \{0, 1\}$, ν is the counting measure, $\phi(x) = x$, and $\Phi(\theta) = \log[1 + \exp(\theta)]$. The natural parameter space Θ is the entire real line. Under this parametrization, we have $p_\theta(X = 1) = e^\theta/(1 + e^\theta)$. □

Example 2 (The Ising Model): Consider an undirected graph $G = (V, E)$ and associate with each vertex $\alpha \in V$ a binary random variable $X_\alpha \in \{-1, +1\}$. In this case, $\mathsf{X} = \{-1, +1\}^V$, ν is the counting measure, and we have the following density for the random variable $X = (X_\alpha : \alpha \in V) \in \{-1, +1\}^V$:

$$p_\theta(x) = \exp\left\{\sum_{\alpha \in V} \theta_\alpha x_\alpha + \sum_{(\alpha,\beta)\in E} \theta_{\alpha\beta} x_\alpha x_\beta - \Phi(\theta)\right\} \quad (5)$$

where θ is the collection of $d = |V| + |E|$ real parameters $(\theta_\alpha : \alpha \in V)$ and $(\theta_{\alpha\beta} : (\alpha, \beta) \in E)$. The log partition function is

$$\Phi(\theta) = \log \sum_{x \in \mathsf{X}} \exp\left\{\sum_{\alpha \in V} \theta_\alpha x_\alpha + \sum_{(\alpha,\beta)\in E} \theta_{\alpha\beta} x_\alpha x_\beta\right\}.$$

The sufficient statistic ϕ is given by the functions $\phi_\alpha : x \mapsto x_\alpha, \alpha \in V$ and $\phi_{\alpha\beta} : x \mapsto x_\alpha x_\beta, (\alpha, \beta) \in E$. Since $\Phi(\theta)$ is finite for any choice of θ, we have $\Theta = \mathbb{R}^d$, with the components of θ appropriately ordered. □

Example 3 [Gaussian Markov Random Fields (MRFs)]: Again, consider an undirected graph $G = (V, E)$. For notational convenience, let us number the vertices as $V = \{1, \ldots, p\}$. A Gaussian MRF on G is a multivariate Gaussian random variable $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, where the covariates X_α and X_β are independent if $(\alpha, \beta) \notin E$. Then $\mathsf{X} = \mathbb{R}^p$ and ν is the Lebesgue measure. The distribution of X can be written exactly as in (5); the log partition function $\Phi(\theta)$ is finite only if the $p \times p$ matrix $\Gamma \stackrel{\triangle}{=} [\theta_{\alpha\beta}]_{\alpha,\beta=1}^p$ is negative definite ($\Gamma \prec 0$), so that the parameter space is $\Theta = \{((\theta_1, \ldots, \theta_p)^T, \Gamma) \in \mathbb{R}^p \times \mathbb{R}^{p\times p} : \Gamma \prec 0, \Gamma = \Gamma^T\}$. □

D. General Properties of Exponential Families

The motivation behind our use of exponential families is twofold. (1) They form a sufficiently rich class of parametric statistical models (which includes MRFs with pairwise interactions) and can be used to describe co-occurrence data, visual scene snapshots, biometric records, and many other categorical and numerical data types. Moreover, they can be used to approximate many nonparametric classes of probability densities [25]. (2) The negative log-likelihood function is *convex* in the natural parameter and *affine* in the sufficient statistic. This structure permits the use of OCP. We will need the following facts about exponential families (proofs can be found in the references listed at the beginning of Section II-C).

1) The log partition function $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is lower semicontinuous on $\mathbb{R}^d$ and infinitely differentiable on Θ.

2) The derivatives of Φ at θ are the cumulants of the random vector $\phi(X) = (\phi_1(X), \ldots, \phi_d(X))$ when $X \sim p_\theta$. In particular,

$$\nabla\Phi(\theta) = \big(\mathbb{E}_\theta\phi_1(X), \ldots, \mathbb{E}_\theta\phi_d(X)\big)^T$$
$$\nabla^2\Phi(\theta) = [\mathrm{Cov}_\theta(\phi_i(X), \phi_j(X))]_{i,j=1}^d .$$

Thus, the Hessian $\nabla^2\Phi(\theta)$, being the covariance matrix of the vector $\phi(X)$, is positive semidefinite, which implies that $\Phi(\theta)$ is a convex function of θ. In particular, Θ, which, by definition, is the essential domain of Φ, is convex.

3) $\Phi(\theta)$ is *steep* (or *essentially smooth*): if $\{\theta_n\} \subset \Theta$ is a sequence converging to some point θ on the boundary of Θ, then $\|\nabla\Phi(\theta_n)\| \to +\infty$ as $n \to \infty$.

4) The relative entropy (Kullback–Leibler divergence) between p_{θ_1} and p_{θ_2} in $\mathcal{P}(\phi)$, defined as $D(p_{\theta_1}\|p_{\theta_2}) = \int_{\mathsf{X}} p_{\theta_1}\log(p_{\theta_1}/p_{\theta_2})d\nu$, can be written as

$$D(p_{\theta_1}\|p_{\theta_2}) = \Phi(\theta_2) - \Phi(\theta_1) - \langle\nabla\Phi(\theta_1), \theta_2 - \theta_1\rangle. \quad (6)$$

From now on, we will use the shorthand $D(\theta_1\|\theta_2)$.

From these properties, it follows that $\Phi : \Theta \to \mathbb{R}$ is a Legendre function, and that the mapping $D_\Phi : \Theta \times \mathrm{Int}\,\Theta \to \mathbb{R}$, defined by $D_\Phi(\theta_1, \theta_2) = D(\theta_2\|\theta_1)$, is a Bregman divergence.

III. Filtering: Sequential Probability Assignment in the Presence of Noise

The first ingredient of FHTAGN is a strategy for assigning a likelihood (or belief) $p_t(\cdot|z^{t-1})$ to the clean symbol x_t based on the past noisy observations z^{t-1}. Alternatively, we can think of the following problem: if x_t is the actual clean symbol that has been generated at time t, then our likelihood $p_t \equiv p_t(x_t|z^{t-1})$, though well defined, is not accessible for observation. Thus, we would like to *estimate* it via some estimator $\widehat{p}_t$, which will depend on the actual observed noisy symbol z_t, as well as on the previously obtained estimates $\widehat{p}^{t-1} = (\widehat{p}_1, \ldots, \widehat{p}_{t-1})$. In the field of signal processing, problems of this kind go under the general heading of *filtering*; this term refers to any situation in which it is desired, at each time t, to obtain an estimate of some clean unobservable quantity *causally* based on noisy past observations.

A. Sequential Probability Assignment: General Formulation

1) Noiseless Observations: Let us first consider the noiseless case, i.e., $z_t = x_t$ for all t. Elements of an arbitrary sequence $\boldsymbol{x} = (x_1, x_2, \ldots) \in \mathsf{X}^\infty$ are revealed to us one at a time, and we make no assumptions on the law that generates $\boldsymbol{x}$. At each time $t = 1, 2, \ldots$, before x_t is revealed, we have to assign a probability density $\widehat{p}_t$ (w.r.t. a fixed dominating measure ν) to the possible values of x_t. When x_t is revealed, we incur the *logarithmic loss* $-\log\widehat{p}_t(x_t)$ (the choice of this loss function is standard and is motivated by information-theoretic considerations; cf., the survey paper by Merhav and Feder [4] for more details). Let $\mathcal{D}$ denote the set of all valid probability densities w.r.t. ν. Then the sequential probability assignment can be represented by a sequence π of mappings $\pi_t : \mathsf{X}^{t-1} \to \mathcal{D}$, so that

$$\widehat{p}_t = \pi_t(x^{t-1}), \qquad \text{or } \widehat{p}_t(x_t) = [\pi_t(x^{t-1})](x_t).$$

We refer to any such sequence of probability assignments π as a *prediction strategy*. Since the probability assignment $\widehat{p}_t$ is a function of the past observations x^{t-1}, we may also view it as a conditional probability density $\widehat{p}_t(\cdot|x^{t-1})$. In the absence of specific probabilistic assumptions on the generation of $\boldsymbol{x}$, it is appropriate to view

$$\widehat{P}_t(A|x^{t-1}) \stackrel{\triangle}{=} \int_A \widehat{p}_t(x|x^{t-1})d\nu(x)$$

as our *belief*, based on the past observations x^{t-1}, that the next observation x_t will lie in a measurable set $A \subseteq \mathsf{X}$. Another way to think about π is as a sequence of joint densities

$$\widehat{p}^T(x^T) = \prod_{t=1}^T \widehat{p}_t(x_t|x^{t-1}), \qquad T = 1, 2, \ldots.$$

In an individual-sequence setting, the performance of a given prediction strategy is compared to the best performance achievable on $\boldsymbol{x}$ by any strategy in some specified comparison class $\mathcal{C}$[4], [16]. Any such comparison strategy is also specified by a sequence of conditional densities $p_t(x_t|x^{t-1})$ of x_t given x^{t-1}. Suppose first that the horizon T is fixed in advance. Given a prediction strategy $\pi = \{\pi_t\}_{t=1}^\infty$, we can define the *regret* w.r.t. $\boldsymbol{p} = \{p_t\} \in \mathcal{C}$ after T time steps as

$$\begin{aligned} &R_T(\pi^T; x^T, p^T) \\ &\quad\stackrel{\triangle}{=} \sum_{t=1}^T \log\frac{1}{[\pi_t(x^{t-1})](x_t)} - \sum_{t=1}^T \log\frac{1}{p_t(x_t|x^{t-1})} \\ &\quad= \sum_{t=1}^T \log\frac{1}{\widehat{p}_t(x_t|x^{t-1})} - \sum_{t=1}^T \log\frac{1}{p_t(x_t|x^{t-1})}. \end{aligned} \quad (7)$$

As before, the distinction between π_t and $\widehat{p}_t$ is that the former is a mapping of X^{t-1} into the space of probability densities $\mathcal{D}$, while the latter is the image of x^{t-1} under π_t.

2) Noisy Observations: We are interested here in a more difficult problem, namely sequential probability assignment in the presence of noise. That is, instead of observing the "clean" symbols $x_t \in \mathsf{X}$, we receive "noisy" symbols $z_t \in \mathsf{Z}$ (where Z is some other observation space). We assume that the noise is stochastic, memoryless and stationary. In other words, at each time t, the noisy observation z_t is given by $z_t = N(x_t, r_t)$, where $\{r_t\}$ is an i.i.d. random sequence and $N(\cdot,\cdot)$ is a fixed deterministic function. There are two key differences between this and the noiseless setting described earlier, namely:

1) the prediction strategy now consists of mappings $\widehat{\pi}_t : \mathsf{Z}^{t-1} \to \mathcal{D}$, where, at each time t, $\widehat{p}_t(\cdot|z^{t-1}) = \widehat{\pi}_t(z^{t-1})$ is the conditional probability density we assign to the *clean* observation x_t at time t given the past *noisy* observations z^{t-1};
2) we cannot compute the true incurred log loss $-\log\widehat{p}_t(x_t|z^{t-1})$.

We are interested in sequential prediction, via the beliefs $\widehat{\pi}_t(z^{t-1}) = \widehat{p}_t(\cdot|z^{t-1})$, of the next clean symbol x_t given the past noisy observations z^{t-1}. We assume, as before, that the clean sequence $\boldsymbol{x}$ is an unknown individual sequence over X. Moreover, under our noise model the noisy observations $\{z_t\}$ are conditionally independent of one another given $\boldsymbol{x}$.

B. Probability Assignment in an Exponential Family via OCP-Based Filtering

We will now show that if the comparison class $\mathcal{C}$ consists of product distributions lying in an *exponential family* with natural parameter $\theta \in \mathbb{R}^d$, then we can use OCP to design a scheme for sequential probability assignment from noisy data. The use of OCP is made possible by the fact that, in an exponential family, the (negative) log likelihood is a *convex* function of the natural parameter and an *affine* function of the sufficient statistic.

Recall from Section II-C that a d-dimensional exponential family consists of probability densities of the form $p_\theta(x) = e^{\langle\theta,\phi(x)\rangle - \Phi(\theta)}$, where the parameter θ lies in a convex subset of $\mathbb{R}^d$. We will consider prediction strategies of the form

$$\widehat{\pi}_t(z^{t-1}) = p_{\widehat{\theta}_t}(\cdot), \qquad t = 1, 2, \ldots \tag{8}$$

where $\widehat{\theta}_t$ is a function of the past noisy observations z^{t-1}. The log-loss function in this particular case takes the form

$$-\log \widehat{p}_t(x_t) = -\langle\widehat{\theta}_t, \phi(x_t)\rangle + \Phi(\widehat{\theta}_t), \qquad x \in \mathsf{X}.$$

Thus, the regret relative to any comparison strategy $\boldsymbol{p}_\theta$ induced by a parameter sequence $\theta = \{\theta_t\} \in \Theta^\infty$ via $p_t(\cdot|x^{t-1}) = p_{\theta_t}(\cdot)$ can be written as

$$\begin{aligned} R_T(\widehat{\pi}^T; x^T, p_\theta^T) &= \sum_{t=1}^T \log \frac{1}{p_{\widehat{\theta}_t}(x_t)} - \sum_{t=1}^T \log \frac{1}{p_{\theta_t}(x_t)} \\ &= \sum_{t=1}^T \Big[\ell(\widehat{\theta}_t, x_t) - \ell(\theta_t, x_t)\Big] \end{aligned}$$

where we have defined the function

$$\ell(\theta, x) \stackrel{\triangle}{=} -\langle\theta, \phi(x)\rangle + \Phi(\theta).$$

Because the log partition function Φ is convex, the function $\theta \mapsto \ell(\theta, x)$ is convex for every fixed $x \in \mathsf{X}$. Therefore, if the observations were noiseless, i.e., $z_t = x_t$ for all t, then MD could have been used to design an appropriate strategy of the form (8). As we will show next, this approach also works in the noisy case, provided an unbiased estimator of $\phi(x)$ based on the noisy observation z is available. In fact, our results in the noisy setting contain the noiseless setting as a special case.

Let us fix an exponential family $\mathcal{P}(\phi)$. We will consider the comparison class consisting of product distributions, where each marginal belongs to a certain subset of $\mathcal{P}(\phi)$. Specifically, let Λ be a closed, convex subset of Θ. We take $\mathcal{C}$ to consist of prediction strategies $\boldsymbol{p}_\theta$, where $\theta = (\theta_1, \theta_2, \ldots) \in \Lambda^\infty$ ranges over all infinite sequences over Λ, and each $\boldsymbol{p}_\theta$ is of the product form $p_{\theta_1} \otimes p_{\theta_2} \otimes \ldots$, i.e.,

$$p_{t,\theta}(\cdot|x^{t-1}) = p_{\theta_t}(\cdot), \qquad x^{t-1} \in \mathsf{X}^{t-1}, t = 1, 2, \ldots. \tag{9}$$

In other words, each prediction strategy in $\mathcal{C}$ corresponds to a time-varying product density whose marginals belong to $\{p_\theta : \theta \in \Lambda\}$. From now on, we will use the term "strategy" to refer to an infinite sequence $\theta \in \Lambda^\infty$; the corresponding object $\boldsymbol{p}_\theta$ will be implied.

Consider the noisy observation model $z_t = N(x_t, r_t)$, where $\{r_t\}$ is the i.i.d. noise process and $N(\cdot,\cdot)$ is a known deterministic function. We make the following assumption.

Assumption 1: There exists a function $h : \mathsf{Z} \to \mathbb{R}^d$, such that $\mathbb{E}[h(z)|x] = \phi(x)$, where the expectation is taken w.r.t. the noise input r. In other words, $h(z)$ is an unbiased estimator of the sufficient statistic $\phi(x)$.

Here, the conditional notation $\mathbb{E}[h(z)|x]$ is shorthand for the fact that the expectation is taken w.r.t. the common distribution Q of $r_1, r_2, \ldots$, while keeping the clean input x fixed:

$$\mathbb{E}[h(z)|x] = \int h(N(x, r))dQ(r). \tag{10}$$

Example 4: In Example 1, $x \in \{0,1\}^d$ and $\phi(x) = x$, $r \in \{0,1\}^d$, where each component $r(i)$ is a $\mathrm{Bernoulli}(p)$ random variable independent of everything else ($p < 1/2$), and

$$z(i) = x(i) \oplus r(i), \qquad i = 1, \ldots, d. \tag{11}$$

In other words, every component of z is independently related to the corresponding component of x via a binary symmetric channel (BSC) with crossover probability p[26]. Then an unbiased estimator for $\phi(x) \equiv x$ is given by $h(z) = (h_1(z), \ldots, h_d(z))^T$, where

$$h_i(z) = \frac{z(i) - p}{1 - 2p}, \qquad i = 1, \ldots, d.$$

Example 5: Consider the Ising model from Example 2 and suppose that each x_α, $\alpha \in V$, is independently corrupted by a BSC with crossover probability p. Then an unbiased estimator for $\phi(x)$ is given by

$$\begin{aligned} h_\alpha(z) &= \frac{z_\alpha - p}{1 - 2p}, \qquad \alpha \in V \\ h_{\alpha\beta}(z) &= \frac{z_\alpha - p}{1 - 2p} \cdot \frac{z_\beta - p}{1 - 2p}, \qquad (\alpha, \beta) \in E \end{aligned}$$

so that we have $\mathbb{E}[h_\alpha(z)|x] = \phi_\alpha(x) = x_\alpha$ and $\mathbb{E}[h_{\alpha\beta}(z)|x] = \phi_{\alpha\beta}(x) = x_\alpha x_\beta$. □

Example 6: Consider the Gaussian MRF from Example 3 and suppose that each x_α, $\alpha \in V$, is independently corrupted by an additive white Gaussian noise (AWGN) channel with noise variance σ^2[26]. In other words, $z = (z_\alpha : \alpha \in V)$, where, for each $\alpha \in V$, $z_\alpha = x_\alpha + r_\alpha$ with $r_\alpha \sim \mathrm{Normal}(0, \sigma^2)$ independent of all other $r_\beta, \beta \neq \alpha$. Then an unbiased estimator for $\phi(x)$ is given by

$$h_\alpha(z) = z_\alpha, h_{\alpha\beta}(z) = \begin{cases} z_\alpha^2 - \sigma^2, & \alpha = \beta \\ z_\alpha z_\beta, & \alpha \neq \beta \end{cases}, \qquad \alpha, \beta \in V$$

so that $\mathbb{E}[h_\alpha(z)|x] = x_\alpha$ and $\mathbb{E}[h_{\alpha\beta}(z)|x] = x_\alpha x_\beta$. □

By virtue of our Assumption 1, the *filtering loss*

$$\widehat{\ell}(\theta, z_t) \stackrel{\triangle}{=} -\langle\theta, h(z_t)\rangle + \Phi(\theta)$$

TABLE I
CORRESPONDENCE BETWEEN THE GENERIC MD AND ALGORITHM 3

	Generic MD	Algorithm 3
Convex set	U	Θ
Cost functions	f_t	$\widehat{\ell}(\cdot, z_t) \equiv -\langle \cdot, h(z_t)\rangle + \Phi(\cdot)$
Project onto	S	Λ
Legendre potential	F	Φ
Bregman divergence	$D_F(\cdot,\cdot)$	$D_\Phi(\cdot,\cdot) \equiv D(\cdot\|\|\cdot)$

is an unbiased estimator of the true log loss $\ell(\theta, x_t)$ for any $\theta \in \Theta$:

$$\mathbb{E}\left[\widehat{\ell}(\theta, z_t)\middle| x_t\right] = -\langle \theta, \mathbb{E}[h(z_t)|x_t]\rangle + \Phi(\theta) = -\langle \theta, \phi(x_t)\rangle + \Phi(\theta) = \ell(\theta, x_t).$$

This leads to the following prediction strategy.

Algorithm 3 Sequential Probability Assignment via Noisy Mirror Descent

Require: A closed, convex set $\Lambda \subset \Theta$; a decreasing sequence of strictly positive step sizes $\{\eta_t\}$

Initialize with $\widehat{\theta}_1 \in \Lambda$

for $t = 1, 2, \ldots$ **do**

Acquire new noisy observation z_t
Compute the filtering loss $\widehat{\ell}_t(\widehat{\theta}_t) = -\langle \widehat{\theta}_t, h(z_t)\rangle + \Phi(\widehat{\theta}_t)$
Output

$$\widehat{\theta}_{t+1} = \arg\min_{\theta\in\Lambda}\left[\eta_t\langle \nabla\widehat{\ell}_t(\widehat{\theta}_t), \theta\rangle + D(\widehat{\theta}_t\|\theta)\right]$$

end for

This induces the following sequential probability assignment strategy:

$$\pi_t = p_{\widehat{\theta}_t}(\cdot) \tag{12a}$$

$$\widehat{\theta}_{t+1} = \arg\min_{\theta\in\Lambda}\left[\left\langle \theta, \nabla\Phi(\widehat{\theta}_t) - h(z_t)\right\rangle + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\theta)\right]. \tag{12b}$$

For the reader's convenience, Table I shows the correspondence between the objects used in Algorithm 3 and the generic MD strategy, i.e., Algorithm 2.

This approach has the following features.

1) The geometry of exponential families leads to a natural choice of the Legendre potential and the corresponding Bregman divergence to be used in the MD updates, namely the log partition function Φ and the Kullback–Leibler divergence $D(\cdot\|\cdot)$.
2) The optimization at each time can be computed using only the current noisy observation z_t and the probability density $\widehat{p}_t$ estimated at the previous time; it is not necessary to keep all observations in memory to ensure strong performance.

Azoury and Warmuth [27] proposed and analyzed an algorithm similar to (12) in the setting of online density estimation over an exponential family. However, they did not consider noisy observations and only proved regret bounds for a couple of specific exponential families. One of the contributions of this paper is to demonstrate that minimax (logarithmic) regret bounds against static strategies can be obtained for a *general* exponential family, subject to mild restrictions on the feasible set $\Lambda \subseteq \Theta$. This provides an answer to the question posed by Azoury and Warmuth about whether it is possible to attain logarithmic regret for a general exponential family.

C. Regret Bounds for OCP-Based Filter

We will now establish the following bounds on the expected regret of Algorithm 3.

1) If the comparison class $\mathcal{C}$ consists of *static* strategies $\theta_1 = \theta_2 = \ldots$ over Λ, then, under certain regularity conditions on Λ and with properly chosen step sizes $\{\eta_t\}$, the expected regret of the strategy in (12) will be $O(\log T)$.
2) Given a strategy θ and a time horizon T, define the *variation* of θ from $t = 1$ to $t = T$ as

$$V_T(\theta) \stackrel{\triangle}{=} \sum_{t=1}^{T}\|\theta_t - \theta_{t+1}\| \tag{13}$$

where $\|\cdot\|$ is taken to be the ℓ_2 norm for concreteness. If the comparison class $\mathcal{C}$ consists of all time-varying strategies $\theta = \{\theta_t\}$ over Λ, then, under certain regularity conditions on Λ and with properly chosen step sizes $\{\eta_t\}$, the expected regret of the algorithm in (12) will be $O\big((V_T(\theta)+1)\sqrt{T}\big)$.

The expectation in both cases is taken w.r.t. the noise process $\{r_t\}$. Moreover, in the absence of noise (i.e., $z_t = x_t$ for all t), the aforementioned regret bounds will hold for all observation sequences $\boldsymbol{x}$.

We will bound the regret of Algorithm 3 in two steps. In the first step, we will obtain bounds on the regret computed using the filtering losses $\widehat{\ell}(\cdot,\cdot)$ that hold for *any* realization of the noisy sequence $\boldsymbol{z} = \{z_t\}$. In the second step, we will use a martingale argument along the lines of Weissman and Merhav [28] to show that the expected "true" regret is bounded by the expected filtering regret.

1) Regret Bounds for the Filtering Loss: We will consider time-varying strategies of the form (9), where the set Λ is restricted in the following way. Given a positive constant $H > 0$, define the set

$$\Theta_H \stackrel{\triangle}{=} \left\{\theta \in \Theta : \nabla^2\Phi(\theta) \succeq 2HI_{d\times d}\right\}$$

where $I_{d\times d}$ denotes the $d\times d$ identity matrix, and the matrix inequality $A \succeq B$ denotes the fact that $A - B$ is positive semidefinite. Note that the Hessian $\nabla^2\Phi(\theta)$ is equal to

$$J(\theta) \stackrel{\triangle}{=} -\mathbb{E}_\theta[\nabla^2_\theta \log p_\theta(X)]$$

which is the Fisher information matrix at θ[23], [24], [29]. In other words, Θ_H consists of all parameter vectors $\theta \in \Theta$, for which the eigenvalues of the Fisher information matrix over Λ are bounded from below by $2H$. Yet another motivation for our definition of Θ_H comes from the fact that $\nabla^2\Phi(\theta)$ is the covariance matrix of the random vector $\phi(X) = (\phi_1(X), \ldots, \phi_d(X))^T$ when $X \sim P_\theta$. Thus, a strictly positive uniform lower bound on the eigenvalues of this covariance matrix implies that any coordinate of $\phi(X)$ is sufficiently

"informative" about (or correlates well with) the remaining coordinates. We will assume in the sequel that Λ is any closed, convex subset of Θ_H.

For any strategy $\theta = \{\theta_t\}$, define the cumulative true and estimated losses

$$L_{\theta,T}(x^T) \stackrel{\triangle}{=} \sum_{t=1}^{T} \ell(\theta_t, x_t)$$
$$\widehat{L}_{\theta,T}(z^T) \stackrel{\triangle}{=} \sum_{t=1}^{T} \widehat{\ell}(\theta_t, z_t)$$

and the difference

$$\Delta_{\theta,T}(x^T, z^T) \stackrel{\triangle}{=} L_{\theta,T}(x^T) - \widehat{L}_{\theta,T}(z^T) = \sum_{t=1}^{T} \langle \theta_t, h(z_t) - \phi(x_t) \rangle.$$

When θ is a static strategy corresponding to $\theta \in \Lambda$, we will write $L_{\theta,T}(x^T)$, $\widehat{L}_{\theta,T}(z^T)$, and $\Delta_{\theta,T}(x^T, z^T)$. We first establish a logarithmic regret bound against static strategies in Λ. The theorem below improves on our earlier result from [8].

Theorem 1 (Logarithmic Regret Against Static Strategies): Let Λ be any closed, convex subset of Θ_H, and let $\widehat{\theta} = \{\widehat{\theta}_t\}$ be the sequence of parameters in Λ computed from the noisy sequence $\boldsymbol{z} = \{z_t\}$ using the OCP procedure shown in Algorithm 3 with step sizes $\eta_t = 1/t$. Then, for any $\theta \in \Lambda$, we have

$$\widehat{L}_{\widehat{\theta},T}(z^T) \le \widehat{L}_{\theta,T}(z^T) + \frac{(K(z^T) + M)^2}{H}(\log T + 1) \quad (14)$$

where

$$K(z^T) \stackrel{\triangle}{=} \frac{1}{2} \max_{1 \le t \le T} \|h(z_t)\| \text{and } M \stackrel{\triangle}{=} \frac{1}{2} \max_{\theta \in \Lambda} \|\nabla \Phi(\theta)\|.$$

Proof: Appendix A. ■

With larger step sizes $\eta_t = 1/\sqrt{t}$, it is possible to compete against *time-varying* strategies $\theta = \{\theta_t\}$, provided the variation is sufficiently slow.

Theorem 2 (Regret Against Time-Varying Strategies): Again, let Λ be any closed, convex subset of Θ_H. Let $\widehat{\theta}$ be the sequence of parameters in Λ computed from the noisy sequence $\boldsymbol{z} = \{z_t\}$ using the OCP procedure shown in Algorithm 3 with step sizes $\eta_t = 1/\sqrt{t}$. Then, for any sequence $\theta = \{\theta_t\}$ over Λ, we have

$$\widehat{L}_{\widehat{\theta},T}(z^T) \le \widehat{L}_{\theta,T}(z^T) + 4M\sqrt{T}V_T(\theta) + \frac{(K(z^T) + M)^2}{H}(2\sqrt{T} - 1)$$

where $K(z^T)$ and M are defined as in Theorem 1, and $V_T(\theta)$ is defined in (13).

Proof: Appendix B. ■

Remark 3: Regret bounds against a dynamically changing reference strategy have been derived in a variety of contexts, including prediction with expert advice with finitely many time-varying experts [30], sequential linear prediction [31], general OCP [5], and sequential universal lossless source coding (which is equivalent to universal prediction with log loss) [32], [33]. It is useful to compare the result of Theorem 2 with some of these bounds. The results of [31] and [5] assume fixed and known horizon T. Furthermore, Herbster and Warmuth [31] assume that the loss function is of *subquadratic type* (see, e.g., [16, Section 11.4]) and use a different scale-dependent notion of regret and a carefully adjusted time-independent step size. On the other hand, Zinkevich [5] uses a constant step size η and obtains a regret bound of the form $O(V_T/\eta + T\eta)$, where the constants implicit in the $O(\cdot)$ notation depend on the diameter of the feasible set and on the maximum norm of the (sub)gradient of the loss function. It is not hard to modify the proof of our Theorem 2 to obtain a similar regret bound in our case, including the constants. The results in [33] (which extend the work of Willems [32]) deal with universal prediction of piecewise-constant memoryless sources under log loss. They propose two types of algorithms — those with linearly growing per-round computational complexity and those with fixed per-round complexity. For the first type, and with $V_T = O(1)$, they obtain $O(\log T)$ regret (which is optimal [34]); for the second type, again with $V_T = O(1)$, they develop two schemes, one of which attains $O(\log T)$ regret for certain sources with "large" jumps and $O(T \log\log T/\log T)$ in general, while the other always achieves $O(\sqrt{T \log T})$ regret. Since our algorithms have fixed per-letter complexity, it is clear that we cannot achieve the optimal $O(\log T)$ regret; however, our $O(\sqrt{T})$ regret in the $V_T = O(1)$ case compares favorably against the second fixed-complexity algorithm of [33]. Of course, the reason why our algorithms have fixed per-letter complexity comes from the special structure of the log-loss function for an exponential family, which maps the problem into an instance of OCP on the underlying parameter space.

2) Bounds on the Expected True Regret: We now proceed to establish regret bounds on $L_{\theta,T}(x^T)$. The bounds of Theorems 1 and 2 reflected how close our cumulative loss might be to that of a competing strategy on noisy data. We now show that our proposed strategy ensures that the *expected* cumulative loss on the *unobserved* clean data is close to that of competing strategies. First, we need the following lemma, which is similar to [28, Lemma 1].

Lemma 2: Let $\boldsymbol{r} = \{r_t\}$ be the i.i.d. observation noise process. For each t, let $\mathcal{R}_t$ denote the σ-algebra generated by $r_1, \ldots, r_t$. Let $\theta = \{\theta_t\}$ be a sequence of probability assignments, such that each $\theta_t = \theta_t(z^{t-1})$. Then, for any individual sequence $\boldsymbol{x} = \{x_t\}$, $\{\Delta_{\theta,t}(x^t, z^t), \mathcal{R}_t\}$ is a martingale, and so $\mathbb{E}\widehat{L}_{\theta,T}(z^T) = \mathbb{E}L_{\theta,T}(x^T)$ for each T. The expectation is conditional on the underlying clean sequence $\boldsymbol{x}$, cf. (10).

Proof: Appendix C. ■

This leads to regret bounds on the proposed OCP-based filter.

Theorem 3: Consider the setting of Theorem 1. Then we have

$$\mathbb{E}[L_{\widehat{\theta},T}(x^T)] \le \inf_{\theta \in \Lambda} L_{\theta,T}(x^T) + \frac{\mathbb{E}[(K(z^T) + M)^2]}{H}(\log T + 1). \quad (15)$$

Likewise, in the setting of Theorem 2, we have

$$\mathbb{E}[L_{\widehat{\theta},T}(x^T)] \leq \inf_\theta \left[L_{\theta,T}(x^T) + 4M\sqrt{T}V_T(\theta)\right] + \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(2\sqrt{T}-1) \quad (16)$$

where the infimum is over all strategies θ over Λ, and the expectation is conditional on the underlying clean sequence $\boldsymbol{x}$.

Proof: We will only prove (16); the proof of (15) is similar. Proceeding analogously to the proof of [28, Th. 4], we have

$$\begin{aligned}
\mathbb{E}L_{\widehat{\theta},T}(x^T) &= \mathbb{E}\widehat{L}_{\widehat{\theta},T}(z^T) \\
&\leq \mathbb{E}\Big\{ \inf_\theta \left[\widehat{L}_{\theta,T}(z^T) + 4M\sqrt{T}V_T(\theta)\right] \\
&\qquad + \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(2\sqrt{T}-1)\Big\} \\
&\leq \inf_\theta \left[\mathbb{E}\widehat{L}_{\theta,T}(z^T) + 4M\sqrt{T}V_T(\theta)\right] \\
&\qquad + \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(2\sqrt{T}-1) \\
&= \inf_\theta \left[L_{\theta,T}(x^T) + 4M\sqrt{T}V_T(\theta)\right] \\
&\qquad + \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(2\sqrt{T}-1)
\end{aligned}$$

where the first step follows from Lemma 2, the second from Theorem 2, the third from the fact that $\mathbb{E}\inf[\cdot] \leq \inf\mathbb{E}[\cdot]$, and the last from Lemma 2 and the fact that the expectation is taken with respect to the distribution of $z_t|x_t$. ■

Remark 4: In the usual regret notation, the bounds of Theorem 3 can be written as follows:

$$\mathbb{E}R_T(\pi^T; x^T, \theta) \leq \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(\log T + 1)$$

and

$$\mathbb{E}R_T(\pi^T; x^T, \theta^T) \leq 4M\sqrt{T}V_T(\theta) + \frac{\mathbb{E}[(K(z^T)+M)^2]}{H}(2\sqrt{T}-1).$$

3) Minimax Optimality and Hannan Consistency: Finally, we make a few comments regarding minimax optimality and Hannan consistency of the strategies described in this section.

Recall the OCP game described in Section II-A. During each round $t = 1, 2, \ldots, T$, the Forecaster plays a point $\widehat{u}_t \in \mathsf{U}$, the Environment responds with a convex function $f_t \in \mathcal{F}$, and the Forecaster incurs the cost $f_t(\widehat{u}_t)$. The Forecaster's goal is to keep the cumulative cost $\sum_{t=1}^T f_t(\widehat{u}_t)$ as low as possible. Let us suppose, moreover, that the Environment is antagonistic in that it tries to choose the functions f_t, so that the current cumulative cost $\sum_{s=1}^t f_s(\widehat{u}_s)$ is as high as possible, given the past moves of the Forecaster $\widehat{u}^t$ and the past cost functions f^{t-1}. To allow the Environment more freedom, we assume that the cost function at time t is selected from a set $\mathcal{F}_t$ which may depend on the current move $\widehat{u}_t$. With this in mind, let us define, following [14], the *minimax value* of the game as

$$R_T^*(\mathsf{U}, \mathcal{F}^T) = \inf_{u_1\in\mathsf{U}} \sup_{f_1\in\mathcal{F}_1} \ldots \inf_{u_T\in\mathsf{U}} \sup_{f_T\in\mathcal{F}_T} \left\{\sum_{t=1}^T f_t(u_t) - \inf_{u\in\mathsf{U}} \sum_{t=1}^T f_t(u)\right\}. \quad (17)$$

In words, $R_T^*(\mathsf{U}, \mathcal{F}^T)$ is the worst-case regret of an *optimal* strategy for the Forecaster. The order of the infima and the suprema in (17) reflects the order of the moves and the causality restrictions. Thus, the Forecaster's move at time t may depend only on his moves and on the cost functions revealed at times $s = 1, \ldots, t-1$; the Environment's cost function at time t may depend only on the Forecaster's moves at times $s = 1, \ldots, t$ and on the cost functions at times $s = 1, \ldots, t-1$. Then we have the following bounds.

Theorem 4 ([14]): Suppose that $\mathsf{U} \subset \mathbb{R}^d$ is compact and convex, and there exist some constants $G, \sigma > 0$ such that at each time t the functions in $\mathcal{F}_t$ satisfy the following conditions:

$$\mathcal{F}_t = \left\{f : \|\nabla f(\widehat{u}_t)\| \leq G, \nabla^2 f \succeq \sigma I_{d\times d}\right\}.$$

Then

$$\frac{G^2}{2\sigma}\log(T+1) \leq R_T^*(\mathsf{U}, \mathcal{F}^T) \leq \frac{G^2}{2\sigma}(\log T + 1).$$

We can particularize this result to our case. Consider the setting of Theorem 1 in the noiseless regime: $x_t = z_t, \forall t$. Let us fix a constant $K > 0$ and let $\mathsf{U} = \Lambda$ and

$$\mathcal{F}_1 = \ldots = \mathcal{F}_T = \Big\{f_x = \ell(\cdot, x) = -\langle\cdot, \phi(x)\rangle + \Phi(\cdot) : x \in \mathsf{X}, \|\phi(x)\| \leq 2K\Big\}.$$

Then, by hypothesis, each $f_x \in \mathcal{F}_t$ satisfies

$$\|\nabla f_x(\theta)\| \leq \|\phi(x)\| + \|\nabla\Phi(\theta)\| \leq 2(K+M).$$

Moreover, each $f_x \in \mathcal{F}_t$ satisfies $\nabla^2 f_x(\theta) = \nabla^2\Phi(\theta) \succeq 2HI_{d\times d}$ at every $\theta \in \Lambda$. Thus, applying Theorem 4 with $G = 2(K+M)$ and $\sigma = 2H$, we get

$$\frac{(K+M)^2}{H}\log(T+1) \leq R_T^*(\Lambda, \mathcal{F}^T) \leq \frac{(K+M)^2}{H}(\log T + 1).$$

On the other hand, with these assumptions we have $K(z^T) = K(x^T) = K$, and so our regret bound from Theorem 1 is of the form

$$R_T \leq \frac{(K+M)^2}{H}(\log T + 1)$$

and thus we attain the minimax value $R_T^*(\Lambda, \mathcal{F}^T)$.

We can also establish a weak form of Hannan consistency for the OCP filters of Theorems 1 and 2. Let $\mathcal{G}$ denote the set of all sequences $\boldsymbol{x} \in \mathsf{X}^\infty$, such that

$$\mathbb{E}[(K(z^T)+M)^2] = o(T/\log T). \quad (18)$$

For example, if h and ϕ have uniformly bounded norms, then $\mathcal{G} \equiv \mathsf{X}^\infty$. Then the filter of Theorem 1 satisfies

$$\limsup_{T\to\infty} \frac{1}{T} \sup_{\theta\in\Lambda} \mathbb{E}[R_T(\pi^T; x^T, \theta)] = 0$$

for any $\boldsymbol{x} \in \mathcal{G}$. As for the filter of Theorem 2, consider any set $\mathcal{C}$ of all time-varying strategies $\theta \in \Lambda^\infty$, such that

$$\sup_{\theta\in\mathcal{C}} V_T(\theta) = o(\sqrt{T}).$$

Then

$$\limsup_{T\to\infty} \frac{1}{T} \sup_{\theta\in\mathcal{C}} \mathbb{E}[R_T(\pi^T; x^T, \theta^T)] = 0 \tag{19}$$

for any $\boldsymbol{x} \in \mathcal{G}$. For example, consider the set $\mathcal{C}$ of all piecewise constant strategies $\theta \in \Lambda^\infty$, such that k_T, the maximum number of switches in any θ^{T+1}, satisfies $k_T = o(\sqrt{T})$. Assume also that Λ is compact. Then for any $\theta \in \mathcal{C}$ we have

$$\begin{aligned}\sup_{\theta\in\mathcal{C}} V_T(\theta) &= \sup_{\theta\in\mathcal{C}} \sum_{t=1}^{T} \|\theta_{t+1} - \theta_t\| \\ &\le \operatorname{diam}(\Lambda) \cdot k_T \\ &= o(\sqrt{T})\end{aligned} \tag{20}$$

so we have Hannan consistency.

Stronger bounds that hold *uniformly* over the choice of the observation sequence $\boldsymbol{x}$ are also possible, provided the convergence in (18) is uniform; in other words, if we have a set $\bar{\mathcal{G}} \subseteq \mathsf{X}^\infty$, such that

$$\limsup_{T\to\infty} \sup_{\boldsymbol{x}\in\bar{\mathcal{G}}} \frac{\mathbb{E}[(K(z^T) + M)^2] \log T}{T} = 0$$

then the convergence in (19) and (20) is uniform in $\boldsymbol{x} \in \bar{\mathcal{G}}$.

IV. Hedging: Sequential Threshold Selection for Anomaly Detection

In the preceding section, we have shown how to perform *filtering*, i.e., how to assign a belief $\widehat{p}_t = \widehat{p}_t(z^t)$ to the clean symbol x_t, such that

$$\sum_{t=1}^{T} \mathbb{E}\left[\log \frac{1}{\widehat{p}_t(z^{t-1})} \middle| x^t \right] \approx \sum_{t=1}^{T} \log \frac{1}{p_t(x_t|x^{t-1})}$$

where $p_t(\cdot|\cdot)$, $t = 1, \ldots, T$, is the optimal sequence of conditional probability assignments, under a (possibly time-varying) exponential family model, for the entire clean observation sequence x^T.

The second ingredient of FHTAGN is *hedging*, i.e., sequential adjustment of the threshold τ_t, such that whenever $\zeta_t \stackrel{\triangle}{=} \zeta(\widehat{p}_t) < \tau_t$, where $\zeta : \mathbb{R}^+ \to \mathbb{R}$ is a user-specified monotonically increasing function, we flag z_t as anomalous.

Remark 5: Note that this formulation is equivalent to sequentially setting a threshold $\widetilde{\tau}_t$ such that, whenever $\widehat{p}_t < \widetilde{\tau}_t$, we flag z_t as anomalous. Using the monotone transformation ζ allows us to sidestep challenging numerical issues when $\widehat{p}_t$ is very small. We will elaborate on this point later.

In order to choose an appropriate τ_t, we rely on feedback from an end user. Specifically, let the end user set the label y_t as 1 if z_t is anomalous and -1 if z_t is not anomalous. However, since it is often desirable to minimize human intervention and analysis of each observation, we seek to limit the amount of feedback received. To this end, two possible scenarios could be considered.

1) At each time t, the Forecaster randomly decides whether to request a label from the end user. A label is requested with probability that may depend on f_t and τ_t.
2) At each time t, the end user arbitrarily chooses whether to provide a label to the Forecaster; the Forecaster has no control over whether or not it receives a label.

As we will see, the advantage of the first approach is that it allows us to bound the average performance over all possible choices of times at which labels are received, resulting in stronger bounds. The advantage of the second approach is that it may be more practical or convenient in many settings. For instance, if an anomaly is by chance noticed by the end user or if an event flagged by the Forecaster as anomalous is, upon further investigation, determined to be nonanomalous, this information is readily available and can easily be provided to the Forecaster. In the sequel, we will develop performance bounds for both of these regimes.

In both settings, we will be interested in the number of mistakes made by the Forecaster over T time steps. At each time step t, let $\widehat{y}_t$ denote the binary label output by the Forecaster, $\widehat{y}_t = \operatorname{sgn}(\tau_t - \zeta_t)$, where we define $\operatorname{sgn}(a) = -1$ if $a \le 0$ and $+1$ if $a > 0$. The number of mistakes over T time steps is given by

$$\sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} \equiv \sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau_t - \zeta_t) \neq y_t\}}. \tag{21}$$

For simplicity, we assume here that the time horizon T is known in advance. We would like to obtain regret bounds relative to any fixed threshold $\tau \in [\tau_{\min}, \tau_{\max}]$ that could be chosen in hindsight after having observed the entire sequence of (ζ-transformed) probability assignments $\{\zeta_t\}_{t=1}^T$ and feedback $\{y_t\}_{t=1}^T$ (note that some y_ts may be "empty," reflecting the lack of availability of feedback at the corresponding times). Here, $\tau_{\min}$ and $\tau_{\max}$ are some user-defined minimum and maximum threshold levels. Ideally, we would like to bound

$$\sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau_t - \zeta_t) \neq y_t\}} - \inf_{\tau\in[\tau_{\min},\tau_{\max}]} \sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau - \zeta_t) \neq y_t\}}. \tag{22}$$

However, analyzing this expression is difficult owing to the fact that the function $\tau \mapsto 1_{\{\operatorname{sgn}(\tau-\zeta)\neq y\}}$ is not convex in τ. To deal with this difficulty, we will use the standard technique of replacing the comparator loss with a convex *surrogate function*. A frequently used surrogate is the *hinge loss*

$$\ell(s, y) \stackrel{\triangle}{=} (1 - sy)_+$$

where $(\alpha)_+ = \max\{0, \alpha\}$. Indeed, for any ζ, τ, and y we have

$$1_{\{\operatorname{sgn}(\tau-\zeta)\neq y\}} \le 1_{\{(\tau-\zeta)y<0\}} \le \big(1 - (\tau - \zeta)y\big)_+.$$

Thus, instead of (22), we will bound the "regret"

$$R_T(\tau) \stackrel{\triangle}{=} \sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} - \sum_{t=1}^{T} \ell_t(\tau) \tag{23}$$

where $\ell_t(\tau)$ is shorthand for $\ell(\tau - \zeta_t, y_t)$. In the following, we show that it is possible to obtain $O(\sqrt{T})$ surrogate regret using a modified MD (more precisely, projected subgradient descent) strategy. The modifications are necessary to incorporate feedback into the updates.

We point out that the algorithms underlying the hedging step may be used in conjunction with any other method for assigning beliefs to incoming observations; however, together with our OCP-based filtering, they result in a low-complexity anomaly detection system with provable performance guarantees.

A. Anomaly Detection With Full Feedback

In order to obtain bounds on the surrogate regret (23), we first analyze the ideal situation in which the Forecaster always receives feedback. Let $\Pi(\cdot)$ denote the projection onto the interval $[\tau_{\min}, \tau_{\max}]$:

$$\Pi(\alpha) \stackrel{\triangle}{=} \underset{\tau \in [\tau_{\min}, \tau_{\max}]}{\arg\min} \; (\tau - \alpha)^2.$$

In this setting, the following simple algorithm, which is essentially the perceptron algorithm (see, e.g., [16, Chapter 12]) with projections, does the job.

Algorithm 4 Anomaly detection with full feedback

Parameters: real numbers $\eta > 0$, $\tau_{\min} < \tau_{\max}$

Initialize: $\tau_1 = \tau_{\min}$

for $t = 1, 2, \ldots, T$ **do**

Receive the estimated likelihood $\widehat{p}_t$, set $\zeta_t = \zeta(\widehat{p}_t)$
if $\zeta_t < \tau_t$ **then** flag z_t as an anomaly: $\widehat{y}_t = 1$ **else** let $\widehat{y}_t = -1$
Obtain feedback y_t
Let $\tau_{t+1} = \Pi\big(\tau_t + \eta y_t 1_{\{\widehat{y}_t \neq y_t\}}\big)$

end for

Intuitively, the idea is this: if the Forecaster correctly assigns the label $\widehat{y}_t$ to z_t, then the threshold stays the same; if the Forecaster incorrectly labels a nominal observation ($y_t = -1$) as anomalous ($\widehat{y}_t = 1$), then the threshold is lowered: $\tau_{t+1} \approx \tau_t - \eta$; if the Forecaster incorrectly labels an anomalous observation ($y_t = 1$) as nominal ($\widehat{y}_t = -1$), then the threshold is raised: $\tau_{t+1} \approx \tau_t + \eta$. We also observe that the aforementioned algorithm is of a MD type with the Legendre potential $F(u) = u^2/2$, with one crucial difference: the current threshold τ_t is updated only when the Forecaster makes a mistake. We obtain the following regret bound:

Theorem 5: Fix a time horizon T and consider the Forecaster acting according to Algorithm 4 with parameter $\eta = (\tau_{\max} - \tau_{\min})/\sqrt{T}$. Then, for any $\tau \in [\tau_{\min}, \tau_{\max}]$, we have

$$R_T(\tau) = \sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} - \sum_{t=1}^{T} \ell_t(\tau) \leq (\tau_{\max} - \tau_{\min})\sqrt{T}. \tag{24}$$

Proof: Appendix D. ■

B. Random, Forecaster-Driven Feedback Times

We can now address the problem of online anomaly detection when the Forecaster has an option to query the end user for feedback when the Forecaster is not sufficiently confident about its own decision [35]. Consider the following *label-efficient* Forecaster for anomaly detection using sequential probability assignments.

Algorithm 5 Label-efficient anomaly detection

Parameters: real numbers $\eta > 0$, $\tau_{\min} < \tau_{\max}$

Initialize: $\tau_1 = \tau_{\min}$

for $t = 1, 2, \ldots$ **do**

Receive the estimated likelihood $\widehat{p}_t$, set $\zeta_t = \zeta(\widehat{p}_t)$
if $\zeta_t < \tau_t$ **then** flag z_t as an anomaly: $\widehat{y}_t = 1$ **else** let $\widehat{y}_t = -1$
Draw a Bernoulli random variable U_t such that $\Pr[U_t = 1 | U^{t-1}] = 1/(1 + |\zeta_t - \tau_t|)$
if $U_t = 1$ **then** request feedback y_t and let $\tau_{t+1} = \Pi\Big(\tau_t + \eta y_t 1_{\{\widehat{y}_t \neq y_t\}}\Big)$ **else** let $\tau_{t+1} = \tau_t$

end for

This algorithm is, essentially the label-efficient perceptron (see [16, Sec. 12.4]) with projections, and it attains the following regret.

Theorem 6: Fix a time horizon T and consider the label efficient Forecaster run with parameter $\eta = 1/\sqrt{T}$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}}\right] \leq \sum_{t=1}^{T} \ell_t(\tau) + (\tau_{\max} - \tau_{\min})\sqrt{T}$$

where the expectation is taken with respect to $\{U_t\}$.

Remark 6 (Computational Issues Involving Very Small Numbers): In some applications, the beliefs $\widehat{p}_t$ may be very small. (For instance, in the Enron example presented in the experimental results, $\widehat{p}_t = O(e^{-100})$.) In such a case, we will have $\Pr[U_t = 1 | U^{t-1}] \approx 1$, and our anomaly detection engine will request feedback almost all the time. To avoid this situation, the monotone transformation ζ is applied to $\widehat{p}_t$ before thresholding. For instance, one might consider $\zeta(s) = Cs$ or $\zeta(s) = C \log s$ for an appropriately chosen positive number C. Note that the choice of ζ changes the form of the surrogate loss function. Thus, care must be taken when choosing ζ to ensure that 1) it

approximates the original comparator loss as accurately as possible; 2) a reasonable number of feedback requests are made; and 3) numerical underflow issues are circumvented.

Proof: Appendix E. ∎

C. Arbitrary Feedback Times

When labels cannot be requested by the Forecaster, but are instead provided arbitrarily by the environment or end user, we use the following algorithm to choose the threshold τ at each time t.

Algorithm 6 Anomaly detection with arbitrarily spaced feedback

Parameters: real number $\eta > 0$, $\tau_{\min} < \tau_{\max}$

Initialize: $\tau_1 = \tau_{\min}$

for $t = 1, 2, \ldots, T$ **do**

Receive the estimated likelihood $\widehat{p}_t$, set $\zeta_t = \zeta(\widehat{p}_t)$
if $\zeta_t < \tau_t$ **then** flag z_t as an anomaly: $\widehat{y}_t = 1$ **else** let $\widehat{y}_t = -1$
if feedback y_t is provided **then** let $\tau_{t+1} = \Pi\big(\tau_t + \eta y_t 1_{\{\widehat{y}_t \neq y_t\}}\big)$ **else** let $\tau_{t+1} = \tau_t$

end for

Under arbitrary feedback, it is meaningful to compare the performance of the Forecaster against a comparator τ only at those times when the feedback is provided. We then have the following performance bound.

Theorem 7: Fix a time horizon T and consider the anomaly detection with arbitrarily spaced feedback Forecaster run with parameter $\eta = 1/\sqrt{T}$. Let $t_1, \ldots, t_m$ denote the time steps at which the Forecaster receives feedback, and let $\epsilon \stackrel{\triangle}{=} m/T$. Then, for any $\tau \in [\tau_{\min}, \tau_{\max}]$, we have

$$\sum_{i=1}^{m} 1_{\{\widehat{y}_{t_i} \neq y_{t_i}\}} \leq \sum_{i=1}^{m} \ell_{t_i}(\tau) + \frac{(1+\epsilon)(\tau_{\max} - \tau_{\min})}{2}\sqrt{T}.$$

Proof: If we consider only the times $\{t_1, \ldots, t_m\}$, then we are exactly in the setting of Theorem 5. This observation leads to the bound

$$\sum_{i=1}^{m} 1_{\{\widehat{y}_{t_i} \neq y_{t_i}\}} \leq \sum_{i=1}^{m} \ell_{t_i}(\tau) + \frac{1}{2\eta}\left[(\tau_{\max} - \tau_{\min})^2 + \epsilon T \eta^2\right].$$

With the choice $\eta = (\tau_{\max} - \tau_{\min})/\sqrt{T}$, we get the bound in the theorem. ∎

D. Arbitrary Horizon

So far, we have considered the case when the horizon T is known in advance. However, it is not hard to modify the proofs of the results of this section to accommodate the case when the horizon is not set beforehand. The only change that is required is to replace the learning rate $\eta = O(1/\sqrt{T})$ with a time-varying sequence $\{\eta_t\}$, where $\eta_t = O(1/\sqrt{t})$. Then the regret bounds remain the same, namely, $O(\sqrt{T})$, possibly with different constants.

V. Experimental Results

In this section, we demonstrate how our proposed anomaly detection approach FHTAGN performs on simulated and real data. We consider four numerical experiments. Experiments 1, 2a, and 2b use simulated data. Experiment 1 tests the filtering component of FHTAGN. Experiments 2a and 2b focus on the hedging component of FHTAGN. Experiment 3 applies FHTAGN to the Enron email database [36].

Experiment 1: For this experiment, we generate simulated data by drawing from a temporally evolving Bernoulli product density. In particular, we first draw i.i.d. samples according to $x_t \sim \prod_{i=1}^{500} \text{Bernoulli}(\beta^*_{i,t})$, where $\beta^*_{i,t} \in [0,1]$ and $1 \leq t \leq 1000$. Our observations $z_t \in \{0,1\}^{500}$ are the noisy versions of x_t, where each bit $x_t(j)$ is passed through a BSC with crossover probability 0.1. The goal is to causally estimate $\{\beta^*_{i,t}\}$ from $\{z_t\}$. Let $\mu^*_t \stackrel{\triangle}{=} (\beta^*_{1,t}, \ldots, \beta^*_{500,t})$. We choose μ^*_t to be piecewise constant in time, with changes at $t = 100$, 500, and 700. In this setting, knowing μ^*_t allows us to compute empirical regret with respect to the known data generation parameters, and to compare it against the theoretical regret bound in Theorem 2.

We apply Algorithm 3 to the aforementioned data with the learning rate set to $\eta_t = 1/\sqrt{t}$. Fig. 1(a) illustrates the ground truth μ^*_t versus the estimated parameter $\widehat{\mu}_t$. Fig. 1(c)–(d) shows that the log loss exhibits pronounced spikes at the jump times, and then subsides as the Forecaster adapts to the new parameters. Note that the variance of the log loss is larger for the noisy case. Finally, Fig. 1(e) shows that the empirical per-round regret is well below the theoretical bound in Theorem 2, with the regret for the noisy case again slightly larger.

Experiment 2a: We now consider the detection of anomalies using Algorithm 5. For this experiment, data are generated using the same model described in Experiment 1. Recall that the goal is to detect anomalies corresponding to observations which would be difficult to predict given past data. In this spirit, we define the 25 observations following each changepoint in μ^*_t to be "true" anomalies, and our goal is to detect them with minimal probability of error. The basic idea is that after this window of 25 time steps, observations corresponding to the new generative model should be predictable based on past data and no longer anomalous.

In this experiment, the Forecaster queries the end user for feedback when $|\zeta(\widehat{p}_t) - \tau_t|$ is small. As discussed in Remark 6, we use the transformation $\zeta(s) = Cs$ with $C = \exp(220)$. Fig. 2(a) shows the result of this experiment. Here, the declared anomalies are shown as black dots superimposed on the log loss, and the feedback times are depicted underneath. This result is described in more detail and compared with the result of Experiment 2b.

Experiment 2b: This experiment is run on exactly the same data as in Experiment 2a; the only difference is the feedback mechanism. In particular, we test Algorithm 6, which is designed for the case when feedback is provided at arbitrary times. In this simulation, feedback is always provided when the Forecaster declares an event anomalous and with 20% probability if it misses an anomaly.

Fig. 2 shows the results of both experiments. For both cases note that after the first jump (when feedback is first received)

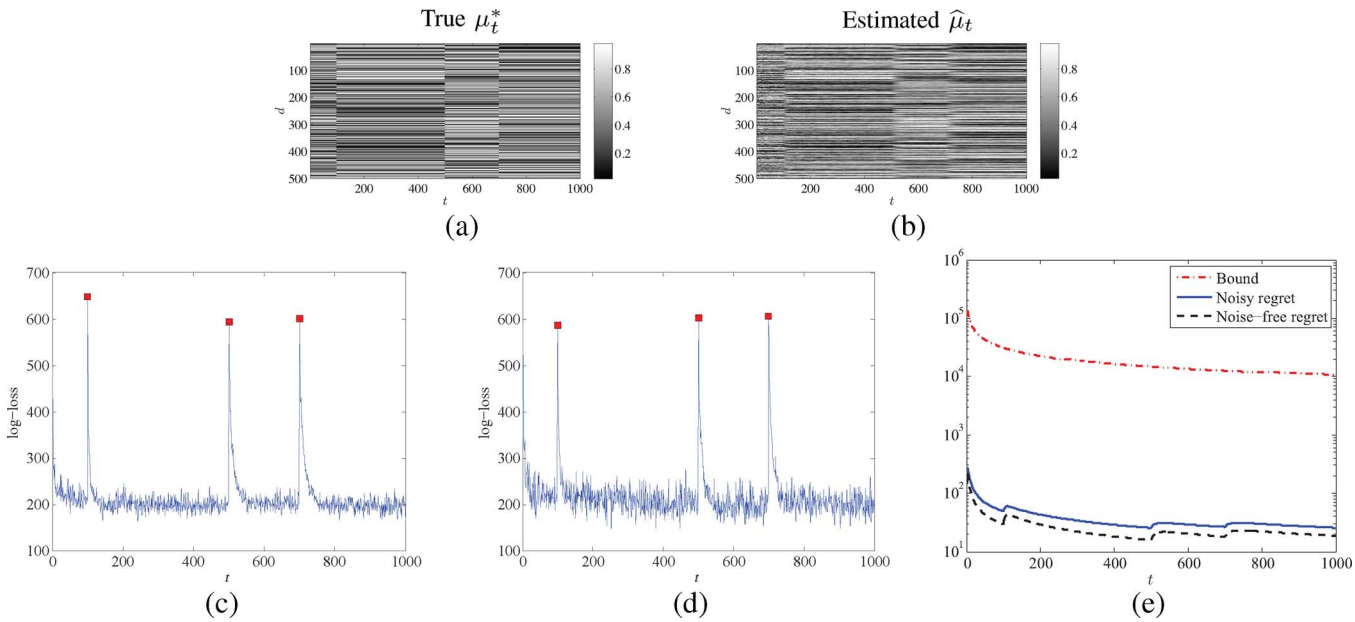

Fig. 1. (a) Ground truth and (b) estimated $\widehat{\mu}_t$. The μ values correspond to Bernoulli means, where lighter colors depict higher probabilities. (c) Evolution of the log loss from noiseless observations. (d) Evolution of the log loss from noisy observations. The spikes at the jump times ($t = 100$, 500, and 700, indicated by red squares) correspond to model changes. Note that the variance of the log loss is larger for the noisy case. (e) Per-round regret compared to theoretical upper bound. Again, the regret is larger for the noisy case.

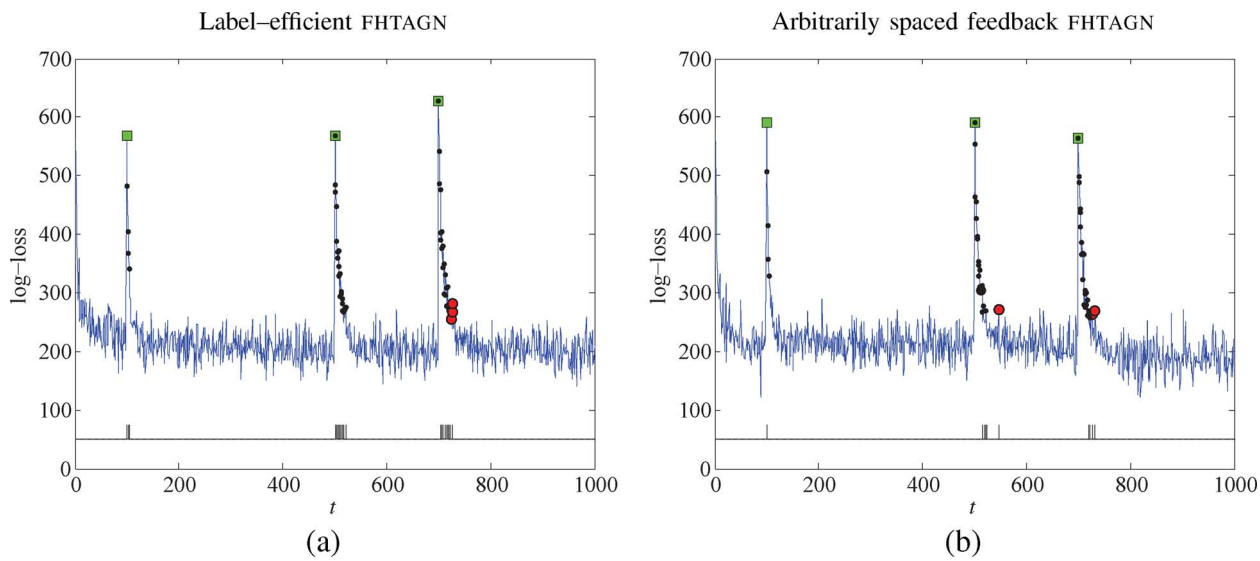

Fig. 2. (a) Anomalies detected (shown as small black dots) and false alarms (large red dots) by Algorithm 5, superimposed on the log loss. Forecaster query times are shown below the log loss. Jump times ($t = 100$, 500, and 700) are indicated by large green squares. (b) Similar plot of anomalies detected by Algorithm 6, with arbitrarily spaced feedback. In both cases, there were 25 true anomalies immediately following each jump. After the first jump, the Forecaster adapted its threshold, enabling it to dramatically increase its detection rate in subsequent jumps.

the Forecaster adapts its threshold, allowing it to dramatically increase its detection rate in subsequent jumps. Specifically, Table II shows the number of detection misses and false alarms for both Algorithms 5 and 6. For comparison, we also show the same performance measures for the best static threshold chosen in hindsight with full knowledge of the anomalies (i.e., $\{y_t\}$). In both experiments, FHTAGN outperforms static thresholding. The number of false alarms is significantly lower than that of detection misses due to the high initial value of the threshold τ_t, which is driven lower as feedback arrives.

Experiment 3: Algorithm 5 was applied to the Enron email database [36], which consists of approximately 500 000 emails involving 151 known employees and more than 75 000 distinct addresses between years 1998 and 2002. We use email time-

TABLE II
PERFORMANCE COMPARISON OF FHTAGN WITH ANOMALY DETECTION USING THE BEST STATIC THRESHOLD FOR EXPERIMENTS 2A AND 2B. FHTAGN COMMITS SIGNIFICANTLY FEWER ERRORS. FOR THE ENRON EXPERIMENT, FEEDBACK WAS REQUESTED FOR 91 OF THE 902 DAYS CONSIDERED, AND ONLY 523 OF THE 902 DAYS HAD THEIR TEXT PARSED

	Label-efficient		*Arbitrary feedback times*	
	FHTAGN	Best static threshold	FHTAGN	Best static threshold
Total Errors	30	44	34	46
False alarms	3	8	3	9
Detection misses	27	36	31	37

stamps in order to record users that were active in each day, either sending or receiving emails. This was done for 1177

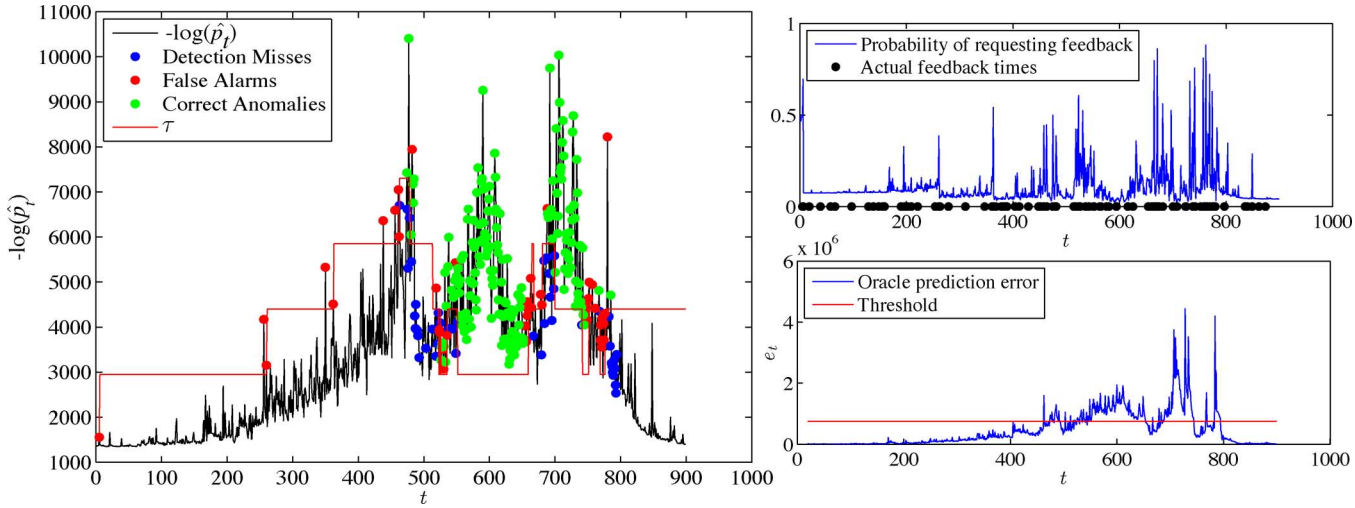

Fig. 3. Online anomaly detection results on Enron corpus. Left plot displays filtering output, locations of missed anomalies (as declared by our oracle), false positives, and correctly identified anomalies, as well as time-varying threshold τ. Upper right plot displays the probability of requesting feedback where black circles indicate the locations where feedback was provided. Lower right plot displays oracle prediction error e_t (from contextual evidence within a sliding window) compared to a static threshold to assign y_t.

days, starting from January 1, 1999. We removed days during which no email correspondence occurred, and we consolidated each weekend's emails into the preceding Friday's observation vector, resulting in a total of 902 days in our dataset. For each day t, $x_t \in \{0,1\}^n$ is a binary vector indicating who sent or received email that day. In this setting, we let $\phi(x) = x$ leading to a dimensionality of $n = d = 75511$. (There is no noise in this case, since we can accurately identify sender and recipient email addresses.)

Algorithms 3 and 5 were applied to this dataset. The goal was to causally determine an accurate predictive model of email sender and recipient co-occurrences, and to identify anomalous periods of email activity using feedback. For the hedging component, we use $\zeta(s) = C \log s$ with $C = 0.0079$, $\eta = 1450$, and $\tau_1 = \widehat{p}_1$. Feedback was received when requested according to Algorithm 5.

The central idea here is that anomalous email discussion topics correspond to anomalous email sender and recipient co-occurrences. In this spirit, we generate oracle or expert feedback (i.e., the $\{y_t\}$) based on the email *text* from the difference in word counts between day t and each of the previous 10 days, and average the result. Upon a feedback request at time t, we generate word count vectors h_t using the 12 000 most frequently appearing words (to avoid memory issues and misspelled words) for days $t-10, \ldots, t-1, t$. Specifically, the mean wordcount deviation e_t is computed as

$$e_t = \frac{1}{10} \sum_{i=t-10}^{t-1} \|h_t - h_i\|_1$$

where $\|\cdot\|_1$ is the ℓ_1 norm. This can be considered a crude measure of temporal variation in text documents. When the deviation e_t is sufficiently high, we consider day t to be anomalous according to our expert system (i.e., $y_t = 1$). The deviation metric e_t and the threshold determining y_t are shown in the right upper plot in Fig. 3, but note that only a fraction of these values need to be computed to run FHTAGN whenever feedback is requested.

TABLE III
PERFORMANCE COMPARISON FOR FHTAGN AND THE BEST STATIC THRESHOLD ON ENRON DATA SET. FEEDBACK WAS REQUESTED FOR 91 OF THE 902 DAYS CONSIDERED, AND ONLY 523 OF THE 902 DAYS HAD THEIR TEXT PARSED

	FHTAGN	Best static threshold
Total Errors	73	143
False alarms	35	96
Detection misses	38	47

The results of Algorithm 5 are summarized in Table III and Fig. 3. FHTAGN performs very well (in terms of detecting anomalies corresponding to the expert system designation with low probability of error) relative to a comparator online anomaly detection method which consists of comparing $\widehat{p}_t$ to the best static threshold, chosen in hindsight with full knowledge of all filtering outputs and feedback. The left plot in Fig. 3 shows the time-varying threshold τ_t in response to user feedback. In this experiment, ground truth anomalies do not always correspond to large values of $-\log \widehat{p}_t$ but rather to the degree to which contextual evidence differs from recent history. With a 10-day memory, the notion of what constitutes an anomaly is constantly evolving, and τ_t adjusts to reflect the pattern. The lower right plot describes the probability of requesting feedback over time with the days on which feedback was requested indicated with black dots. This plot suggests that feedback is more likely to be requested on days where $\zeta(\widehat{p}_t)$ and τ_t have similar magnitude, which is expected. Feedback was requested 91 out of 902 days, and because of the sliding window used by our oracle to determine the true labels y_t, a total of 523 of the 902 days required text parsing (and, generally speaking, any overhead associated with decrypting, transcribing, or translating documents).

TABLE IV
SIGNIFICANT DATES IN ENRON'S HISTORY AND OUR ANALYSIS

Date	Significance
Dec. 1, 2000	Days before "California faces unprecedented energy alert" (Dec. 7) and energy commodity trading deregulated in Congress. (Dec. 15) [37].
May 9, 2001	"California Utility Says Prices of Gas Were Inflated" by Enron collaborator El Paso [38], blackouts affect upwards of 167,000 Enron customers [39].
Oct. 18, 2001	Enron reports $618M third quarter loss, followed by later major correction [40].

Some of the most anomalous events detected by our proposed approach correspond to historical events, as summarized in Table IV. These examples indicate that the anomalies in social network communications detected by FHTAGN are indicative of anomalous events of interest to the social network members.

VI. CONCLUSION

We have proposed and analyzed a methodology for sequential (or online) anomaly detection from an individual sequence of potentially noisy observations in the setting when the anomaly detection engine can receive external feedback confirming or disputing the engine's inference on whether or not the current observation is anomalous relative to the past. Our methodology, dubbed FHTAGN for Filtering and Hedging for Time-varying Anomaly recoGNition, is based on the filtering of noisy observations to estimate the belief about the next clean observation, followed by a threshold test. The threshold is dynamically adjusted, whenever feedback is received and the engine has made an error, which constitutes the hedging step. Our analysis of the performance of FHTAGN was carried out in the individual sequence framework, where no assumptions were made on the mechanism underlying the evolving observations. Thus, performance was measured in terms of *regret* against the best *offline* (nonsequential) method for assigning beliefs to the entire sequence of *clean* observations and then using these beliefs and the feedback (whenever available) to set the best critical threshold. The design and analysis of both filtering and hedging were inspired by recent developments in OCP.

One major drawback of the proposed filtering step is the need to compute the log partition function. While closed-form expressions are available for many frequently used models (such as Gaussian MRFs), computing log partitions of general pairwise MRFs is intractable [24]. While there exist a variety of techniques for approximate computation of log partition functions, such as the log determinant relaxation [41], [42], these techniques themselves involve solving convex programs. This may not be an issue in the offline (batch) setting; however, in a sequential setting, computations may have to be performed in real time. Therefore, an important direction for future research is to find ways to avoid computing (or approximating) log partition functions in the filtering step, perhaps by replacing the full likelihood with an appropriate "pseudolikelihood" [43], [44].

APPENDIX

A) Proof of Theorem 1: For each t, let us use the shorthand $\widehat{\ell}_t(\theta)$ to denote the filtering loss $\widehat{\ell}(\theta, z_t)$, $\theta \in \Lambda$. We start by observing that for any $\theta, \theta' \in \Theta$ we have[4]

$$\begin{aligned}
&\widehat{\ell}_t(\theta) - \left[\widehat{\ell}_t(\theta') + \left\langle \nabla\widehat{\ell}_t(\theta'), \theta - \theta' \right\rangle\right] \\
&= -\langle \theta, h(z_t)\rangle + \Phi(\theta) - \big[- \langle \theta', h(z_t)\rangle + \Phi(\theta') \\
&\qquad - \langle h(z_t), \theta - \theta'\rangle + \langle \nabla\Phi(\theta'), \theta - \theta'\rangle\big] \\
&= \underbrace{\langle \theta' - \theta, h(z_t)\rangle + \langle \theta - \theta', h(z_t)\rangle}_{=0} \\
&\qquad + \Phi(\theta) - \Phi(\theta') - \langle \nabla\Phi(\theta'), \theta - \theta'\rangle \\
&\equiv D(\theta' \| \theta). \qquad (25)
\end{aligned}$$

In particular, using (25) with $\theta' = \widehat{\theta}_t$, we can write

$$\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) = -\langle \nabla\widehat{\ell}_t(\widehat{\theta}_t), \theta - \theta_t\rangle - D(\widehat{\theta}_t \| \theta). \qquad (26)$$

Now, by our hypothesis on Λ, the Legendre potential Φ is strongly convex w.r.t. the Euclidean norm $\|\cdot\|$ with constant $\alpha = 2H$. Indeed, for any $\theta, \theta' \in \Lambda$ we have

$$\begin{aligned}
&\Phi(\theta) - \Phi(\theta') - \langle \nabla\Phi(\theta'), \theta - \theta'\rangle \\
&\quad = \frac{1}{2}\langle \theta - \theta', \nabla^2\Phi(\theta'')(\theta - \theta')\rangle \\
&\quad \geq H\|\theta - \theta'\|^2
\end{aligned}$$

where in the second step θ'' is some point on the line segment joining θ and θ', and the last step follows from the fact that $\nabla^2\Phi(\theta'') \succeq 2HI_{d\times d}$ for any $\theta'' \in \Lambda \subset \Theta_H$. Thus, we can apply Lemma 1 to(26) to get

$$\begin{aligned}
&\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) \\
&= -\langle \nabla\widehat{\ell}_t(\widehat{\theta}_t), \theta - \theta_t\rangle - D(\widehat{\theta}_t \| \theta) \\
&\leq \frac{1}{\eta_t}\Big(D(\widehat{\theta}_t \| \theta) - D(\widehat{\theta}_{t+1} \| \theta)\Big) + \frac{\eta_t}{4H}\|\nabla\widehat{\ell}_t(\widehat{\theta}_t)\|^2 - D(\widehat{\theta}_t \| \theta). \qquad (27)
\end{aligned}$$

Now, if we define

$$\Delta_t \stackrel{\triangle}{=} \begin{cases} 0, & t = 1 \\ \frac{1}{\eta_{t-1}} D(\widehat{\theta}_t \| \theta), & t \geq 2 \end{cases}$$

then we can rewrite (27) as

$$\begin{aligned}
\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) &\leq \Delta_t - \Delta_{t+1} + \frac{\eta_t}{4H}\|\nabla\widehat{\ell}_t(\widehat{\theta}_t)\|^2 \\
&\qquad - D(\widehat{\theta}_t \| \theta) + \frac{1}{\eta_t} D(\widehat{\theta}_t \| \theta) - \Delta_t \\
&= \Delta_t - \Delta_{t+1} + \frac{\eta_t}{4H}\|\nabla\widehat{\ell}_t(\widehat{\theta}_t)\|^2
\end{aligned}$$

[4]In the terminology of [13], (25) means that the function $\theta \mapsto \widehat{\ell}_t(\theta)$ is *strongly convex* w.r.t. the Bregman divergence $D_\Phi(\theta, \theta') \equiv D(\theta' \| \theta)$ with constant 1. In fact, their condition for strong convexity holds here with equality.

where in the last step we have used the fact that with $\eta_t = 1/t$, $\frac{1}{\eta_t} D(\widehat{\theta}_t \| \theta) - \Delta_t = D(\widehat{\theta}_t \| \theta)$ for all t. Moreover, because

$$\|\nabla \widehat{\ell}_t(\widehat{\theta}_t)\| \leq \|h(z_t)\| + \|\nabla \Phi(\widehat{\theta}_t)\| \leq 2(K(z^T) + M)$$

we get

$$\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) \leq \Delta_t - \Delta_{t+1} + \frac{(K(z^T)+M)^2 \eta_t}{H}.$$

Summing from $t = 1$ to $t = T$, we obtain

$$\begin{aligned}
&\sum_{t=1}^T \widehat{\ell}(\widehat{\theta}_t, z_t) - \sum_{t=1}^T \widehat{\ell}(\theta, z_t) \\
&\quad \leq \sum_{t=1}^T (\Delta_t - \Delta_{t+1}) + \frac{(K(z^T)+M)^2}{H} \sum_{t=1}^T \eta_t \\
&\quad = \Delta_1 - \Delta_{T+1} + \frac{K+L)^2}{H} \sum_{t=1}^T \frac{1}{t} \\
&\quad \leq \frac{(K(z^T)+M)^2}{H} \log(T+1)
\end{aligned}$$

where in the last line we have used the estimate $\sum_{t=1}^T t^{-1} \leq 1 + \int_1^T t^{-1} dt = \log T + 1$.

B) Proof of Theorem 2: The main idea of the proof is similar to that in Theorem 1, except that now care must be taken in dealing with the time-varying comparison strategy $\theta = \{\theta_t\}$. Using (27) with $\theta = \theta_t$ and the fact that $D(\cdot\|\cdot) \geq 0$, we can write

$$\begin{aligned}
\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta_t) \leq \frac{1}{\eta_t} \Big(D(\widehat{\theta}_t \| \theta_t) - D(\widehat{\theta}_{t+1} \| \theta_t) \Big) \\
+ \frac{\eta_t}{4H} \|\nabla \widehat{\ell}_t(\widehat{\theta}_t)\|^2 - D(\widehat{\theta}_t \| \theta_t). \quad (28)
\end{aligned}$$

Let us define

$$\Delta'_t \stackrel{\Delta}{=} \begin{cases} 0, & t = 1 \\ \frac{1}{\eta_{t-1}} D(\widehat{\theta}_t \| \theta_t), & t \geq 2 \end{cases}$$

and $\Gamma_t \stackrel{\Delta}{=} D(\widehat{\theta}_{t+1} \| \theta_{t+1}) - D(\widehat{\theta}_{t+1} \| \theta_t)$. Then we can rewrite (28) as

$$\begin{aligned}
\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta_t) &\leq \Delta'_t - \Delta'_{t+1} + \frac{1}{\eta_t} \Gamma_t + \frac{\eta_t}{4H} \|\nabla \widehat{\ell}_t(\widehat{\theta}_t)\|^2 \\
&\quad - D(\widehat{\theta}_t \| \theta_t) + \frac{1}{\eta_t} D(\widehat{\theta}_t \| \theta_t) - \Delta'_t \\
&\leq \Delta'_t - \Delta'_{t+1} + \frac{1}{\eta_t} \Gamma_t + \frac{\eta_t}{4H} \|\nabla \widehat{\ell}_t(\widehat{\theta}_t)\|^2
\end{aligned}$$

where the last step uses the easily checked fact that, with the choice $\eta_t = 1/\sqrt{t}$,

$$\frac{1}{\eta_t} D(\widehat{\theta}_t \| \theta_t) - \Delta'_t = (\sqrt{t} - \sqrt{t-1}) D(\widehat{\theta}_t \| \theta_t) \leq D(\widehat{\theta}_t \| \theta_t).$$

Next, we have

$$\begin{aligned}
\Gamma_t &= \Phi(\theta_{t+1}) - \Phi(\widehat{\theta}_{t+1}) - \Big\langle \nabla \Phi(\widehat{\theta}_{t+1}), \theta_{t+1} - \widehat{\theta}_{t+1} \Big\rangle \\
&\quad - \Phi(\theta_t) + \Phi(\widehat{\theta}_{t+1}) + \Big\langle \nabla \Phi(\widehat{\theta}_{t+1}), \theta_t - \widehat{\theta}_{t+1} \Big\rangle \\
&= \Phi(\theta_{t+1}) - \Phi(\theta_t) + \Big\langle \nabla \Phi(\widehat{\theta}_{t+1}), \theta_t - \theta_{t+1} \Big\rangle \\
&\leq 4M \|\theta_t - \theta_{t+1}\|.
\end{aligned}$$

Moreover, just as in the proof of Theorem 1, we have

$$\|\nabla \widehat{\ell}_t(\widehat{\theta}_t)\|^2 \leq 4(K(z^T) + M)^2.$$

Combining everything and summing from $t = 1$ to $t = T$, we obtain

$$\begin{aligned}
&\sum_{t=1}^T \widehat{\ell}(\widehat{\theta}_t, z_t) - \sum_{t=1}^T \widehat{\ell}(\theta_t, z_t) \\
&\quad \leq \sum_{t=1}^T \left(\Delta'_t - \Delta'_{t+1}\right) + 4M \sum_{t=1}^T \frac{1}{\eta_t} \|\theta_t - \theta_{t+1}\| \\
&\qquad + \frac{(K(z^T)+M)^2}{H} \sum_{t=1}^T \eta_t \\
&\quad \leq \frac{4M}{\eta_T} V_T(\theta) + \frac{(K(z^T)+M)^2}{H} \sum_{t=1}^T \eta_t \\
&\quad \leq 4M\sqrt{T} V_T(\theta) + \frac{(K(z^T)+M)^2}{H} (2\sqrt{T} - 1).
\end{aligned}$$

In the last line, we have used the estimate $\sum_{t=1}^T t^{-1/2} \leq 1 + \int_1^T t^{-1/2} dt = 2\sqrt{T} - 1$.

C) Proof of Lemma 2: For each t, we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{\theta,t+1}(x^{t+1}, z^{t+1}) | \mathcal{R}_t\right] \\
&\quad = \mathbb{E}\left[\sum_{s=1}^{t+1} \langle \theta_s, h(z_s) - \phi(x_s) \rangle \,\middle|\, \mathcal{R}_t \right] \\
&\quad = \mathbb{E}\left[\langle \theta_{t+1}, h(z_{t+1}) - \phi(x_{t+1}) \rangle | \mathcal{R}_t\right] \\
&\qquad + \mathbb{E}\left[\sum_{s=1}^{t} \langle \theta_s, h(z_s) - \phi(x_s) \rangle \,\middle|\, \mathcal{R}_t \right] \\
&\quad = \langle \theta_{t+1}, \mathbb{E}[h(z_{t+1}) | \mathcal{R}_t] - \phi(x_{t+1}) \rangle \\
&\qquad + \sum_{s=1}^{t} \langle \theta_s, h(z_s) - \phi(x_s) \rangle \\
&\quad = 0 + \Delta_{\theta,t}(x^t, z^t)
\end{aligned}$$

where in the third step we used the fact that θ_{t+1}, $\{\theta_s\}_{s \leq t}$, and $\{z_s\}_{s \leq t}$ are $\mathcal{R}_t$-measurable, and in the last step we used the fact that $\mathbb{E}[h(z_{t+1}) | \mathcal{R}_t] = \phi(x_{t+1})$. Thus, $\{\Delta_{\theta,t}(x^t, z^t), \mathcal{R}_t\}_{t \geq 0}$, with $\Delta_{\theta,0}(x^0, z^0) \equiv 0$ and $\mathcal{R}_0$ the trivial σ-algebra, is a zero-mean martingale, and the desired result follows.

D) Proof of Theorem 5: We closely follow the proof of Theorem 12.1 in [16], except that we use Lemma 1 to highlight the role of MD and to streamline the argument. Let $\ell'_t(\tau)$ denote the subgradient of $\tau \mapsto \ell_t(\tau)$. Note that when $\ell_t(\tau) > 0$, $\ell'_t(\tau) = -y_t$. Thus, when $\widehat{y}_t \neq y_t$, the Forecaster implements the projected subgradient update $\tau_{t+1} = \Pi\big(\tau_t - \eta \ell'_t(\tau_t)\big)$. Thus, whenever $\widehat{y}_t \neq y_t$, we may use Lemma 1 :

$$\begin{aligned}
\ell'_t(\tau)(\tau_t - \tau) &= (\tau - \tau_t) y_t \\
&\leq \frac{1}{2\eta} \Big((\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 \Big) + \frac{\eta}{2} |\ell'_t(\tau_t)|^2 \\
&= \frac{1}{2\eta} \Big((\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 \Big) + \frac{\eta}{2} \quad (29)
\end{aligned}$$

where the inequalities hold for every τ. Now, at any step at which $\widehat{y}_t \neq y_t$, i.e., $\mathrm{sgn}(\tau_t - \zeta_t) \neq y_t$, the hinge loss $\ell_t(\tau) = (1 - (\tau - \zeta_t)y_t)_+$ obeys the bound

$$\begin{aligned} 1 - \ell_t(\tau) &= 1 - (1 - (\tau - \zeta_t)y_t)_+ \\ &\leq (\tau - \zeta_t)y_t \\ &= -(\tau - \zeta_t)\ell_t'(\tau_t). \end{aligned} \tag{30}$$

Therefore, when $\widehat{y}_t \neq y_t$, we have

$$\begin{aligned} 1 - \ell_t(\tau) &\leq (\tau - \tau_t)y_t + \underbrace{(\tau_t - \zeta_t)y_t}_{<0} \\ &\leq \frac{1}{2\eta}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2\right] + \frac{\eta}{2} \end{aligned} \tag{31}$$

where we used the fact that $\widehat{y}_t \neq y_t$ implies that $\mathrm{sgn}(\tau_t - \zeta_t) \neq y_t$, so that $(\tau_t - \zeta_t)y_t < 0$. Note also that when $\widehat{y}_t = y_t$, we will have $\widetilde{\tau}_{t+1} = \tau_t$, and since $\tau_t \in [\tau_{\min}, \tau_{\max}]$, $\tau_{t+1} = \Pi(\widetilde{\tau}_{t+1}) = \tau_t$. Thus, the very last expression in (31) is identically zero when $\widehat{y}_t = y_t$. Hence, we get the bound

$$(1 - \ell_t(\tau))1_{\{\widehat{y}_t \neq y_t\}} \leq \frac{1}{2\eta}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2\right] + \frac{\eta}{2}$$

that holds for all t. Summing from $t = 1$ to $t = T$ and rearranging, we get

$$\begin{aligned} \sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} &\leq \sum_{t=1}^{T} \ell_t(\tau) + \frac{1}{2\eta}(\tau - \tau_1)^2 + \frac{T\eta}{2} \\ &\leq \sum_{t=1}^{T} \ell_t(\tau) + \frac{(\tau_{\max} - \tau_{\min})^2}{2\eta} + \frac{T\eta}{2}. \end{aligned}$$

Choosing $\eta = (\tau_{\max} - \tau_{\min})/\sqrt{T}$, we obtain the regret bound (24).

E) Proof of Theorem 6: We closely follow the proof of Theorem 12.5 in [16], but, again, we use Lemma 1 to streamline and simplify the argument. Introduce the random variables $M_t = 1_{\{\widehat{y}_t \neq y_t\}}$. Then, whenever $M_t = 1$, we have $1 - \ell_t(\tau) \leq (\tau - \zeta_t)y_t$. When $M_t U_t = 1$ (i.e., when τ_t is updated to τ_{t+1}), we can use Lemma 1 and obtain

$$\begin{aligned} 1 - \ell_t(\tau) &\leq (\tau_t - \zeta_t)y_t + (\tau - \tau_t)y_t \\ &\leq (\tau_t - \zeta_t)y_t + \frac{1}{2\eta}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2\right] + \frac{\eta}{2}. \end{aligned}$$

From this, we obtain the inequality

$$\begin{aligned} &(1 + |\zeta_t - \tau_t|)M_t U_t \\ &\quad \leq \ell_t(\tau) + \frac{1}{2\eta}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2\right] + \frac{\eta}{2} \end{aligned}$$

which holds for all t. Indeed, if $M_t U_t = 0$, the left-hand side is zero, while the right-hand side is greater than zero since $\ell_t(\tau) \geq 0$ and $\tau_t = \tau_{t+1}$. If $M_t U_t = 1$, then $y_t(\tau_t - \zeta_t) = -(\tau_t - \zeta_t)\,\mathrm{sgn}(\tau_t - \zeta_t) = -|\zeta_t - \tau_t|$. Summing over t, we get

$$\begin{aligned} \sum_{t=1}^{T}(1 + |\zeta_t - \tau_t|)M_t U_t &\leq \sum_{t=1}^{T} \ell_t(\tau) \\ &\quad + \frac{1}{2\eta}\sum_{t=1}^{T}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2\right] + \frac{T\eta}{2}. \end{aligned}$$

We now take expectation of both sides. Let $\mathcal{U}_t$ denote the σ-algebra generated by $U_1, \ldots, U_t$, and let $\mathbb{E}_t[\cdot]$ denote the conditional expectation $\mathbb{E}[\cdot|\mathcal{U}_{t-1}]$. Note that M_t and $|\zeta_t - \tau_t|$ are measurable w.r.t. $\mathcal{R}_{t-1}$, since both of them depend on $U_1, \ldots, U_{t-1}$, and that $\mathbb{E}_t U_t = 1/(1 + |\zeta_t - \tau_t|)$. Hence

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T}(1 + |\zeta_t - \tau_t|)M_t U_t\right] &= \mathbb{E}\left[\sum_{t=1}^{T}(1 + |\zeta_t - \tau_t|)M_t \mathbb{E}_t U_t\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} M_t\right]. \end{aligned}$$

Using the same argument as before with $\eta = (\tau_{\max} - \tau_{\min})/\sqrt{T}$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq Y_t\}}\right] \leq \sum_{t=1}^{T} \ell_t(\tau) + (\tau_{\max} - \tau_{\min})\sqrt{T}$$

and the theorem is proved.

ACKNOWLEDGMENT

The authors would like to thank S. Rakhlin for helpful discussions, and two anonymous referees for their suggestions and comments that helped improve the presentation.

REFERENCES

[1] H. P. Lovecraft, "The call of Cthulhu," *Weird Tales*, vol. 11, no. 2, pp. 159–178, Feb. 1928.
[2] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*. New York: Springer-Verlag, 2009.
[3] C. R. Shalizi, "Dynamics of Bayesian updating with dependent data and mispecified models," *Electron. J. Statist.*, vol. 3, pp. 1039–1074, 2009.
[4] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[5] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 928–936.
[6] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley, 1983.
[7] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Oper. Res. Lett.*, vol. 31, pp. 167–175, 2003.
[8] M. Raginsky, R. Marcia, J. Silva, and R. Willett, "Sequential probability assignment via online convex programming using exponential families," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 1338–1342.
[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection—A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
[10] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, 2005.
[11] C. Scott and G. Blanchard, D. van Dyk and M. Welling, Eds., "Novelty detection: Unlabeled data definitely help," in *Proc. 12th Int. Conf. Artif. Intell. Stat.*, 2009, pp. 464–471.
[12] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Ann. Statist.*, vol. 25, no. 3, pp. 948–969, 1997.

[13] P. Bartlett, E. Hazan, and A. Rakhlin, "Adaptive online gradient descent," in *Advance in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2008, vol. 20, pp. 65–72.
[14] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari, "Optimal strategies and minimax lower bounds for online convex games," in *Proc. Int. Conf. Learn. Theory*, 2008, pp. 415–423.
[15] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[16] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. New York: Cambridge Univ. Press, 2006.
[17] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.
[18] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Berlin, Germany: Springer-Verlag, 2001.
[19] L. M. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *Comput. Math. Math. Phys.*, vol. 7, pp. 200–217, 1967.
[20] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*. Oxford, U.K.: Oxford Univ. Press, 1997.
[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[22] M. Raginsky, A. Rakhlin, and S. Yüksel, "Online convex programming and regularization in adaptive control," in *IEEE Conf. Decis. Control*, Atlanta, GA, Dec. 2010, pp. 1957–1962.
[23] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence: American Mathematical Society, 2000.
[24] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, 2008.
[25] 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, 1991.
[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[27] K. S. Azoury and M. K. Warmuth, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Mach. Learn.*, vol. 43, pp. 211–246, 2001.
[28] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2151–2173, Sep. 2001.
[29] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[30] M. Herbster and M. K. Warmuth, "Tracking the best expert," *Mach. Learn.*, vol. 32, no. 2, pp. 151–178, 1998.
[31] M. Herbster and M. K. Warmuth, "Tracking the best linear predictor," *J. Mach. Learn. Res.*, vol. 1, pp. 281–309, 2001.
[32] F. M. J. Willems, "Coding for a binary independent piecewise-identically-distributed source," *IEEE Trans. Inf. Theory*, vol. 42, no. 11, pp. 2210–2217, Nov. 1996.
[33] G. I. Shamir and N. Merhav, "Low-complexity sequential lossless coding for piecewise-stationary memoryless sources," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 1498–1519, Jul. 1999.
[34] N. Merhav, "On the minimum description length principle with piecewise constant parameters," *IEEE Trans. Inf. Theory*, vol. 39, no. 11, pp. 1962–1967, Nov. 1993.
[35] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Minimizing regret with label efficient prediction," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2152–2162, Jun. 2005.
[36] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proc. Eur. Conf. Mach. Learn.*, 2004, pp. 217–226.
[37] WGBH educational foundation, 2010 [Online]. Available: http://www.pbs.org/wgbh/pages/frontline/shows/blackout/california/timeline.html
[38] R. A. Oppel Jr. and L. Bergman, "California utility says prices of gas were inflated," *New York Times*, May 9, 2001.
[39] cnn.com, 2001 [Online]. Available: http://archives.cnn.com/2001/US/05/08/calif.power.crisis.02/
[40] J. Earle, "Analysts vent anger at 'hidden' Enron charge," *Financial Times*, Oct. 18, 2001.
[41] M. J. Wainwright and M. I. Jordan, "Log-determinant relaxation for approximate inference in discrete Markov random fields," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2099–2109, Jun. 2006.
[42] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.
[43] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Statist.*, vol. 38, no. 3, pp. 1287–1319, 2010.
[44] H. Höfling and R. Tibshirani, "Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods," *J. Mach. Learn. Res.*, vol. 10, pp. 883–906, 2009.

Maxim Raginsky (S'99–M'00) received the B.S. and M.S. degrees in 2000 and the Ph.D. degree in 2002 from Northwestern University, Evanston, IL, all in electrical engineering. He has held research positions with Northwestern, the University of Illinois at Urbana-Champaign (where he was a Beckman Foundation Fellow from 2004 to 2007), and Duke University. In 2012, he has returned to UIUC, where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory. His research interests lie at the intersection of information theory, machine learning, and control.

Rebecca M. Willett (S'01–M'05–SM'11) received the Ph.D. degree in electrical and computer engineering from Rice University, Houston, TX, in 2005. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. She has also held visiting researcher positions with the Institute for Pure and Applied Mathematics, University of California, Los Angeles, in 2004, the University of Wisconsin-Madison, from 2003 to 2005, the French National Institute for Research in Computer Science and Control (INRIA), Paris, France, in 2003, and the Applied Science Research and Development Laboratory, GE Healthcare, in 2002. Her research interests include network and imaging science with applications in medical imaging, wireless sensor networks, astronomy, and social networks. Prof. Willett is a member of the Defense Advanced Research Projects Agency Computer Science Study Group. She was the recipient of the National Science Foundation CAREER Award in 2007 and the Air Force Office of Scientific Research Young Investigator Program Award in 2010.

Corinne Horn received the B.S.E. degree in electrical and computer engineering from Duke University, Durham, NC, in 2011. She is currently a graduate student in the Department of Electrical Engineering, Stanford University, Stanford, CA.

Jorge Silva (M'00) received his EE, M.Sc., and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico (IST), Lisbon, Portugal, in 1993, 1999, and 2007, respectively. He was a Researcher at Instituto de Engenharia de Sistemas e Computadores (INESC) in 1993–1996, and at the Instituto de Sistemas e Robótica (ISR), Lisbon, in 2003–2007. He held teaching positions at Instituto Superior de Engenharia de Lisboa (ISEL) in 1996–2007. In the same period, he did consulting and R&D work for major Portuguese utility and transportation companies. He is now a Research Scientist, Sr., at Duke University, where he is developing estimation methods for very high-dimensional spaces. His research interests include signal processing, manifold learning, computer vision and social network analysis.

Roummel F. Marcia (M'08) received the Ph.D. degree in mathematics from the University of California, San Diego, in 2002. He is currently an Assistant Professor of applied mathematics with the School of Natural Sciences, University of California, Merced. He was a Computation and Informatics in Biology and Medicine Postdoctoral Fellow with the Department of Biochemistry, University of Wisconsin-Madison, and a Research Scientist with the Department of Electrical and Computer Engineering, Duke University. His research interests include nonlinear optimization, numerical linear algebra, signal and image processing, and computational biology.