Sequential anomaly detection in the presence of noise and limited feedback

Maxim Raginsky and Rebecca Willett*
Department of Electrical and Computer Engineering
Duke University, Durham NC, 27708

August 13, 2009

Abstract

Recent work has established the efficacy of using online convex programming methods on exponential families in the context of sequential probability assignment. This paper describes methods which build upon that framework to handle noisy observations. Furthermore, the problem of detecting anomalous (i.e. rare) events by using the sequential probability assignments and limited feedback is presented.

Keywords: Filtering, universal prediction, individual sequences, anomaly detection

1 Filtering and anomaly detection from noisy individual sequences

This paper is concerned with the problem of detecting anomalies in a series of sequentially arriving observations. 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. There are several key challenges that we consider in this report:

- The observations cannot be assumed to be independent, identically distributed, or even 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.
- Observations may be contaminated by noise or be observed through an imperfect communication channel.
- 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.

The method described in this paper is designed to address all these challenges effectively. In particular, we adopt an "individual sequence" perspective in which, instead of modeling the data generation process and designing methods for that process, we rather design methods which perform provably well for any individual sequence in the problem domain. 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.

In addition, we present a "label-efficient" anomaly detection methodology, in which the forecaster receives limited feedback on the correctness of its declarations of anomalous or non-anomalous (i.e. nominal). Access to such limited feedback is a natural assumption in many settings. For instance, imagine that the proposed

*This work was supported by NSF CAREER Award No. CCF-06-43947, NSF Award No. DMS-08-11062, and DARPA Grant No. HR0011-07-1-003.

anomaly detection method is used as a first-pass analysis of the received data, and that anything declared anomalous is more carefully analyzed by experts. In that case, the expert naturally determines the veracity of the anomaly label, and the expert feedback can easily be provided to the anomaly detection algorithm when it becomes available. Alternatively, consider a setting where the anomaly detection algorithm may request limited feedback when it will be the most informative to the anomaly detection engine. Specifically, if the feedback is being used to alter the anomalousness threshold, then observations with belief levels near the threshold should occasionally be tested further to ensure the threshold is at a meaningful level. In these settings, both of which are analyzed in this paper, feedback on the anomaly labels is only provided for a small subset of the observations, and only when it is either (a) convenient for an observer or (b) most useful for the anomaly detection engine.

This paper builds upon our previous work [1] on online convex programming over exponential families of probability distributions. In addition to tightening the bounds presented in that work, we extend the results to more general settings in which data may be noisy and build upon them for our anomaly detection framework.

For a motivating example, we consider the problem of monitoring interactions within a social network of p people. Each time a group of people in the social network meet, we record who was in attendance. This observation can be recorded as a binary string of length p, with 1's representing people at the meeting and 0's representing other people. Observations in this setting can easily be corrupted, as we may not notice some people in attendance, or we might mistake one person for another. Ultimately, using these imperfect observations of meeting patterns, and allowing that these patterns continuously change and evolve, we wish to detect unusual or anomalous meetings to be as helpful as possible to any human observer.

1.1 Problem formulation

This paper focuses on the problems of predicting elements of an arbitrary sequence $\boldsymbol{x} = x_1, x_2, \ldots$ over some set $\mathcal{X}$ and flagging elements in the sequence which are rare or anomalous. Moreover, rather than observing each x_t directly, we observe a noisy version, denoted $z_t \in \mathcal{Z}$. At each time $t = 1, 2, \ldots$, before z_t is revealed, we first assign a probability density p_t to the possible values of the underlying x_t. After making our prediction, we observe z_t and use this to evaluate a loss function. Ideally, we would like to evaluate and minimize the *logarithmic loss* $-\log p_t(x_t)$, but this loss cannot be computed directly since we do not have access to x_t. One of the challenges we face, therefore, is the selection of a loss function on z_t and an associated *prediction strategy* (i.e. sequence of probability assignments $\boldsymbol{p} = \{p_t\}_{t=1}^{\infty}$) which will help ensure that the logarithmic loss is small.

After making the probability assignment at time t, we want to flag any observation z_t for which $p_t(z_t) \leq \tau$ for some critical level τ. The key challenge here is to determine the level τ which will be most useful for the end user – that is, the level at which flagged events are also considered anomalous from the end user's perspective. However, without feedback from the end user about whether the detected anomalies are useful or whether important anomalies are missed, setting this threshold and potentially having it adapt to a changing environment is not possible. To address this challenge, we describe a novel method for data-adaptively choosing a threshold at each time t based on limited feedback from an observer.

2 Online convex programming and the method of mirror descent

The philosophy advocated in the present 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* (OCP) [2–4]. An OCP problem is specified by a convex feasible set $\mathcal{U} \subseteq \mathbb{R}^d$ and a family of convex functions $\mathcal{F} = \{f : \mathcal{U} \to \mathbb{R}\}$, and is described as follows:

1: The Forecaster picks an arbitrary initial point $\widehat{u}_1 \in \mathcal{U}$
2: **for** $t = 1, 2, ...$ **do**
3: The Environment picks a convex function $f_t \in \mathcal{F}$
4: The Forecaster observes f_t and incurs the cost $f_t(\widehat{u}_t)$
5: The Forecaster picks a new point $\widehat{u}_{t+1} \in \mathcal{U}$
6: **end for**

The total cost incurred by the Forecaster after T rounds is given by $\sum_{t=1}^{T} f_t(\widehat{u}_t)$. We will refer to any sequence $\boldsymbol{u} = \{u_t\}_{t=1}^{\infty}$ over $\mathcal{U}$ as a *strategy*. Of particular interest is the strategy $\widehat{\boldsymbol{u}} = \{\widehat{u}_t\}_{t=1}^{\infty}$ composed of the points selected by the Forecaster in the course of the game. Observe that $\widehat{\boldsymbol{u}}$ obeys the following *causality constraint*:

$$\widehat{u}_t = \widehat{u}_t(\widehat{u}_1, f_1, \ldots, \widehat{u}_{t-1}, f_{t-1}), \qquad t = 1, 2, \ldots.$$

that is, $\widehat{u}_t$ depends only on past instances of f and $\widehat{u}$. Given the cost function sequence $\boldsymbol{f} = \{f_t\}_{t=1}^{\infty}$ and any other (not necessarily causal) strategy $\boldsymbol{u} = \{u_t\}_{t=1}^{\infty}$, we can define the *regret* of the Forecaster w.r.t. $\widehat{\boldsymbol{u}}$ by

$$R_T(\boldsymbol{u}) \triangleq \sum_{t=1}^{T} f_t(\widehat{u}_t) - \sum_{t=1}^{T} f_t(u_t). \tag{2.1}$$

(Strictly speaking, the regret depends not only on the comparator strategy $\boldsymbol{u}$, but also on the proposed strategy $\widehat{\boldsymbol{u}}$ and on the function sequence $\boldsymbol{f}$, but we will not write this out in the interest of brevity.) Given a particular comparison class $\mathcal{C}$ of strategies, the objective of the Forecaster is to ensure that

$$R_T(\mathcal{C}) \triangleq \sup_{\boldsymbol{f}} \sup_{\boldsymbol{u} \in \mathcal{C}} R_T(\boldsymbol{u}) = o(T). \tag{2.2}$$

A strategy $\widehat{\boldsymbol{u}}$ that achieves (2.2) over a comparison class $\mathcal{C}$ is said to be *Hannan-consistent* w.r.t. $\mathcal{F}$ and $\mathcal{C}$ (see the text of Cesa-Bianchi and Lugosi [5] for a thorough discussion of Hannan consistency).

One important comparison class is composed of all *static* strategies, i.e., all strategies $\boldsymbol{u} = \{u_t\}$ for which $u_1 = u_2 = \ldots$. The class of all static strategies is in one-to-one correspondence with the feasible set $\mathcal{U}$, and we can define the regret relative to $u \in \mathcal{U}$

$$R_T(u) \triangleq \sum_{t=1}^{T} f_t(\widehat{u}_t) - \sum_{t=1}^{T} f_t(u)$$

and the worst-case regret

$$R_T(\mathcal{U}) \triangleq \sup_{\boldsymbol{f}} \sup_{u \in \mathcal{U}} R_T(u).$$

Note that we can write

$$\sup_{u \in \mathcal{U}} R_T(u) = \sum_{t=1}^{T} f_t(\widehat{u}_t) - \inf_{u \in \mathcal{U}} \sum_{t=1}^{T} f_t(u).$$

In other words, the worst-case regret $\sup_u R_T(u)$ is the difference between the actual cost incurred by the Forecaster after T rounds of the game and the smallest cumulative cost the Forecaster could achieve *in hindsight* (with full knowledge of the first T cost functions selected by the Environment) using a single feasible point.

2.1 The mirror descent procedure

A generic procedure for constructing OCP strategies is inspired by the so-called *method of mirror descent*, which was originally introduced by Nemirovski and Yudin [6] in the context of mathematical programming over Banach spaces (we refer the reader to the papers by Beck and Teboulle [7] and Nemirovski et al. [8]

for simple expositions). The main benefit of mirror descent methods is that they incorporate more general measures of "proximity" between pairs of points in the feasible set $\mathcal{U}$ than the Euclidean norm on $\mathbb{R}^d$. For instance, as we will see below, when the set $\mathcal{U}$ is the canonical parameter space of an exponential family of probability densities, the KL divergence between two such densities is a more natural measure of proximity than the Euclidean distance between the canonical parameter vectors. These general measures of proximity are given by the so-called *Bregman divergences* [9, 10]:

Definition 2.1 *Let $\mathcal{U} \subseteq \mathbb{R}^d$ be a nonempty set with convex interior (which we denote by $\operatorname{Int}\mathcal{U}$). A function $F : \mathcal{U} \to \mathbb{R}$ is called* Legendre *if it is:*

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

Definition 2.2 *The* Bregman divergence *induced by a Legendre function $F : \mathcal{U} \to \mathbb{R}$ is the nonnegative function $D_F : \mathcal{U} \times \operatorname{Int}\mathcal{U} \to \mathbb{R}$, given by*

$$D_F(u, v) \triangleq F(u) - F(v) - \langle \nabla F(v), u - v \rangle, \qquad \forall u \in \mathcal{U}, v \in \operatorname{Int}\mathcal{U}. \tag{2.3}$$

For example, if $\mathcal{U} = \mathbb{R}^d$ and $F(u) = (1/2)\|u\|^2$, where $\|\cdot\|$ is the Euclidean norm, then $D_F(u, v) = (1/2)\|u - v\|^2$. In general, for a fixed $v \in \operatorname{Int}\mathcal{U}$, $D_F(\cdot, v)$ gives the tail of the first-order Taylor expansion of $F(\cdot)$ around v. We will use the following properties of Bregman divergences (for proofs, see [5]):

Lemma 2.1 (Properties of Bregman divergences) *The Bregman divergence D_F enjoys the following properties:*

1. *For all $u \in \mathcal{U}$ and all $v, w \in \operatorname{Int}\mathcal{U}$, we have*

$$D_F(u, v) + D_F(v, w) = D_F(u, w) + \langle \nabla F(w) - \nabla F(v), u - v \rangle. \tag{2.4}$$

2. *Let $\mathcal{S} \subseteq \mathcal{U}$ be a closed convex set. The* Bregman projection *of $w \in \operatorname{Int}\mathcal{U}$ onto $\mathcal{S}$ is defined by*

$$\Pi_{F,\mathcal{S}}(w) \triangleq \operatorname*{arg\,min}_{u \in \mathcal{S}} D_F(u, w).$$

For any $w \in \operatorname{Int}\mathcal{U}$, $\Pi_{F,\mathcal{S}}(w)$ exists, is unique, and satisfies the following generalized Pythagorean inequality*:*

$$D_F(u, w) \geq D_F(u, \Pi_{F,\mathcal{S}}(w)) + D_F(\Pi_{F,\mathcal{S}}(w), w), \qquad \forall u \in \mathcal{S}, w \in \operatorname{Int}\mathcal{U}. \tag{2.5}$$

We now present the general mirror descent scheme. We allow the possibility of restricting the feasible points to a closed, convex subset $\mathcal{S}$ of $\operatorname{Int}\mathcal{U}$. The mirror descent updates make use of F^*, the *Legendre–Fenchel dual* of F [11], which is given by

$$F^*(z) \triangleq \sup_{u \in \mathcal{U}} \left\{ \langle u, z \rangle - F(u) \right\}.$$

Since F is steep and strictly convex, F^* is finite everywhere and differentiable. Let $\mathcal{U}^*$ denote the image of $\operatorname{Int}\mathcal{U}$ under the gradient mapping ∇F:

$$\mathcal{U}^* \triangleq \left\{ w \in \mathbb{R}^d : w = \nabla F(u) \text{ for some } u \in \operatorname{Int}\mathcal{U} \right\}.$$

From the fact that F is Legendre, it can be shown that the function $F^* : \mathcal{U}^* \to \mathbb{R}$ is also Legendre, and that the gradient mappings ∇F and ∇F^* are inverses of one another:

$$\nabla F^*(\nabla F(u)) = u \qquad \text{and} \qquad \nabla F(\nabla F^*(w)) = w, \qquad \forall u \in \operatorname{Int}\mathcal{U}, w \in \operatorname{Int}\mathcal{U}^*.$$

Following the terminology of [5], we will refer to the points in $\mathrm{Int}\,\mathcal{U}$ as the *primal points* and to their images under ∇F as the *dual points*. In the context of mirror descent schemes, the Legendre function F is referred to as the *potential function*.

The mirror descent strategy for OCP, presented below as Algorithm 1, assumes that, for every f_t, we can compute a subgradient $g_t(u)$ of f_t at any $u \in \mathrm{Int}\,\mathcal{U}$ (recall that a subgradient of a convex function $f : \mathcal{U} \to \mathbb{R}$ at a point $u \in \mathrm{Int}\,\mathcal{U}$ is any vector $g \in \mathbb{R}^d$, such that

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

holds for all $v \in \mathrm{Int}\,\mathcal{U}$). 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.

Algorithm 1 A Generic Mirror Descent Strategy for OCP

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

1: The Forecaster picks an arbitrary initial point $\widehat{u}_1 \in \mathcal{S}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Observe the cost function $f_t \in \mathcal{F}$
4: Incur the cost $f_t(\widehat{u}_t)$
5: Compute $\xi_t = \nabla F(\widehat{u}_t)$
6: Dual update: $\widetilde{\xi}_{t+1} = \xi_t - \eta_t g_t(\widehat{u}_t)$
7: Projected primal update: compute $\widetilde{u}_{t+1} = \nabla F^*(\widetilde{\xi}_{t+1})$ and

$$\widehat{u}_{t+1} = \Pi_{F,\mathcal{S}}(\widetilde{u}_{t+1}) \equiv \operatorname*{arg\,min}_{u \in \mathcal{S}} D_F(u, \widetilde{u}_{t+1})$$

8: **end for**

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

$$\begin{aligned} \widetilde{u}_{t+1} &= \widehat{u}_t - \eta g_t(\widehat{u}_t) \\ \widehat{u}_{t+1} &= \operatorname*{arg\,min}_{u \in \mathcal{S}} \|u - \widetilde{u}_{t+1}\| . \end{aligned}$$

It can be also shown that, for a general Legendre potential F, the mirror descent strategy is equivalent to solving, at each time t, the optimization problem

$$\widehat{u}_{t+1} = \operatorname*{arg\,min}_{u \in \mathcal{S}} \Big[\eta_t \big\langle g_t(\widehat{u}_t), u \big\rangle + D_F(u, \widehat{u}_t) \Big]$$

(see, e.g., [7] or [5, Chap. XI]). The following lemma is a key ingredient in bounding the regret of the mirror descent strategy:

Lemma 2.2 *For any $u \in \mathcal{S}$ and any t, we have the bound*

$$f_t(\widehat{u}_t) \leq f_t(u) + \frac{1}{\eta_t} \Big(D_F(u, \widehat{u}_t) - D_F(u, \widehat{u}_{t+1}) + D_F(\widehat{u}_t, \widetilde{u}_{t+1}) \Big). \tag{2.6}$$

Proof: By the convexity of f_t, we can write

$$f_t(\widehat{u}_t) \leq f_t(u) - \big\langle g_t(\widehat{u}_t), u - \widehat{u}_t \big\rangle. \tag{2.7}$$

Now, using the definition of the dual update and the fact that $\widetilde{\xi}_{t+1} = \nabla F(\widetilde{u}_t)$, we have

$$g_t(\widehat{u}_t) = \frac{1}{\eta_t}(\xi_t - \widetilde{\xi}_{t+1}) = \frac{1}{\eta_t}(\nabla F(\widehat{u}_t) - \nabla F(\widetilde{u}_{t+1})).$$

Substituting this into (2.7) and using (2.4), we obtain

$$\begin{aligned} f_t(\widehat{u}_t) &\le f_t(u) + \frac{1}{\eta_t}\Big\langle \nabla F(\widetilde{u}_{t+1}) - \nabla F(\widehat{u}_t), u - \widehat{u}_t \Big\rangle \\ &= f_t(u) + \frac{1}{\eta_t}\Big(D_F(u,\widehat{u}_t) + D_F(\widehat{u}_t,\widetilde{u}_{t+1}) - D_F(u,\widetilde{u}_{t+1}) \Big). \end{aligned} \tag{2.8}$$

Now, since $u \in \mathcal{S}$ and $\widehat{u}_{t+1} = \Pi_{F,\mathcal{S}}(\widetilde{u}_{t+1})$, we can use the generalized Pythagorean inequality (2.5) to obtain

$$D_F(u,\widetilde{u}_{t+1}) \ge D_F(u,\widehat{u}_{t+1}) + D_F(\widehat{u}_{t+1},\widetilde{u}_{t+1}) \ge D_F(u,\widehat{u}_{t+1})$$

(the second inequality is due to the fact that $D_F(\cdot,\cdot)$ is always nonnegative). Using this in (2.8), we get (2.6). ■

As we will see shortly, this basic bound, combined with additional assumptions on the functions in $\mathcal{F}$, will allow us to construct Hannan–consistent OCP strategies.

2.2 Illustration for constant step sizes

As a preliminary illustration, consider the case when the step sizes are constant, $\eta_t = \eta$ for all t. Then, for any static strategy $u \in \mathcal{S}$ we have, by Lemma 2.2,

$$\begin{aligned} R_T(u) &= \sum_{t=1}^T f_t(\widehat{u}_t) - \sum_{t=1}^T f_t(u) \\ &\le \frac{1}{\eta}\sum_{t=1}^T \Big(D_F(u,\widehat{u}_t) - D_F(u,\widehat{u}_{t+1}) \Big) + \frac{1}{\eta}\sum_{t=1}^T D_F(\widehat{u}_t,\widetilde{u}_{t+1}) \\ &= \frac{1}{\eta}D_F(u,\widehat{u}_1) - \frac{1}{\eta}D_F(u,\widehat{u}_{T+1}) + \frac{1}{\eta}\sum_{t=1}^T D_F(\widehat{u}_t,\widetilde{u}_{t+1}) \\ &\le \frac{1}{\eta}D_F(u,\widehat{u}_1) + \frac{1}{\eta}\sum_{t=1}^T D_F(\widehat{u}_t,\widetilde{u}_{t+1}). \end{aligned}$$

The goal then is to arrange that the potential function F is *strongly convex* on $\mathcal{U}$ w.r.t. to some norm $\|\cdot\|$, that is

$$F(u) \ge F(v) + \langle \nabla F(v), u - v\rangle + \frac{1}{2}\|u - v\|^2, \qquad \forall u, v \in \operatorname{Int}\mathcal{U}.$$

From this it can be shown that

$$D_F(\widehat{u}_t,\widetilde{u}_{t+1}) \le \frac{\eta^2}{2}\,\|g_t(\widehat{u}_t)\|_*^2, \qquad \forall t$$

where $\|\cdot\|_*$ is the norm dual to $\|\cdot\|$ (i.e., $\|u\|_* \triangleq \sup_{\|v\|\le 1}\langle u, v\rangle$). Thus we obtain the regret bound

$$R_T(u) \le \frac{D}{\eta} + \frac{G\eta T}{2},$$

where G is an upper bound on the dual norms of the subgradients of $f \in \mathcal{F}$ at the points in $\mathcal{S}$ and $D \triangleq \max_{u,v\in\mathcal{S}} D_F(u,v)$. If the time horizon T is fixed, we can optimize over η to get $O(\sqrt{T})$ regret against any static strategy over $\mathcal{S}$.

3 Sequential probability assignment in the presence of noise

The problem of *sequential probability assignment* appears in such contexts as universal data compression, online learning, and sequential investment [5,12]. It is defined as follows. Elements of an arbitrary sequence $\boldsymbol{x} = x_1, x_2, \ldots$ over some set $\mathcal{X}$ are revealed to us one at a time. We make no assumptions on the structure of $\boldsymbol{x}$. At time $t = 1, 2, \ldots$, before x_t is revealed, we have to assign a probability density $\widehat{p}_t$ to the possible values of x_t. When x_t is revealed, we incur the *logarithmic loss* $-\log \widehat{p}_t(x_t)$. We refer to any such sequence of probability assignments $\widehat{\boldsymbol{p}} = \{\widehat{p}_t\}_{t=1}^{\infty}$ as a *prediction strategy*. Since the probability assignment $\widehat{p}_t$ is a function of the past observations $x^{t-1} \stackrel{\triangle}{=} (x_1, x_2, \ldots, x_{t-1}) \in \mathcal{X}^{t-1}$, we may view it as a conditional probability density $\widehat{p}_t(\cdot|x^{t-1})$. One way to view $\widehat{p}_t(x|x^{t-1})$, $x \in \mathcal{X}$, is as our *belief*, based on the past observations x^{t-1}, that the next observation x_t will be equal to x.

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 $\mathcal{C}$ [5, 12]. Thus, given a prediction strategy $\widehat{\boldsymbol{p}}$, let us define the *regret* of $\widehat{\boldsymbol{p}}$ w.r.t. some $\boldsymbol{p} = \{p_t\} \in \mathcal{C}$ after T time steps as

$$R_T(\boldsymbol{p}) \stackrel{\triangle}{=} \sum_{t=1}^{T} \log \frac{1}{\widehat{p}_t(x_t|x^{t-1})} - \sum_{t=1}^{T} \log \frac{1}{p_t(x_t|x^{t-1})}. \tag{3.9}$$

The goal is to design $\widehat{\boldsymbol{p}}$ to be Hannan-consistent, i.e.,

$$R_T(\mathcal{C}) \stackrel{\triangle}{=} \sup_{\boldsymbol{x}} \sup_{\boldsymbol{p} \in \mathcal{C}} R_T(\boldsymbol{p}) = o(T).$$

If we are interested in predicting only the first T elements of $\boldsymbol{x}$, we could consider approaches based on maximum likelihood estimation or mixture strategies. For example, a fundamental result due to Shtarkov [13] says that the *minimax regret* $R_T^*(\mathcal{C}) \stackrel{\triangle}{=} \inf_{\widehat{\boldsymbol{p}}} R_T(\mathcal{C})$, where the infimum is over all prediction strategies, is achieved by the normalized maximum likelihood estimator (MLE) over $\mathcal{C}$. The latter is computed from the joint density

$$p^*(x^T) \stackrel{\triangle}{=} \frac{\sup_{\boldsymbol{q} \in \mathcal{C}} \prod_{t=1}^{T} q_t(x_t|x^{t-1})}{\int_{\mathcal{X}^T} \sup_{\boldsymbol{q} \in \mathcal{C}} \prod_{t=1}^{T} q_t(z_t|z^{t-1}) d\nu^T(z^T)}$$

by letting

$$p_t^*(\cdot|x^{t-1}) = \frac{p^*(\cdot, x^{t-1})}{p^*(x^{t-1})}, \qquad t = 1, \ldots, T.$$

However, practical use of the normalized MLE strategy is limited since it requires solving an optimization problem over $\mathcal{C}$ whose complexity increases with T.

Mixture strategies provide a more easily computable alternative: if the reference class $\mathcal{C}$ is parametrized, $\mathcal{C} = \{\boldsymbol{p}_\theta : \theta \in \Theta\}$ with $\boldsymbol{p}_\theta = \{p_{\theta,t}\}_{t=1}^{\infty}$, then we can pick a prior probability measure w on Θ and consider a strategy induced by the joint densities

$$p(x^t) = \int_\Theta \prod_{s=1}^{t} p_{\theta,s}(x_s|x^{s-1}) dw(\theta)$$

via the posterior

$$p_t(\cdot|x^{t-1}) \stackrel{\triangle}{=} \frac{p(\cdot, x^{t-1})}{p(x^{t-1})}, \qquad t = 1, \ldots, T.$$

For instance, when the underlying observation space $\mathcal{X}$ is finite and the reference class $\mathcal{C}$ consists of all product distributions of the form $p(x^t) = \prod_{s=1}^{t} p_0(x_s)$, where p_0 is some probability mass function on $\mathcal{X}$, the well-known Krichevsky–Trofimov (KT) predictor [14]

$$p_t(a|x^{t-1}) = \frac{N(a|x^{t-1}) + 1/2}{(t-1) + |\mathcal{X}|/2}, \qquad a \in \mathcal{X}$$

where $N(a|x^{t-1})$ is the number of times a occurs in x^{t-1}, is a mixture strategy induced by a Dirichlet prior on the probability simplex over $\mathcal{X}$ [12]. It can be shown that the regret of the KT predictor is $O(|\mathcal{X}| \log T)$ [5,14].

The computational cost of updating the probability assignment using a mixture strategy is independent of T. However, as can be seen in the case of the KT predictor, the dependence of the regret on the cardinality of $\mathcal{X}$ still presents certain difficulties. For example, consider the case where $\mathcal{X} = \{0,1\}^d$ for some large positive integer d. If we wish to bring the per-round regret $T^{-1}R_T$ down to some given $\epsilon > 0$, we must have $T/\log T = \Omega(2^d/\epsilon)$. Moreover, when $\mathcal{X} = \{0,1\}^d$, the KT predictor will assign extremely small probabilities (on the order of $1/2^d$) to all as yet unseen binary strings $x \in \{0,1\}^d$. This is undesirable in settings where prior knowledge about the "smoothness" of the relative frequencies of $\boldsymbol{x}$ is available. Of course, if the dimensionality k of the underlying parameter space Θ is much lower than the cardinality of $\mathcal{X}$, mixture strategies lead to $O(k \log T)$ regret, which is minimax optimal [5]. This can be thought of as a generalization of the MDL-type regret bounds of Rissanen [12, 15] to the online, individual-sequence setting. However, the predictive distributions output by a mixture strategy will not, in general, lie in $\mathcal{C}$, which is often a reasonable requirement. In addition, implementation of a mixture strategy requires obtaining the posterior $p_t(\cdot|x^{t-1})$ (also known as the predictive distribution) at each time t, which can be computationally expensive.

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 $x_t \in \mathcal{X}$, we receive "noisy" symbols $z_t \in \mathcal{Z}$ (where $\mathcal{Z}$ 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 some 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 above, namely:

1. The prediction strategy is now of the form $\{\widehat{p}_t(\cdot|z^{t-1})\}_{t=1}^{\infty}$, where, at each time t, $\widehat{p}_t(x|z^{t-1})$ is our belief, given the past observations z^{t-1}, that x_t, the *clean* observation at time t, will be x.
2. We cannot observe the true incurred log-likelihood $-\log \widehat{p}_t(x_t|z^{t-1})$.

Problems of this kind are known in statistical signal processing under the name of *filtering*. In our case, we are interested in sequential prediction, via the beliefs $\widehat{p}_t(x_t|z^{t-1})$, of the next element of the clean sequence 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 $\mathcal{X}$, and the noisy observations $\{z_t\}_t$ are conditionally independent of one another given $\boldsymbol{x}$.

As we will show in the next section, if the comparison class $\mathcal{C}$ consists of product distributions lying in an *exponential family* parametrized by a real vector $\theta \in \mathbb{R}^d$, then we can cast the problem of sequential probability assignment from noisy data as an instance of OCP. Recall that a d-dimensional exponential family is characterized by the 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 $\mathbb{R}^d$, the function $\phi : \mathcal{X} \to \mathbb{R}^d$ is referred to as the *sufficient statistic*, and $\Phi(\theta)$ is the normalization constant known as the *log partition function* (see Section 4.1 for details). We will consider probability assignment strategies of the form

$$\widehat{p}_t(\cdot|z^{t-1}) = p_{\widehat{\theta}_t}(\cdot),$$

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

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

Thus, the regret relative to any comparison strategy $\boldsymbol{p}$ induced by a parameter sequence $\boldsymbol{\theta} = \{\theta_t\}$ via

$$p_t(x) = p_{\theta_t}(x), \qquad \forall x \in \mathcal{X}$$

can be written as

$$R_T(\boldsymbol{p}) = \sum_{t=1}^{T} \log \frac{1}{p_{\widehat{\theta}_t}(x_t)} - \sum_{t=1}^{T} \log \frac{1}{p_{\theta_t}(x_t)} = \sum_{t=1}^{T} \Big[\ell(\widehat{\theta}_t, x_t) - \ell(\theta_t, x_t) \Big],$$

where we have defined the function $\ell(\theta, x) \triangleq -\langle \theta, \phi(x) \rangle + \Phi(\theta)$. The resulting approach will have the following benefits:

1. The loss function $\ell(\theta, x)$ is convex in θ. Moreover, if there exists an unbiased estimator $h(z_t)$ of the sufficient statistic $\phi(x_t)$, the mirror descent procedure can use the estimated losses $\widehat{\ell}(\theta, z) = -\langle \theta, h(z) \rangle + \Phi(\theta)$.

2. The geometry of exponential families leads to a natural choice of the Legendre potential to be used in the mirror-descent updates, namely the log partition function Φ.

3. 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.

4 Probability assignment in an exponential family via OCP-based filtering

In this section, we detail our OCP filtering approach to sequential probability assignment in the presence of noise.

4.1 Background on exponential families

We begin by briefly recalling the basics of exponential families (see, e.g., [16, 17] and references therein). We assume that the observation space $\mathcal{X}$ is equipped with a σ-algebra $\mathcal{B}$ and a dominating σ-finite measure ν on $(\mathcal{X}, \mathcal{B})$. From now on, all densities will be defined w.r.t. ν. Given a positive integer d, let $\phi_k : \mathcal{X} \to \mathbb{R}$, $k = 1, \ldots, d$, be a given set of measurable functions. Define a vector-valued function $\phi : \mathcal{X} \to \mathbb{R}^d$ by $\phi(x) \triangleq \big(\phi_1(x), \ldots, \phi_d(x)\big)^T$ and the set

$$\Theta = \left\{ \theta \in \mathbb{R}^d : \Phi(\theta) \triangleq \log \int_{\mathcal{X}} e^{\langle \theta, \phi(x) \rangle} d\nu(x) < +\infty \right\},$$

where $\langle \theta, \phi(x) \rangle = \theta_1 \phi_1(x) + \ldots + \theta_d \phi_d(x)$. Then the *exponential family* $\mathcal{P}(\phi) = \{p_\theta : \theta \in \Theta\}$ consists of densities $p_\theta(x) \triangleq \exp\big(\langle \theta, \phi(x) \rangle - \Phi(\theta)\big)$. In the following, we will denote by $\mathbb{E}_\theta[\cdot]$ the expectation w.r.t. p_θ. The function Φ is the so-called *log partition function*. We state without proof the following facts:

1. The log partition function $\Phi(\theta)$ 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 $\phi(X) = (\phi_1(X), \ldots, \phi(X))$ when $X \sim p_\theta$. In particular,

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

   Thus, the Hessian $\nabla^2 \Phi(\theta)$, being a 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 mapping $\theta \mapsto \nabla \Phi(\theta)$ is invertible; the inverse mapping is $\mu \mapsto \nabla \Phi^*(\mu)$, where

$$\Phi^*(\mu) \triangleq \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - \Phi(\theta)\}$$

   is the Legendre–Fenchel dual of Φ. The gradient mapping $\nabla \Phi$ maps the *primal parameter* $\theta \in \Theta$ to the corresponding *dual parameter* $\mu \in \Theta^* \triangleq \nabla \Phi(\Theta)$.

5. The relative entropy (Kullback–Leibler divergence) between p_θ and $p_{\theta'}$ in $\mathcal{P}(\phi)$, defined as $D(p_\theta \| p_{\theta'}) = \int_{\mathcal{X}} p_\theta \log(p_\theta / p_{\theta'}) d\nu$, can be written as

$$D(p_\theta \| p_{\theta'}) = \Phi(\theta) - \Phi(\theta') - \langle \nabla\Phi(\theta'), \theta - \theta' \rangle \tag{4.10}$$

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

From these properties, it follows that $\Phi : \Theta \to \mathbb{R}$ is a Legendre function, and that the mapping $D_\Phi : \Theta \times \operatorname{Int}\Theta \to \mathbb{R}$, defined by $D_\Phi(\theta, \theta') = D(\theta \| \theta')$, is a Bregman divergence. Hence, the KL divergence $D(\theta \| \theta')$ enjoys all the properties listed in Lemma 2.1.

4.2 Sequential probability assignment with exponential families via OCP

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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ ranges over all infinite sequences over Λ, and each $\boldsymbol{p_\theta}$ is of the form

$$p_{t,\boldsymbol{\theta}}(x_t | x^{t-1}) = p_{\theta_t}(x_t), \qquad t = 1, 2, \ldots; x^t \in \mathcal{X}^t. \tag{4.11}$$

In other words, each prediction strategy in $\mathcal{C}$ is 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 a sequence $\boldsymbol{\theta} = \{\theta_t\}_t$ over Λ; the corresponding object $\boldsymbol{p_\theta}$ will be implied. Given a strategy $\boldsymbol{\theta}$ and a time horizon T, we define the *variation* of $\boldsymbol{\theta}$ from $t = 1$ to $t = T$ as

$$V_T(\boldsymbol{\theta}) \triangleq \sum_{t=1}^{T} \|\theta_t - \theta_{t+1}\|. \tag{4.12}$$

Consider a density p_θ in $\mathcal{P}(\phi)$. The negative log-likelihood

$$-\log p_\theta(x) = -\langle \theta, \phi(x) \rangle + \Phi(\theta)$$

is a convex function of $\theta \in \Theta$. Hence, we can define the loss function

$$\ell(\theta, x) \triangleq -\langle \theta, \phi(x) \rangle + \Phi(\theta).$$

Consider first the noiseless case: $z_t = x_t$ for all t. In a previous publication [1], we have analyzed the following mirror descent scheme that uses the log partition function as the Legendre potential:

Algorithm 2 Sequential Probability Assignment via OCP

1: Initialize with $\widehat{\theta}_1 \in \Lambda$
2: **for** $t = 1, 2, \ldots$ **do**
3: Acquire new observation x_t
4: Incur the cost $\ell_t(\widehat{\theta}_t) = -\langle \widehat{\theta}_t, \phi(x_t) \rangle + \Phi(\widehat{\theta}_t)$
5: Compute $\widehat{\mu}_t = \nabla\Phi(\widehat{\theta}_t)$
6: Dual update: compute $\widehat{\mu}'_{t+1} = \widehat{\mu}_t - \eta_t \nabla \ell_t(\widehat{\theta}_t)$
7: Projected primal update: compute $\widetilde{\theta}_{t+1} = \nabla\Phi^*(\widehat{\mu}'_{t+1})$ and

$$\widehat{\theta}_{t+1} = \Pi_{\Phi,\Lambda}(\widetilde{\theta}_{t+1}) \equiv \arg\min_{\theta \in \Lambda} D(\theta \| \widetilde{\theta}_{t+1})$$

8: **end for**

We will now extend this approach to the case of noisy observations. Consider the observation model

$$z = N(x, r),$$

where r is the random noise input and $N(\cdot,\cdot)$ is a known deterministic function. For example, if $x \in \{0,1\}^d$ and $\phi(x) = x$, we can consider $r \in \{0,1\}^d$, where each component $r(i)$ is a Bernoulli-p random variable independent of everything else ($p < 1/2$), and

$$z(i) = x(i) \oplus r(i), \qquad i = 1, \ldots, d. \tag{4.13}$$

Here, $\oplus$ denotes modulo 2 addition. 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.

We assume that there exists a function $h : \mathcal{Z} \to \mathbb{R}^d$, 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 $\phi(x)$. As an example, consider the BSC noise model (4.13). Then an unbiased estimator for $\phi(x) \equiv x$ would be given by $h(z) = (h_1(z), \ldots, h_d(z))$, where

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

Alternatively, consider the Ising model set-up, where $x \in \{0,1\}^m$ and $\phi_i(x) = x(i)$ for $i = 1, \ldots, m$ and $\phi_{kp+j}(x) = x(k)x(j)$ for $k, j = 1, \ldots, m$. In other words, $\phi(x) \in \{0,1\}^d$, where $d = m(m+1)/2$, captures all first- and second-order interactions among the elements of x. Then an unbiased estimator for $\phi_i(x)$ would be

$$h_i(z) = \frac{z(i) - p}{1 - 2p}$$

for $i = 1, \ldots, m$ and

$$h_{kd+j}(z) = \frac{z(k) - p}{1 - 2p} \frac{z(j) - p}{1 - 2p}$$

for $k, j = 1, \ldots, m$.

If we had access to the true x_t's, we would measure the loss at time t using $\ell(\widehat{\theta}_t, x_t)$. Since x_t is unavailable, however, we *estimate* $\ell(\widehat{\theta}_t, x_t)$ by the *filtering loss*

$$\widehat{\ell}(\widehat{\theta}_t, z_t) \triangleq \langle \widehat{\theta}_t, h(z_t) \rangle - \Phi(\theta_t).$$

This leads to the following prediction strategy:

Algorithm 3 Noisy Sequential Probability Assignment via OCP

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

1: Initialize with $\widehat{\theta}_1 \in \Lambda$
2: **for** $t = 1, 2, \ldots$ **do**
3: Acquire new noisy observation z_t
4: Compute the filtering loss $\widehat{\ell}_t(\widehat{\theta}_t) = -\langle \widehat{\theta}_t, h(z_t) \rangle + \Phi(\widehat{\theta}_t)$
5: Compute $\widehat{\mu}_t = \nabla\Phi(\widehat{\theta}_t)$
6: Dual update: compute $\widehat{\mu}'_{t+1} = \widehat{\mu}_t - \eta_t \nabla\widehat{\ell}_t(\widehat{\theta}_t)$
7: Projected primal update: compute $\widetilde{\theta}_{t+1} = \nabla\Phi^*(\widehat{\mu}'_{t+1})$ and

$$\widehat{\theta}_{t+1} = \Pi_{\Phi,\Lambda}(\widetilde{\theta}_{t+1}) \equiv \arg\min_{\theta \in \Lambda} D(\theta \| \widetilde{\theta}_{t+1})$$

8: **end for**

In other words (see the remark after Algorithm 1), at time t we choose the parameter $\widehat{\theta}_{t+1}$ and the corresponding distribution $\widehat{p}_{t+1}$ according to

$$\widehat{\theta}_{t+1} = \arg\min_{\theta \in \Lambda} \left[\left\langle \theta, \Phi(\widehat{\theta}_t) - h(z_t) \right\rangle + \frac{1}{\eta_t} D(\theta \| \widehat{\theta}_t) \right] \tag{4.14}$$

$$\widehat{p}_{t+1} = p_{\widehat{\theta}_{t+1}} \tag{4.15}$$

It should be pointed out that an algorithm similar to (4.14) and (4.15) was suggested by Azoury and Warmuth [18] for the problem of sequential probability assignment over an exponential family, but they did not consider noisy observations and only proved regret bounds for a couple of specific exponential families. One of the contributions of the present paper is to demonstrate that near-minimax regret bounds can be obtained for a *general* exponential family, subject to mild restrictions on the parameter space Θ.

In the following, we will 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 algorithm in (4.14) and (4.15) will be
$$R_T(\theta) = O(\log T).$$

2. If the comparison class $\mathcal{C}$ consists of all time-varying strategies $\boldsymbol{\theta} = \{\theta_t\}$ over Λ, then, under certain regularity conditions on Λ and with properly chosen step sizes $\{\eta_t\}$, the expected regret of the algorithm in (4.14) and (4.15) will be
$$R_T(\boldsymbol{\theta}) = O\left((V_T(\boldsymbol{\theta}) + 1)\sqrt{T}\right),$$
where $V_T(\boldsymbol{\theta})$ is defined in (4.12).

Moreover, in the absence of noise (i.e., $z_t = x_t$ for all t), the above regret bounds will hold for all observation sequences $\boldsymbol{x}$.

4.3 Regret bounds for OCP-based filter

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

$$L_{\boldsymbol{\theta},t}(x^t) \triangleq \sum_{s=1}^{t} \ell(\theta_s, x_s),$$
$$\widehat{L}_{\boldsymbol{\theta},t}(z^t) \triangleq \sum_{s=1}^{t} \widehat{\ell}(\theta_s, z_s),$$

and the difference

$$\Delta_{\boldsymbol{\theta},t}(x^t, z^t) \triangleq L_{\boldsymbol{\theta},t}(x^t) - \widehat{L}_{\boldsymbol{\theta},t}(z^t) = \sum_{s=1}^{t} \langle \theta_s, h(z_s) - \phi(x_s) \rangle.$$

When $\boldsymbol{\theta}$ 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 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 $\boldsymbol{z} = \{z_t\}_t$. In the second step, we will use a martingale argument along the lines of Weissman and Merhav [19] to show that the expected "true" regret is bounded by the expected filtering regret.

4.3.1 Regret bounds for the filtering loss

We will consider time-varying strategies of the form (4.11), where the set Λ is restricted in the following way. Given a positive constant $H > 0$, define the set

$$\Theta_H \triangleq \left\{\theta \in \Lambda : \nabla^2 \Phi(\theta) \succeq 2H I_{d\times d}\right\},$$

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

$$J(\theta) \triangleq -\mathbb{E}_\theta[\nabla_\theta^2 \log p_\theta(X)],$$

which is the Fisher information matrix at θ [16]. Our assumption on Λ thus stipulates that the eigenvalues of the Fisher information matrix are bounded from below by $2H$ over Λ.

We first establish a logarithmic regret bound against static strategies in Λ. The theorem below is an improved version of our result from [1]:

Theorem 4.1 (Logarithmic regret against static strategies) *Let Λ be any closed, convex subset of Θ_H, and let $\widehat{\boldsymbol{\theta}} = \{\widehat{\theta}_t\}$ be the sequence of parameters in Λ computed from the noisy sequence $\boldsymbol{z} = \{z_t\}_t$ using the OCP procedure shown in Algorithm 3 with step sizes $\eta_t = 1/t$. Then, for any $\theta \in \Lambda$, we have*

$$\widehat{L}_{\widehat{\boldsymbol{\theta}},T}(z^T) \leq \widehat{L}_{\theta,T}(z^T) + D(\theta\|\widehat{\theta}_1) + \frac{(K+L)^2}{H}(\log T + 1), \tag{4.16}$$

where

$$K \triangleq (1/2)\max_{1\leq t\leq T}\|h(z_t)\| \qquad \text{and} \qquad L \triangleq (1/2)\max_{\theta\in\Lambda}\|\nabla\Phi(\theta)\|.$$

Proof: 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

$$\begin{aligned}
&\widehat{\ell}_t(\theta) - \left[\widehat{\ell}_t(\theta') + \left\langle \nabla\widehat{\ell}_t(\theta'), \theta - \theta' \right\rangle\right] \\
&= -\langle\theta, h(z_t)\rangle + \Phi(\theta) - [-\langle\theta', h(z_t)\rangle + \Phi(\theta') - \langle h(z_t), \theta - \theta'\rangle + \langle\nabla\Phi(\theta'), \theta - \theta'\rangle] \\
&= \underbrace{\langle\theta' - \theta, h(z_t)\rangle + \langle\theta - \theta', h(z_t)\rangle}_{=0} + \Phi(\theta) - \Phi(\theta') - \langle\nabla\Phi(\theta'), \theta - \theta'\rangle \\
&\equiv D(\theta\|\theta').
\end{aligned} \tag{4.17}$$

In the terminology of [3], 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, the condition for strong convexity from [3] holds here with equality. Moreover, for any $\theta, \theta' \in \Lambda$ we will have

$$D(\theta\|\theta') = \Phi(\theta) - \Phi(\theta') - \left\langle\nabla\Phi(\theta'), \theta - \theta'\right\rangle = \frac{1}{2}\left\langle\theta - \theta', \nabla^2\Phi(\theta'')(\theta - \theta')\right\rangle \geq H\|\theta - \theta'\|^2, \tag{4.18}$$

where θ'' is some point on the line segment joining θ and θ'.

Now let us linearize $\widehat{\ell}_t(\theta)$ around $\widehat{\theta}_t$ by defining

$$\widetilde{\ell}_t(\theta) \triangleq \widehat{\ell}_t(\widehat{\theta}_t) + \left\langle\nabla\widehat{\ell}_t(\widehat{\theta}_t), \theta - \widehat{\theta}_t\right\rangle, \qquad \forall\theta \in \Lambda.$$

Note that $\widetilde{\ell}_t(\widehat{\theta}_t) = \widehat{\ell}_t(\widehat{\theta}_t)$ and $\widetilde{\ell}_t(\widehat{\theta}_t) - \widetilde{\ell}_t(\theta) = -\left\langle\nabla\widehat{\ell}_t(\widehat{\theta}_t), \theta - \widehat{\theta}_t\right\rangle$. Hence, substituting $\theta' = \widehat{\theta}_t$ into (4.17) and applying (4.18), we obtain

$$\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) = \widetilde{\ell}_t(\widehat{\theta}_t) - \widetilde{\ell}_t(\theta) - D(\theta\|\widehat{\theta}_t). \tag{4.19}$$

Next, we note that if we were to replace $\widehat{\ell}_t(\cdot)$ with $\widetilde{\ell}_t(\cdot)$ throughout Algorithm 3, then the updates in the algorithm would not change. Therefore, applying Lemma 2.2, we can write

$$\widetilde{\ell}_t(\widehat{\theta}_t) - \widetilde{\ell}_t(\theta) \leq \frac{1}{\eta_t}\left(D(\theta\|\widehat{\theta}_t) - D(\theta\|\widehat{\theta}_{t+1}) + D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1})\right) \tag{4.20}$$

Combining (4.19) and (4.20) and rearranging yields

$$\begin{aligned}
\widehat{\ell}_t(\widehat{\theta}_t) - \widehat{\ell}_t(\theta) &\leq \frac{1}{\eta_t}\left(D(\theta\|\widehat{\theta}_t) - D(\theta\|\widehat{\theta}_{t+1})\right) - D(\theta\|\widehat{\theta}_t) + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}) \\
&= \underbrace{\left(\frac{1}{\eta_t} - 1\right)}_{=1/\eta_{t-1}} D(\theta\|\widehat{\theta}_t) - \frac{1}{\eta_t}D(\theta\|\widehat{\theta}_{t+1}) + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}) \\
&= \Delta_t - \Delta_{t+1} + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}),
\end{aligned} \tag{4.21}$$

where in the second line we have used the definition $\eta_t = 1/t$, $t \geq 1$, and $1/\eta_0 \equiv 0$, and in the third line we have defined $\Delta_t \triangleq (1/\eta_{t-1}) D(\theta \| \widehat{\theta}_t)$. We now bound the third term on the RHS of (4.21). To that end, we use a small trick (see, e.g., Section 4 in [3]). First,

$$\begin{aligned} D(\widehat{\theta}_t \| \widetilde{\theta}_{t+1}) + D(\widetilde{\theta}_{t+1} \| \widehat{\theta}_t) &= \left\langle \nabla\Phi(\widetilde{\theta}_{t+1}) - \nabla\Phi(\widehat{\theta}_t), \widetilde{\theta}_{t+1} - \widehat{\theta}_t \right\rangle \\ &= \left\langle \widehat{\mu}'_{t+1} - \widehat{\mu}_t, \widetilde{\theta}_{t+1} - \widehat{\theta}_t \right\rangle \\ &= -\eta_t \left\langle \nabla \widehat{\ell}_t(\widehat{\theta}_t), \widetilde{\theta}_{t+1} - \widehat{\theta}_t \right\rangle \\ &\leq \frac{\eta_t^2}{4H} \left\| \nabla \widehat{\ell}_t(\widehat{\theta}_t) \right\|^2 + H \left\| \widetilde{\theta}_{t+1} - \widehat{\theta}_t \right\|^2, \end{aligned}$$

where the first line is due to (4.10), the second line uses the primal-dual relations $\widehat{\mu}'_{t+1} = \nabla\Phi(\widetilde{\theta}_{t+1})$ and $\widehat{\mu}_t = \nabla\Phi(\widehat{\theta}_t)$, the third line uses the definition of the dual update, and the final line is an application of the inequality

$$\langle u, v \rangle \leq \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2$$

that holds for all $u, v \in \mathbb{R}^d$ to the vectors $u = -\eta_t \nabla \widehat{\ell}_t(\widehat{\theta}_t)/\sqrt{2H}$ and $v = \sqrt{2H}(\widetilde{\theta}_{t+1} - \widehat{\theta}_t)$. The goal of this exercise is to be able to write

$$D(\widehat{\theta}_t \| \widetilde{\theta}_{t+1}) \leq \frac{\eta_t^2}{4H} \left\| \nabla \widehat{\ell}_t(\widehat{\theta}_t) \right\|^2 + \underbrace{H \left\| \widetilde{\theta}_{t+1} - \widehat{\theta}_t \right\|^2 - D(\widetilde{\theta}_{t+1} \| \widehat{\theta}_t)}_{\leq 0},$$

where we have used (4.18). Moreover,

$$\left\| \nabla \widehat{\ell}_t(\widehat{\theta}_t) \right\| \leq \|h(z_t)\| + \left\| \nabla\Phi(\widehat{\theta}_t) \right\| \leq 2(K + L).$$

Hence,

$$\frac{1}{\eta_t} D(\widehat{\theta}_t \| \widetilde{\theta}_{t+1}) \leq \frac{(K+L)^2 \eta_t}{H}. \tag{4.22}$$

Substituting (4.22) into (4.21) and summing from $t = 1$ to $t = T$, we obtain

$$\begin{aligned} \sum_{t=1}^T \widehat{\ell}(\widehat{\theta}_t, z_t) - \sum_{t=1}^T \widehat{\ell}(\theta, z_t) &\leq \sum_{t=1}^T (\Delta_t - \Delta_{t+1}) + \frac{(K+L)^2}{H} \sum_{t=1}^T \eta_t \\ &= \Delta_1 - \Delta_{T+1} + \frac{(K+L)^2}{H} \sum_{t=1}^T \frac{1}{t} \\ &\leq D(\theta \| \widehat{\theta}_1) + \frac{(K+L)^2}{H} \log(T+1), \end{aligned}$$

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

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

Theorem 4.2 (Regret against time-varying strategies) *Again, let Λ be any closed, convex subset of Θ_H. Let $\widehat{\boldsymbol{\theta}}$ be the sequence of parameters in Λ computed from the noisy sequence $\boldsymbol{z} = \{z_t\}_t$ using the OCP procedure shown in Algorithm 3 with step sizes $\eta_t = 1/\sqrt{t}$. Then, for any sequence $\boldsymbol{\theta} = \{\theta_t\}_t$ over Λ, we have*

$$\widehat{L}_{\widehat{\boldsymbol{\theta}},T}(z^T) \leq \widehat{L}_{\boldsymbol{\theta},T}(z^T) + D(\theta_1 \| \widehat{\theta}_1) + D\sqrt{T+1} + L\sqrt{T} V_T(\boldsymbol{\theta}) + \frac{(K+L)^2}{H}(2\sqrt{T} - 1), \tag{4.23}$$

where K and L are defined as in Theorem 4.1, $D \triangleq \max_{\theta,\theta' \in \Lambda} D(\theta \| \theta')$, and $V_T(\boldsymbol{\theta})$ is defined in (4.12).

Proof: Applying Lemma 2.2 to $\widehat{\ell}_t(\cdot)$, we write

$$\widehat{\ell}_t(\widehat{\theta}_t) \leq \widehat{\ell}_t(\theta_t) + \frac{1}{\eta_t}\left(D(\theta_t\|\widehat{\theta}_t) - D(\theta_t\|\widehat{\theta}_{t+1}) + D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1})\right). \tag{4.24}$$

Adding and subtracting $D(\theta_{t+1}\|\widehat{\theta}_{t+1})$ inside the parentheses and rearranging, we get

$$\begin{aligned}
\widehat{\ell}_t(\widehat{\theta}_t) &\leq \widehat{\ell}_t(\theta_t) + \frac{1}{\eta_t}\left(D(\theta_t\|\widehat{\theta}_t) - D(\theta_{t+1}\|\widehat{\theta}_{t+1}) + D(\theta_{t+1}\|\widehat{\theta}_{t+1}) - D(\theta_t\|\widehat{\theta}_{t+1}) + D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1})\right) \\
&= \Delta_t - \Delta_{t+1} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) D(\theta_{t+1}\|\widehat{\theta}_{t+1}) + \frac{1}{\eta_t}\Gamma_t + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}) \\
&\leq \Delta_t - \Delta_{t+1} + D\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \frac{1}{\eta_t}\Gamma_t + \frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}),
\end{aligned}$$

where we have defined $\Delta_t = (1/\eta_t)D(\theta_t\|\widehat{\theta}_t)$ and $\Gamma_t = D(\theta_{t+1}\|\widehat{\theta}_{t+1}) - D(\theta_t\|\widehat{\theta}_{t+1})$. Next, we have

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

Moreover, just as in the proof of Theorem 4.1, we have

$$\frac{1}{\eta_t}D(\widehat{\theta}_t\|\widetilde{\theta}_{t+1}) \leq \frac{(K+L)^2\eta_t}{H}.$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}\widehat{\ell}(\widehat{\theta}_t, z_t) - \sum_{t=1}^{T}\widehat{\ell}(\theta_t, z_t) \\
&\quad\leq \sum_{t=1}^{T}(\Delta_t - \Delta_{t+1}) + D\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + L\sum_{t=1}^{T}\frac{1}{\eta_t}\|\theta_t - \theta_{t+1}\| + \frac{(K+L)^2}{H}\sum_{t=1}^{T}\eta_t \\
&\quad\leq \Delta_1 - \Delta_{T+1} + D\left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1}\right) + \frac{L}{\eta_T}V_T(\boldsymbol{\theta}) + \frac{(K+L)^2}{H}\sum_{t=1}^{T}\eta_t \\
&\quad\leq D(\theta_1\|\widehat{\theta}_1) + D\sqrt{T+1} + L\sqrt{T}V_T(\boldsymbol{\theta}) + \frac{(K+L)^2}{H}(2\sqrt{T}-1).
\end{aligned}$$

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

4.3.2 Bounds on the expected true regret

We now proceed to establish regret bounds on $L_{\boldsymbol{\theta},T}(x^T)$. First, we need the following lemma, which is similar to Lemma 1 in [19]:

Lemma 4.1 *Let $\boldsymbol{r} = \{r_t\}_t$ 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 $\boldsymbol{\theta} = \{\theta_t\}_{t\geq 1}$ be a sequence of probability assignments, such that each $\theta_t = \theta_t(z^{t-1})$. Then, for any individual sequence $\boldsymbol{x} = \{x_t\}$, $\{\Delta_{\boldsymbol{\theta},t}(x^t, z^t), \mathcal{R}_t\}_t$ is a martingale, and so we have $\mathbb{E}\widehat{L}_{\boldsymbol{\theta},T}(z^T) = \mathbb{E}L_{\boldsymbol{\theta},T}(x^T)$ for each T.*

Proof: For each t, define

$$m_t \triangleq \sum_{s=1}^{t} \langle \theta_s, h(z_s) - \phi(x_s) \rangle,$$

so that

$$\Delta_{\boldsymbol{\theta},T}(x^T) = \sum_{t=1}^{T} m_t, \qquad T = 1, 2, \ldots.$$

We wish to prove that $\mathbb{E}[m_{t+1}|\mathcal{R}_t] = m_t$ for each t. To that end, we have

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

where in the third step we used the fact that θ_{t+1}, $\{\theta_s\}_{s\le t}$, and $\{z_s\}_{s\le t}$ 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, $\{\Delta_{\boldsymbol{\theta},t}(x^t, z^t), \mathcal{R}_t\}_{t\ge 0}$, with $\Delta_{\boldsymbol{\theta},0}(x^0, z^0) \equiv 0$ and $\mathcal{R}_0$ the trivial σ-algebra, is a zero-mean martingale, and the desired result follows. ■

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

Theorem 4.3 *Consider the setting of Theorem 4.1. Then we have*

$$\mathbb{E}L_{\widehat{\boldsymbol{\theta}},T}(x^T) \le \inf_{\theta\in\Lambda} \left[L_{\theta,T}(x^T) + D(\theta\|\widehat{\theta}_1)\right] + \frac{(K+L)^2}{H}(\log T + 1). \tag{4.25}$$

Likewise, in the setting of Theorem 4.2, we have

$$\mathbb{E}L_{\widehat{\boldsymbol{\theta}},T}(x^T) \le \inf_{\boldsymbol{\theta}} \left[L_{\boldsymbol{\theta},T}(x^T) + D(\theta_1\|\widehat{\theta}_1) + L\sqrt{T}V_T(\boldsymbol{\theta})\right] + D\sqrt{T+1} + \frac{(K+L)^2}{H}(2\sqrt{T} - 1), \tag{4.26}$$

where the infimum is over all strategies $\boldsymbol{\theta}$ over Λ.

Proof: We will only prove (4.26); the proof of (4.25) is similar. Following the proof of Theorem 4 in [19], we have

$$\begin{aligned}
\mathbb{E}L_{\widehat{\boldsymbol{\theta}},T}(x^T) =& \mathbb{E}\widehat{L}_{\widehat{\boldsymbol{\theta}},T}(z^T) \\
&\le \mathbb{E}\left(\inf_{\boldsymbol{\theta}} \left[\widehat{L}_{\boldsymbol{\theta},T}(z^T) + D(\theta_1\|\widehat{\theta}_1) + L\sqrt{T}V_T(\boldsymbol{\theta})\right] + D\sqrt{T+1} + \frac{(K+L)^2}{H}(2\sqrt{T} - 1)\right) \\
&\le \inf_{\boldsymbol{\theta}} \left[\mathbb{E}\widehat{L}_{\boldsymbol{\theta},T}(z^T) + D(\theta_1\|\widehat{\theta}_1) + L\sqrt{T}V_T(\boldsymbol{\theta})\right] + D\sqrt{T+1} + \frac{(K+L)^2}{H}(2\sqrt{T} - 1) \\
&= \inf_{\boldsymbol{\theta}} \left[L_{\boldsymbol{\theta},T}(x^T) + D(\theta_1\|\widehat{\theta}_1) + L\sqrt{T}V_T(\boldsymbol{\theta})\right] + D\sqrt{T+1} + \frac{(K+L)^2}{H}(2\sqrt{T} - 1),
\end{aligned}$$

where the first step follows from Lemma 4.1, the second step from Theorem 4.2, the third step from the fact that $\mathbb{E}\inf[\cdot] \le \inf\mathbb{E}[\cdot]$, and the last step from Lemma 4.1 and the fact that the expectation is taken with respect to the distribution of $z_t|x_t$. ■

Thus, we see that the OCP filter described in Theorem 4.1 is Hannan-consistent (in expectation) w.r.t. all static strategies in a suitably chosen convex subset $\Lambda \subset \Theta$. For any such strategy $\theta \in \Lambda$ we will have

$$\frac{\mathbb{E}R_T(\theta)}{T} = O\left(\frac{\log T}{T}\right).$$

The filter of Theorem 4.2 is Hannan-consistent (in expectation) w.r.t. all strategies $\boldsymbol{\theta}$ over Λ that satisfy $V_T(\boldsymbol{\theta}) = o(\sqrt{T})$: for any such strategy we will have

$$\frac{\mathbb{E}R_T(\boldsymbol{\theta})}{T} = O\left(\frac{1}{\sqrt{T}}\right).$$

5 Anomaly detection using feedback

In this section we consider the problem of determining a threshold at time t, denoted τ_t, such that whenever $\widehat{p}_t = \widehat{p}_t(z_t) < \tau_t$, we flag z_t as anomalous. In order to choose an appropriate level τ_t, we rely on feedback from an end-user who will indicate whether our anomaly flag is accurate. 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 we receive from the end user. To this end, two possible scenarios could be considered:

- 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 $\widehat{p}_t$ and τ_t.
- 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 is 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 non-anomalous, this information is readily available and can easily be provided to the forecaster. In the sequel, we will develop performance bounds in both settings.

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,

$$\widehat{y}_t = \operatorname{sgn}(\tau_t - \widehat{p}_t),$$

where we define $\operatorname{sgn}(a) = -1$ if $a \leq 0$ and $+1$ if $a > 0$. The number of mistakes over T time steps is given by

$$\sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} \equiv \sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau_t - \widehat{p}_t) \neq y_t\}}. \tag{5.27}$$

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 $\tau \in [0,1]$ that could be chosen in hindsight after having observed the entire sequence of probability assignments $\{\widehat{p}_t\}_{t=1}^T$. Ideally, we would like to bound

$$\sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau_t - \widehat{p}_t) \neq y_t\}} - \inf_{\tau \in [0,1]} \sum_{t=1}^{T} 1_{\{\operatorname{sgn}(\tau - \widehat{p}_t) \neq y_t\}}. \tag{5.28}$$

However, analyzing this expression is difficult owing to the fact that the function $\tau \mapsto 1_{\{\operatorname{sgn}(\tau - p) \neq y\}}$ 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* (see Chapter 12 in [5]). A frequently used surrogate is the *hinge loss* $\ell(s,y) \triangleq (1-sy)_+$, where $(\alpha)_+ = \max\{0,\alpha\}$. Indeed, for any p, τ and y we have

$$1_{\{\mathrm{sgn}(\tau-p)\neq y\}} \leq 1_{\{(\tau-p)y<0\}} \leq \big(1-(\tau-p)y\big)_+.$$

Thus, instead of (5.28), we will bound the "regret"

$$R_T(\tau) \triangleq \sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} - \sum_{t=1}^{T} \ell_t(\tau), \tag{5.29}$$

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

5.1 Anomaly detection with full feedback

In order to obtain bounds on the surrogate regret (5.29), we first need to analyze the ideal situation in which the forecaster always receives feedback from the end-user. Let $\Pi(\cdot)$ denote the projection onto the unit interval: $\Pi(\alpha) = \arg\min_{\tau\in[0,1]}(\tau-\alpha)^2$. In this setting, the following simple algorithm does the job:

Algorithm 4 Anomaly detection with full feedback

Parameters: real number $\eta > 0$.
Initialize: $\tau_t = 0$.
for $t = 1, 2, \ldots$ **do**
  if $p_t(z_t) < \tau_t$ **then**
    Flag z_t as an anomaly: $\widehat{y}_t = 1$.
  else
    Let $\widehat{y}_t = -1$.
  end if
  Obtain 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 (anomalous vs. not) 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 above algorithm is of a mirror descent 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. Nevertheless, we will be able to use the basic bound of Lemma 2.2 to analyze the regret. To that end, we have the following:

Theorem 5.1 *Fix a time horizon T and consider the forecaster acting according to Algorithm 4 with parameter $\eta = 1/\sqrt{T}$. Then, for any $\tau \in [0,1]$, we have*

$$R_T(\tau) = \sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} - \sum_{t=1}^{T} \ell_t(\tau) \leq \sqrt{T}. \tag{5.30}$$

Proof: The following is a modification of the proof of Theorem 12.1 in [5]. Let $\ell_t'(\tau)$ denote the value of the subgradient of the function $\tau \mapsto \ell_t(\tau)$. 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

$$\tau_{t+1} = \Pi\big(\tau_t - \eta \ell_t'(\tau_t)\big).$$

Define the *unprojected* update $\widetilde{\tau}_{t+1} = \tau_t - \eta \ell_t'(\tau_t)$. Then, whenever $\widehat{y}_t \neq y_t$, we may use Lemma 2.2 to write

$$\eta\big(\ell_t(\tau_t) - \ell_t(\tau)\big) \leq \eta(\tau - \tau_t)y_t \leq \frac{1}{2}\Big((\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 + (\tau_t - \widetilde{\tau}_{t+1})^2\Big). \tag{5.31}$$

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

$$1 - \ell_t(\tau) = 1 - (1 - (\tau - \widehat{p}_t)y_t)_+ \leq (\tau - \widehat{p}_t)y_t = -(\tau - \widehat{p}_t)\ell_t'(\tau_t). \tag{5.32}$$

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

$$\begin{aligned}
\eta(1 - \ell_t(\tau)) &\leq -\eta(\tau - \widehat{p}_t)\ell_t'(\tau_t) \\
&= \eta(\tau - \widehat{p}_t)y_t \\
&= \eta(\tau - \tau_t + \tau_t - \widehat{p}_t)y_t \\
&= \eta(\tau - \tau_t)y_t + \underbrace{\eta(\tau_t - \widehat{p}_t)y_t}_{<0} \\
&< \frac{1}{2}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 + (\tau_t - \widetilde{\tau}_{t+1})^2\right],
\end{aligned} \tag{5.33}$$

where in the fourth line we use the fact that $\widehat{y}_t \neq y_t$ implies that $\mathrm{sgn}(\tau_t - \widehat{p}_t) \neq y_t$, so that $(\tau_t - \widehat{p}_t)y_t < 0$. Note also that when $\widehat{y}_t = y_t$, we will have $\widetilde{\tau}_{t+1} = \tau_t$, and since $\tau_t \in [0,1]$, $\tau_{t+1} = \Pi(\widetilde{\tau}_{t+1}) = \tau_t$. Thus, the very last expression in (5.33) is identically zero when $\widehat{y}_t = y_t$. Hence, we get the bound

$$\eta(1 - \ell_t(\tau))1_{\{\widehat{y}_t \neq y_t\}} \leq \frac{1}{2}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 + (\tau_t - \widetilde{\tau}_{t+1})^2\right] \tag{5.34}$$

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

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

Now, since $\widetilde{\tau}_{t+1} = \tau_t + \eta y_t 1_{\{\widehat{y}_t \neq y_t\}}$, we can bound this further by

$$\sum_{t=1}^{T} 1_{\{\widehat{y}_t \neq y_t\}} \leq \sum_{t=1}^{T} \ell_t(\tau) + \frac{1}{2\eta}\left[1 + T\eta^2\right].$$

Choosing $\eta = 1/\sqrt{T}$, we obtain the regret bound (5.30). ■

5.2 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. Consider the following *label-efficient* forecaster for anomaly detection using sequential probability assignments:

Algorithm 5 Label-efficient anomaly detection

Parameters: real number $\eta > 0$.
Initialize: $\tau_1 = 0$.
for $t = 1, 2, \ldots$ **do**
  if $\widehat{p}_t(z_t) < \tau_t$ **then**
    Flag z_t as an anomaly: let $\widehat{y}_t = 1$.
  else
    Let $\widehat{y}_t = -1$.
  end if
  Draw a Bernoulli random variable U_t such that $\mathbb{P}[U_t = 1 | U^{t-1}] = 1/(1 + |\widehat{p}_t - \tau_t|)$.
  if $U_t = 1$ **then**
    Obtain 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 if
end for

Theorem 5.2 *Fix a time horizon T and consider the label efficient forecaster run with parameter $\eta = 1/\sqrt{T}$. Then*

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

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

Proof: The following is a modification of the proof of Theorem 12.5 in [5]. Introduce the Bernoulli random variables $M_t = 1_{\{\widehat{y}_t \neq y_t\}}$. Then the following inequality holds whenever $M_t = 1$:

$$\eta(1 - \ell_t(\tau)) \leq \eta(\tau - \widehat{p}_t) y_t.$$

We can rewrite it as

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

From this, we obtain the inequality

$$(1 + |p_t - \tau_t|) M_t U_t \leq \ell_t(\tau) + \frac{1}{2\eta}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 + (\tau_t - \widetilde{\tau}_{t+1})^2\right],$$

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 $\tau_t = \widetilde{\tau}_{t+1} = \tau_{t+1}$. If $M_t U_t = 1$, then $y_t(\tau_t - \widehat{p}_t) = -(\tau_t - \widehat{p}_t)\operatorname{sgn}(\tau_t - \widehat{p}_t) = -|\widehat{p}_t - \tau_t|$. Summing over t, we get

$$\sum_{t=1}^{T} (1 + |\widehat{p}_t - \tau_t|) M_t U_t \leq \sum_{t=1}^{T} \ell_t(\tau) + \frac{1}{2\eta}\sum_{t=1}^{T}\left[(\tau - \tau_t)^2 - (\tau - \tau_{t+1})^2 + (\tau_t - \widetilde{\tau}_{t+1})^2\right].$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T} (1 + |\widehat{p}_t - \tau_t|) M_t U_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} (1 + |\widehat{p}_t - \tau_t|) M_t \mathbb{E}_t U_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} M_t\right].$$

Using the same argument as before with $\eta = 1/\sqrt{T}$, we obtain

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

and the theorem is proved. ■

5.3 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 $\eta > 0$.
Initialize: $\tau_1 = 0$.
for $t = 1, 2, \ldots, T$ **do**
  if $\widehat{p}_t(z_t) < \tau_t$ **then**
    Flag z_t as an anomaly: let $\widehat{y}_t = 1$.
  else
    Let $\widehat{y}_t = -1$.
  end if
  if Label 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 if
end for

In this setting of 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 5.3 *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 $\epsilon \triangleq m/T$. Then, for any $\tau \in [0, 1]$, we have*

$$\sum_{i=1}^{m} 1_{\{\widehat{y}_{t_i} \neq y_{t_i}\}} \leq \sum_{i=1}^{m} \ell_{t_i}(\tau) + \frac{1+\epsilon}{2}\sqrt{T}. \tag{5.35}$$

Proof: The same proof technique as used for Theorem 5.1 yields the bound

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

With the choice $\eta = 1/\sqrt{T}$, we get the bound in the theorem. ■

Acknowledgments

The authors would like to thank Sasha Rakhlin for helpful discussions.

References

[1] M. Raginsky, R. Marcia, J. Silva, and R. Willett, "Sequential probability assignment via online convex programming using exponential families," in *Proc. of IEEE International Symposium on Information Theory*, 2009.

[2] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient descent," in *Proc. Int. Conf. on Machine Learning*, 2003, pp. 928–936.

[3] P. Bartlett, E. Hazan, and A. Rakhlin, "Adaptive online gradient descent," in *Adv. Neural Inform. Processing Systems*, vol. 20. Cambridge, MA: MIT Press, 2008, pp. 65–72.

[4] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari, "Optimal strategies and minimax lower bounds for online convex games," in *Proc. Int. Conf. on Learning Theory*, 2008, pp. 415–423.

[5] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. Cambridge Univ. Press, 2006.

[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," *Operations Res. Lett.*, vol. 31, pp. 167–175, 2003.

[8] 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.

[9] 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. Mathematics and Math. Phys.*, vol. 7, pp. 200–217, 1967.

[10] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*. Oxford Univ. Press, 1997.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrdige Univ. Press, 2004.

[12] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[13] Y. Shtarkov, "Universal sequential coding of single messages," *Problems Inform. Transmission*, vol. 23, pp. 3–17, 1987.

[14] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, pp. 199–207, March 1981.

[15] A. Barron, J. Rissanen, and B. Yu, "Minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2743–2760, October 1998.

[16] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence: American Mathematical Society, 2000.

[17] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," UC Berkeley, Dept. of Statistics, Tech. Rep. 649, 2003.

[18] K. S. Azoury and M. K. Warmuth, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Machine Learning*, vol. 43, pp. 211–246, 2001.

[19] 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, 2001.