Sequential Probability Assignment Via Online Convex Programming Using Exponential Families

Maxim Raginsky
ECE Department
Duke University
Durham, NC 27708, USA
Email: m.raginsky@duke.edu

Roummel F. Marcia
ECE Department
Duke University
Durham, NC 27708, USA
Email: roummel@ee.duke.edu

Jorge Silva
ECE Department
Duke University
Durham, NC 27708, USA
Email: jg.silva@duke.edu

Rebecca M. Willett
ECE Department
Duke University
Durham, NC 27708, USA
Email: willett@duke.edu

***Abstract*— This paper considers the problem of sequential assignment of probabilities (likelihoods) to elements of an individual sequence using an exponential family of probability distributions. We draw upon recent work on online convex programming to devise an algorithm that does not require computing posterior distributions given all current observations, involves simple primal-dual parameter updates, and achieves minimax per-round regret against slowly varying product distributions with marginals drawn from the same exponential family. We validate the theory on synthetic data drawn from a time-varying distribution over binary vectors of high dimensionality.**

I. Introduction

The problem of *sequential probability assignment* appears in such contexts as universal data compression, online learning, and sequential investment [1], [2]. 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 p_t to the possible values of x_t. When x_t is revealed, we incur the *logarithmic loss* $-\log p_t(x_t)$. We refer to any such sequence of probability assignments $\boldsymbol{p} = \{p_t\}_{t=1}^{\infty}$ as a *prediction strategy*. Since the probability assignment p_t is a function of the past observations $x^{t-1} \triangleq (x_1, x_2, \ldots, x_{t-1}) \in \mathcal{X}^{t-1}$, we may view it as a conditional probability density $p(\cdot|x^{t-1})$.

In this paper, we analyze the following prediction strategy. We restrict our attention to an exponential family of distributions $\{p_\theta\}$, where the parameter θ ranges over a convex set Λ in a Euclidean space. At time t, we choose the parameter θ_{t+1} and the corresponding distribution p_{t+1} according to

$$\theta_{t+1} \approx \operatorname*{arg\,min}_{\theta \in \Lambda} \left[-\log p_\theta(x_t) + \frac{1}{\lambda_t} D(p_\theta \| p_t) \right] \quad (1.1)$$

$$p_{t+1} = p_{\theta_{t+1}} \quad (1.2)$$

where $\lambda_t > 0$ is a regularization parameter and $D(\cdot\|\cdot)$ is the relative entropy (Kullback–Leibler divergence). We will show that this approach has several key advantages:

- The sequence $\{p_t\}_t$ achieves minimax per-round regret (see definitions below) with respect to any prediction strategy in a comparison class consisting of time-varying product distributions with marginals in $\{p_\theta\}_{\theta\in\Lambda}$, provided the variation in time is sufficiently slow. A fortiori, we achieve minimax regret relative to the best time-varying sequence that can be fitted to the entire data sequence (after a finite number of rounds) in hindsight. This is proved using recently developed theory of online convex programming [3]–[5].
- The optimization at each time can be computed using only the current observation and the probability density estimated at the previous time; it is not necessary to keep all observations in memory to ensure strong performance.
- The Kullback–Leibler regularization term induces a *Bregman divergence* [2, Ch. 11] on the parameter space of the exponential family; this allows the parameter updates to be computed using an efficient primal-dual "mirror descent" algorithm of Nemirovsky and Yudin [6], [7].

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* $\mathcal{F}$ [1], [2]. Thus, given a prediction strategy $\boldsymbol{p}$, let us define the *regret* of $\boldsymbol{p}$ w.r.t. some $\boldsymbol{f} \in \mathcal{F}$ after T time steps as

$$R_T(\boldsymbol{f}) \triangleq \sum_{t=1}^{T} \log \frac{1}{p(x_t|x^{t-1})} - \sum_{t=1}^{T} \log \frac{1}{f(x_t|x^{t-1})}. \quad (1.3)$$

The goal is to design $\boldsymbol{p}$ in such a way that

$$R_T(\boldsymbol{p}, \mathcal{F}) \triangleq \sup_{\boldsymbol{x}} \sup_{\boldsymbol{f}\in\mathcal{F}} R_T(\boldsymbol{f}) = o(T).$$

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; both, however, have certain disadvantages compared to the approach proposed in this paper. For example, a fundamental result due to Shtarkov [8] says that the *minimax regret* $R_T^*(\mathcal{F}) \triangleq \inf_{\boldsymbol{p}} R_T(\boldsymbol{p}, \mathcal{F})$, where the infimum is over all prediction strategies, is achieved by the normalized maximum likelihood estimator (MLE) over $\mathcal{F}$. However, practical use of the normalized MLE strategy is limited since it requires solving an optimization problem over $\mathcal{F}$ whose complexity increases with T.

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

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

via the posterior $p(a|x^{t-1}) \triangleq \frac{p(a,x^{t-1})}{p(x^{t-1})}$. For instance, when the underlying observation space $\mathcal{X}$ is finite and the reference class $\mathcal{F}$ consists of all product distributions of the form $f(x^t) = \prod_{s=1}^t f_0(x_s)$, where f_0 is some probability mass function on $\mathcal{X}$, the well-known Krichevsky–Trofimov (KT) predictor [9]

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

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

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 [2]. This can be thought of as a generalization of the MDL-type regret bounds of Rissanen [1], [10] to the online, individual-sequence setting. However, the predictive distributions output by a mixture strategy will not, in general, lie in $\mathcal{F}$, which is often a reasonable requirement.

In this paper, we show that the prediction strategy based on (1.1) and (1.2) leads to an algorithm that does not require choosing a prior distribution over the comparison class (thus avoiding the need to compute posteriors conditioned on observed sequences of increasing length), has simple update rules, and does not rely on empirical frequencies. In addition to proving a regret bound for our algorithm, we demonstrate its empirical performance in the sequential probability assignment for high-dimensional binary vectors. An algorithm similar to (1.1) and (1.2) was suggested by Azoury and Warmuth [11] for the problem of sequential probability assignment over an exponential family, but they 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.

II. Exponential families and prediction strategies

Our main contribution is an online convex programming (OCP) algorithm for sequential probability assignment that can compete with time-varying prediction strategies induced by an exponential family of probability densities. Exponential families (see, e.g., [12], [13] and references therein) are a natural choice because (a) their log-likelihood functions are convex in the underlying parameter, and (b) their geometric properties immediately lead to an efficient implementation of OCP using the method of mirror descent [6], [7] (see [2, Chap. 11] for a detailed description of the mirror descent algorithm in the context of sequential linear prediction).

More specifically, we construct the class of candidate distributions from which we select each p_t and the comparison class of distributions $\mathcal{F}$ as follows. 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 $\langle \theta, \phi(x) \rangle = \theta_1\phi_1(x) + \ldots + \theta_d\phi_d(x)$. The function Φ is the so-called *log partition function*. Finally, let p_0 be a fixed reference density. Then the *exponential family* induces by ϕ is

$$\mathcal{P}(\phi) = \left\{ p_\theta(\cdot) \triangleq p_0(\cdot) \exp\big(\langle \theta, \phi(\cdot) \rangle - \Phi(\theta)\big) : \theta \in \Theta \right\}.$$

We will use $\mathbb{E}_\theta[\cdot]$ to denote expectations w.r.t. p_θ.

Our comparison class of prediction strategies $\mathcal{F} = \{\boldsymbol{f_\theta} : \theta \in \Lambda\}$ will be made up of product distributions where each marginal belongs to a certain subset of $\mathcal{P}(\phi)$. We assume that the functions ϕ_k are bounded: $|\phi_k(x)| \le G/2$ for some $G < +\infty$. To define $\mathcal{F}$, we choose a closed, convex set $\Lambda \subseteq \Theta$ satisfying the following condition: there exist constants $H_1, H_2 > 0$, such that, for every $\theta \in \Lambda$, $2H_1 I_d \preceq \nabla^2\Phi(\theta) \preceq 2H_2 I_d$, where I_d denotes the $d \times d$ identity matrix. ($A \preceq B$ means $B - A$ is positive semidefinite.) Since $\Phi(\theta)$ is a convex function of θ, this condition amounts to placing upper and lower bounds on the curvature of Φ, which will guarantee the convexity of Λ. 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 [12]. Our assumption on Λ thus stipulates that the eigenvalues of the Fisher information matrix are bounded between $2H_1$ and $2H_2$ on Λ. Moreover, $\kappa \triangleq H_2/H_1$ can be viewed as a bound on the *condition number* of $J(\theta)$, $\theta \in \Lambda$.

Then let $\mathcal{F}$ consist of prediction strategies $\boldsymbol{f_\theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ ranges over all infinite sequences over Λ, and each $\boldsymbol{f_\theta}$ is of the form

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

In other words, each prediction strategy in $\mathcal{F}$ is a time-varying product density whose marginals belong to $\{p_\theta : \theta \in \Lambda\}$.

III. Sequential probability assignment using OCP and exponential families

Recent results on online convex programming (OCP) [3]–[5] make it possible to analyze the performance of a Forecaster who is continually predicting changes in a dynamic Environment. 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 running cost as low as possible. This framework has not been used conventionally in the context of sequential probability assignment, but it is a natural fit in that the assigned probabilities p_t are essentially forecasts of changing environmental variables x_t, and the exponential-family negative log-likelihoods $-\log p_\theta(x_t)$ are convex functions of θ. Let us define the loss function $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ by

$$\ell(x,\theta) \triangleq -\log p_\theta(x) = \Phi(\theta) - \langle \theta, \phi(x)\rangle - \log p_0(x). \quad (3.4)$$

In an OCP setting, this loss is referred to as the *cost* incurred by the Forecaster's choice of θ when the Environment produces x. Owing to the convexity of Φ, $\theta \mapsto \ell(x,\theta)$ is convex for any $x \in \mathcal{X}$. The convexity of Φ can be established by considering its derivatives; because $\mathcal{P}(\phi)$ is an exponential family, the log partition function $\Phi(\theta)$ is lower semicontinuous on $\mathbb{R}^d$ and infinitely differentiable on Θ. The derivatives of Φ at θ are the cumulants of the random vector $\phi(X) = \big(\phi_1(X), \dots, \phi(X)\big)^T$ when $X \sim p_\theta$. In particular, $\nabla\Phi(\theta) = \big(\mathbb{E}_\theta\, \phi_1(X), \dots, \mathbb{E}_\theta\, \phi_d(X)\big)^T$ and

$$\nabla^2\Phi(\theta)_{i,j} = \mathrm{Cov}_\theta\big(\phi_i(X), \phi_j(X)\big), \qquad 1 \le i,j \le d.$$

The latter property implies that $\Phi(\theta)$ is a convex function of θ. Therefore, the set Θ, which is the essential domain of Φ, is convex. We denote by Θ^* the image of Θ under the gradient mapping $\theta \mapsto \nabla\Phi(\theta)$, which maps the *primal parameter* $\theta \in \Theta$ to the corresponding *dual parameter* $\mu \in \Theta^*$. The gradient mapping is invertible, with the inverse $\mu \mapsto \nabla\Phi^*(\mu)$, where

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

is the Legendre–Fenchel dual of Φ [13].

Using (3.4), we can express the regret (1.3) of any $\boldsymbol{f_\theta} \in \mathcal{F}$ w.r.t. some other $\boldsymbol{f_{\theta^*}} \in \mathcal{F}$ after T time steps as

$$R_T(\boldsymbol{f_\theta}) = \sum_{t=1}^T \ell(x_t, \theta_t) - \sum_{t=1}^T \ell(x_t, \theta_t^*).$$

A. Proposed algorithm

We now present our OCP approach to sequential probability assignment using exponential families. Our scheme is described below as Algorithm 1; $\{\lambda_t\}_{t=1}^\infty$ is a decreasing sequence of step sizes. The algorithm is essentially a mirror descent procedure [6], [7] (see also [2, Chap. 11]): after seeing each new observation x_t, we compute the probability assignment $p_{\theta_{t+1}}$ for x_{t+1} by first performing gradient descent in the space Θ^* of the dual parameters (thought of as the "mirror image" of the primal space Θ), then mapping back to the primal space Θ via the inverse gradient mapping $\nabla\Phi^*$, and finally by projecting onto Λ. This is illustrated in Fig. 1. It can be shown [2], [7] that the combination of the dual update and the projected primal update is equivalent to finding

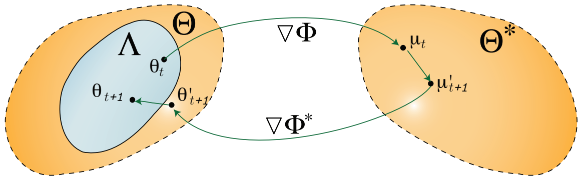

Fig. 1. Graphical depiction of mirror descent.

$$\min_{\theta\in\Lambda} \left[-\langle \theta, \nabla_\theta \log p_{\theta_t}(x_t)\rangle + \frac{1}{\lambda_t} D(\theta \| \theta_t) \right].$$

This corresponds to regularized minimization of the first-order Taylor approximation to $-\log p_\theta(x_t)$ around $\theta = \theta_t$, as shown in (1.1) and (1.2).

Algorithm 1

```
Initialize with $\theta_1 \in \Lambda$
for $t = 1, 2, \dots$ do
    Acquire new observation $x_t$
    Incur the cost $\ell_t(\theta_t) = -\log p_{\theta_t}(x_t)$
    Compute $\mu_t = \nabla\Phi(\theta_t)$
    Dual update: compute $\mu'_{t+1} = \mu_t - \lambda_t \nabla \ell_t(\theta_t)$
    Projected primal update: compute $\theta'_{t+1} = \nabla\Phi^*(\mu'_{t+1})$ and
        $\theta_{t+1} = \arg\min_{\theta\in\Lambda} D(\theta \| \theta'_{t+1})$
end for
```

In mirror descent, the dual variables are related to the primal variables via the gradient of the so-called *potential function*, which in the current setting is the log partition function Φ. The role of the potential function is to induce a regularization functional on the underlying feasible set [7]. In the context of exponential families, this regularization functional turns out to be the relative entropy. It can be shown [13] that the relative entropy 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

$$D(p_\theta \| p_{\theta'}) = \Phi(\theta) - \Phi(\theta') - \langle \nabla\Phi(\theta'), \theta - \theta'\rangle \qquad (3.5)$$

From now on, we will use the shorthand $D(\theta\|\theta')$. The fact that the relative entropy can be written in the form (3.5), together with the analytic properties of the log partition function [13], implies that the mapping $D(\cdot\|\cdot) : \Theta \times \mathrm{Int}\,\Theta \to \mathbb{R}$ is a *Bregman divergence* (see, e.g., [2, Ch. 11]) on Θ.

As a Bregman divergence, $D(\cdot\|\cdot)$ satisfies the *generalized Pythagorean inequality*: Let Λ be a closed convex subset of Θ. Given any $\theta_0 \in \Theta$, let $\widetilde{\theta}_0 \in \Lambda$ denote the *Bregman projection* of θ_0 onto Λ, defined by $\widetilde{\theta}_0 \triangleq \arg\min_{\xi\in\Lambda} D(\xi\|\theta_0)$. Then for all $\theta \in \Lambda$ we have

$$D(\theta\|\theta_0) \ge D(\theta\|\widetilde{\theta}_0) + D(\widetilde{\theta}_0\|\theta_0). \qquad (3.6)$$

B. Tracking regret of the proposed method

The fact that $D(\cdot\|\cdot)$ is a Bregman divergence is a key aspect of the proof of our main result: a bound on the regret of Algorithm 1 (with suitably chosen λ_t's) w.r.t. any prediction strategy in $\mathcal{F}$. Following the terminology of [2], we refer to this regret as a *tracking regret* since it quantifies how well the proposed prediction strategy can "track" a time-varying sequence of probability distributions drawn from our comparison class $\mathcal{F}$.

Theorem 3.1. Let $\mathcal{F}$ be the comparison class described above. Suppose that Algorithm 1 is run with $\lambda_t = \kappa/t$, $t = 1, 2, \ldots$, where $\kappa = H_2/H_1$. Then for any $\boldsymbol{f}_{\boldsymbol{\theta}^*} \in \mathcal{F}$ we have

$$\sum_{t=1}^{T} \ell(x_t, \theta_t) \leq \sum_{t=1}^{T} \ell(x_t, \theta_t^*) + \frac{LT \cdot V_T(\boldsymbol{\theta}^*)}{\kappa} + \frac{\kappa L^2}{4H_1}(\log T + 1), \tag{3.7}$$

where $L = \sqrt{d}G$ and $V_T(\boldsymbol{\theta}^*) \triangleq \sum_{t=1}^{T} \|\theta_t^* - \theta_{t+1}^*\|$ is the *variation* of the sequence $\boldsymbol{\theta}^*$ from $t = 1$ to $t = T + 1$.

Proof: The proof combines the technique used in [2] to analyze tracking regret of linear prediction strategies with OCP-type arguments [4]. Fix some sequence $\boldsymbol{\theta}^*$ over Λ. Given $x_t \in \mathcal{X}$ at time t and any $\theta \in \Theta$, we use the shorthand $\ell_t(\theta) \equiv \ell(x_t, \theta)$ and $\nabla \ell_t(\theta) \equiv \nabla_\theta \ell(x_t, \theta)$. Note that $\nabla_\theta^2 \ell_t(\theta) = \nabla^2 \Phi(\theta)$. Hence, $\ell_t(\theta)$ is $2H_1$-strongly convex [14]: for all $\theta, \theta' \in \Lambda$

$$\ell_t(\theta') \geq \ell_t(\theta) + \langle \nabla \ell_t(\theta), \theta' - \theta \rangle + H_1 \|\theta' - \theta\|^2.$$

Using this fact, the definition of the dual update ($\mu'_{t+1} = \mu_t - \lambda_t \nabla \ell_t(\theta_t)$), and the properties of the gradient map $\nabla \Phi$, we have

$$\ell_t(\theta_t) - \ell_t(\theta_t^*) \leq \frac{1}{\lambda_t} \langle \nabla \Phi(\theta'_t) - \nabla \Phi(\theta_t), \theta_t^* - \theta_t \rangle - H_1 \|\theta_t^* - \theta_t\|^2.$$

Next, using Eq. (3.5), we can write the inner product above as $D(\theta_t^* \| \theta_t) - D(\theta_t^* \| \theta'_{t+1}) + D(\theta_t \| \theta'_{t+1})$. Since θ_{t+1} is the Bregman projection of θ'_{t+1} onto the closed, convex set Λ, it follows from the generalized Pythagorean inequality (3.6) that

$$D(\theta_t^* \| \theta'_{t+1}) \geq D(\theta_t^* \| \theta_{t+1}) + D(\theta_{t+1} \| \theta'_{t+1}) \geq D(\theta_t^* \| \theta_{t+1}),$$

where the second step uses the fact that $D(\cdot\|\cdot) \geq 0$. Also, our assumption on $\nabla^2 \Phi(\theta)$ over Λ implies that for all $\theta, \theta' \in \Lambda$

$$H_1 \|\theta - \theta'\|^2 \leq D(\theta \| \theta') \leq H_2 \|\theta - \theta'\|^2 \tag{3.8}$$

Hence, we can bound $\ell_t(\theta_t) - \ell_t(\theta_t^*)$ from above by

$$\frac{1}{\lambda_t} \left[D(\theta_t^* \| \theta_t) - D(\theta_t^* \| \theta_{t+1}) \right] - \frac{1}{\kappa} D(\theta_t^* \| \theta_t) + \frac{1}{\lambda_t} D(\theta_t \| \theta'_{t+1}).$$

Adding and subtracting $D(\theta_{t+1}^* \| \theta_{t+1})$ inside the brackets and using the definition of λ_t, we rewrite the first three terms as

$$\frac{1}{\lambda_{t-1}} D(\theta_t^* \| \theta_t) - \frac{1}{\lambda_t} D(\theta_{t+1}^* \| \theta_{t+1}) + \frac{1}{\lambda_t} \Delta_t,$$

where we have set $1/\lambda_0 \equiv 0$ and $\Delta_t \triangleq D(\theta_{t+1}^* \| \theta_{t+1}) - D(\theta_t^* \| \theta_{t+1})$. Since $|\phi_k(x)| \leq G/2$ for all x, k, $\|\nabla \Phi(\theta)\| = \|\mathbb{E}_\theta \, \phi(X)\| \leq L/2$ for all θ. Using this and Eq. (3.5),

$$\begin{aligned} \Delta_t &= \Phi(\theta_{t+1}^*) - \Phi(\theta_t^*) - \langle \nabla \Phi(\theta_{t+1}), \theta_t^* - \theta_{t+1}^* \rangle \\ &\leq L \|\theta_t^* - \theta_{t+1}^*\|. \end{aligned}$$

Finally, we deal with the term involving $D(\theta_t \| \theta'_{t+1})$. Using (3.5), the property of the gradient mapping $\nabla \Phi$, and the definition of the dual update rule, we get

$$D(\theta_t \| \theta'_{t+1}) + D(\theta'_{t+1} \| \theta_t) = \lambda_t \langle \nabla \ell_t(\theta_t), \theta_t - \theta'_{t+1} \rangle.$$

Using Cauchy–Schwarz and $(a + b)^2 \geq 0$, we further get

$$D(\theta_t \| \theta'_{t+1}) + D(\theta'_{t+1} \| \theta_t) \leq \frac{\lambda_t^2 \|\nabla \ell_t(\theta_t)\|^2}{4H_1} + H_1 \|\theta_t - \theta'_{t+1}\|^2.$$

Using this, (3.8), and $\|\nabla \ell_t(\theta_t)\| \leq \|\nabla \Phi(\theta_t)\| + \|\phi(x_t)\| \leq L$, we get $D(\theta_t \| \theta'_{t+1}) \leq \frac{\lambda_t^2 L^2}{4H_1}$. Combining everything and summing from $t = 1$ to T gives

$$\begin{aligned} \sum_{t=1}^{T} [\ell_t(\theta_t) - \ell_t(\theta_t^*)] &\leq \sum_{t=1}^{T} \left(\frac{D(\theta_t^* \| \theta_t)}{\lambda_{t-1}} - \frac{D(\theta_{t+1}^* \| \theta_{t+1})}{\lambda_t} \right) \\ &\quad + \sum_{t=1}^{T} \frac{L \|\theta_t^* - \theta_{t+1}^*\|}{\lambda_t} + \sum_{t=1}^{T} \frac{\lambda_t \|\nabla \ell_t(\theta_t)\|^2}{4H_1} \\ &\leq \sum_{t=1}^{T} \frac{LT \|\theta_t^* - \theta_{t+1}^*\|}{\kappa} + \sum_{t=1}^{T} \frac{\kappa L^2}{4H_1 t} \\ &\leq \frac{LT \cdot V_T(\boldsymbol{\theta}^*)}{\kappa} + \frac{\kappa L^2}{4H_1}(\log T + 1). \end{aligned}$$

This finishes the proof. ■

We immediately get the following bounds on the per-round regret against all constant strategies and against all sufficiently slowly varying strategies in $\mathcal{F}$:

Corollary 3.2. Let $\mathcal{F}_{\text{const}}$ be the subset of $\mathcal{F}$ consisting of all constant strategies $\boldsymbol{f}_{\boldsymbol{\theta}^*}$, where $\theta_1^* = \theta_2^* = \ldots$. Then

$$\sup_{\boldsymbol{f} \in \mathcal{F}_{\text{const}}} \frac{1}{T} R_T(\boldsymbol{f}) \leq \frac{\kappa L^2}{4H_1} \frac{\log T + 1}{T}.$$

Let $\mathcal{F}_{\text{slow}}$ be any subset of $\mathcal{F}$ consisting of slowly varying prediction strategies $\boldsymbol{f}_{\boldsymbol{\theta}^*}$, such that $V_T(\boldsymbol{\theta}^*) = o(1)$. Then

$$\sup_{\boldsymbol{f} \in \mathcal{F}_{\text{slow}}} \frac{1}{T} R_T(\boldsymbol{f}) \leq \frac{\kappa L^2}{4H_1} \frac{\log T + 1}{T} + \frac{L}{\kappa} o(1).$$

Remark 3.1. The $O(T^{-1} \log T)$ per-round regret is minimax-optimal for OCP with strongly convex cost functions [5].

IV. Experiments

In this section, we show how our approach performs on simulated data drawn from time-varying Bernoulli product densities. This allows us to compute the empirical regret with respect to the known data generation parameters, and to compare it against the theoretical regret bound of Theorem 3.1. We draw i.i.d. observations from d-dimensional Bernoulli product densities with exponential-family parameter vectors $\theta_t^* = (\alpha_{1,t}^*, \ldots, \alpha_{d,t}^*)^T \in \mathbb{R}^d$, i.e.,

$$\log p_{\theta_t^*}(x_t) = \langle \theta_t^*, x_t \rangle - \sum_{i=1}^{d} \log[1 + \exp(\alpha_{i,t}^*)].$$

Note that the corresponding mean parameters are $\mu_t^* = (\beta_{1,t}^*, \ldots, \beta_{d,t}^*)^T$ with $\beta_{i,t}^* = \exp(\alpha_{i,t}^*)/(1 + \exp(\alpha_{i,t}^*))$. That is, $x_t \sim \prod_{i=1}^{d} \text{Bernoulli}(\beta_{i,t}^*)$. The initial θ_1^* is drawn at random from the cube $[-\alpha_0, \alpha_0]^d$ for some $\alpha_0 > 0$, and at certain jump times we draw new parameter vectors. The value of θ_t^* (and therefore μ_t^*) is kept constant between jumps.

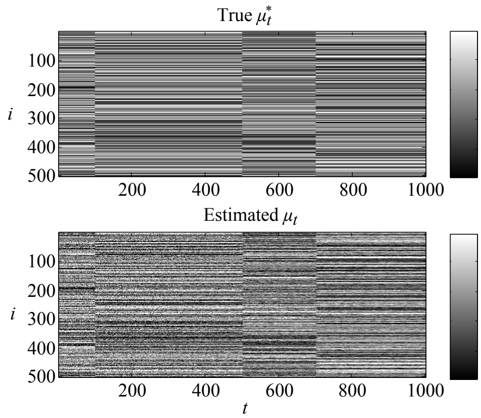

Fig. 2. Estimated μ vs. ground truth. The μ values correspond to Bernoulli means. Lighter colors depict higher probabilities.

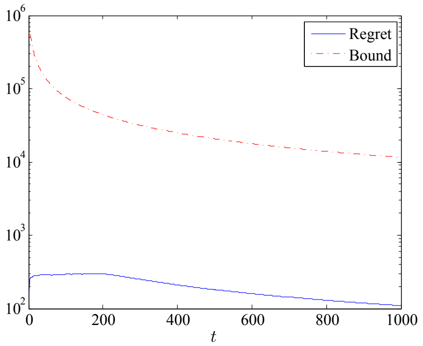

Fig. 3. Per-round regret compared to upper bound.

For product Bernoulli densities with canonical parameter $\theta = (\alpha_1, \ldots, \alpha_d)^T$, the Hessian of $\Phi(\theta)$ is a diagonal matrix with elements $\nabla^2\Phi(\theta)_{i,j} = \delta_{i,j} \cdot \exp(\alpha_i)/(1+\exp(\alpha_i))^2$. The maximum value of the function $\alpha \mapsto \exp(\alpha)/(1+\exp(\alpha))^2$ is $\frac{1}{4}$ (attained at $\alpha = 0$), and its infimum is zero (attained as $|\alpha| \to \infty$). To define our feasible set Λ, we set $H_2 = \frac{1}{8}$, $H_1 = \frac{\exp(\alpha_0)}{2(1+\exp(\alpha_0))^2}$. Thus, $\Lambda = [-\alpha_0, \alpha_0]^d$, which is closed and convex. For our results, we use $\alpha_0 = 4$, so that $\kappa \approx 193$. We also use $d = 500$, $1 \leq t \leq 1000$, and generate three jumps at $t = 100$, 500 and 700.

Figure 2 illustrates the estimated parameter vector vs. the ground truth. We show μ_t, which is easier to interpret than θ_t, since its components are numbers between 0 and 1, corresponding to Bernoulli parameters assigned to the components of x_t. Figure 3 shows that the empirical per-round regret is well below the theoretical bound (3.7) of Theorem 3.1. Finally, Figure 4 shows that the log-loss exhibits pronounced spikes at the jump times, and then subsides as the forecaster adapts to the new parameters. It can also be observed that, for $t < \kappa \approx 193$, the log-loss is higher than for subsequent t. In this transient period, the dual update must be followed by clipping to keep each component of μ_t in the interval $[0, 1]$.

V. Summary

The online convex programming framework offers a natural set of tools for addressing the sequential probability assignment problem. The conventional perspective of OCP using a Forecaster to make predictions about a changing Environment translates into using sequential probability assignment to make predictions about a changing individual sequence. In this paper, we explored the theoretical ramifications of this connection, and in particular demonstrated that OCP leads to a sequential probability assignment scheme that (a) performs (in a minimax sense) as well as the very best (even clairvoyant) predictor in a broad comparison class of time-varying product exponential family distributions and (b) has high computational efficiency and minimal memory requirements.

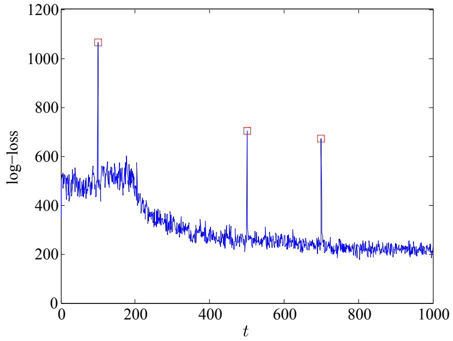

Fig. 4. Evolution of the log-loss. Note the spikes at the jump times ($t = 100$, 500 and 700), and the transient for $t < \kappa \approx 193$.

Acknowledgment

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.

References

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

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

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

[4] 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.

[5] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

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

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