Adaptive Estimation of Small Event Rates

Proceedings of the 55th IEEE Conference on Decision and Control, Las Vegas, NV, December 12 - 14, 2016

Niklas Karlsson[1]

***Abstract*— Motivated by a need in online advertising, where control systems often involve estimators of very small event rates, we propose an adaptive algorithm that regulates the stiffness of an otherwise time-invariant Bayesian event rate estimator to maintain a desired relative steady-state standard deviation of the event rate estimate. The result is an estimator that is fast (agile) when permitted by the observed input data, and that is slow (stiff) only when necessary to maintain the desired relative steady-state standard deviation of the estimate.**

I. Introduction

Event rate estimation is required in many engineering systems. The properties of the event rate estimate often have a major impact on performance and optimality of the overall system, but frequently also on stability and robustness. In some systems the unknown event rate p is constant, but in general it is time-varying. The problem considered in this paper is to estimate $p(k)$ given observations of $n_E(k)$ and $n_I(k)$, where $n_E(k)$ is a realization of $N_E(k) \sim \text{Binomial}(n_I(k), p(k))$.

One industry in which event rate estimation is a great challenge and of high importance is that of online advertising. The online advertising industry has over the last decade grown dramatically in size, significance, and complexity. In short, the goal of online advertising is to use data-driven automation and optimization to manage the marketing budgets of advertisers [1], [2]. The optimization involves constraints imposed by the advertisers, as well as dynamics, uncertainties, and noise resulting from interactions across ad campaigns and between ad campaigns and Internet users. The optimization problem can be turned into separate problems of event rate estimation and feedback control, where the event rates relate to the probability a user will click on, or otherwise respond to, an ad [3]. Typical event rates in online advertising are between 10^{-6} and 10^{-4}; i.e., on average only one *impression* (ad view) out of 10,000-1,000,000 impressions results in a click or a conversion (a sale).

For less challenging event rate estimation problems, an acceptable estimator of p can often be selected for example as a moving average or a Kalman filter of the observed event rate $n_E(k)/n_I$, where $n_E(k)$ is a realization of $\text{Binomial}(n_I, p)$ [4]. However, for problems involving small event rates and insufficient number of experiments n_I, or time-varying values of n_I or p, the estimator must be designed more carefully. Indeed, it is important to account for the non-Gaussian error distribution and the time-varying error variance resulting from $n_I(k)$ or $p(k)$ changing over time. A more sound approach is based on the Bayesian framework where the Binomial measurement model is explicitly used in the estimator design [5].

An additional challenge in the online advertising use case is that the same algorithm usually must work well for all ads, hence must work seamlessly for a range of p spanning $10^{-6} - 10^{-4}$, and a wide range of impression counts $n_I(k)$, say, $10^2 - 10^6$. The relative variance of event observations $N_E(k)$, defined by $\text{Var}(N_E(k)/\text{E}(N_E(k)))$, is dramatically larger when $p = 10^{-6}$ and $n_I(k) = 10^2$, than when $p = 10^{-4}$ and $n_I(k) = 10^6$. Fast response and small variance are conflicting properties of an estimator, hence it is desired to have an event rate estimator trading these properties against each other in a sound and automated manner. The present paper is addressing this need by developing an algorithm that adjusts the gain of an estimator in such a way that the relative variance of the event rate estimate maintains a desired value [6].

The paper is organized as follows. We begin in Section II by formulating the problem. Thereafter, in Section III we derive a static estimator, which is a building block of the ultimate adaptive estimator. The behavior of the static estimator and its limitations which motivates adaptation are demonstrated via simulations in Section IV. In Section V the adaptive algorithm is developed, and in Section VI its behavior is demonstrated via simulations. We conclude the paper in Section VII with some final remarks.

II. Problem Formulation

Let $n_I(k)$ and $n_E(k)$ denote observed counts of impressions and *events* at time k. The event count $n_E(k)$ is a realization of a random variable $N_E(k)$, where

$$N_E(k) \quad \sim \quad \text{Binomial}(n_I(k), p) \tag{1}$$

for some unknown event rate p independent of $n_I(k)$. Assume $n_I(k) \geq 100$ and $p \ll 1$. Furthermore, suppose p is constant or at most slowly varying and subject to step changes at most rarely. Derive a recursive estimator of p using a loss function $L(p, \hat{p}) = (p - \hat{p})^2$, where the estimator trades responsiveness and volatility adaptively based on observed data.

III. Static Event Rate Estimation

We first derive an optimal static Bayesian event rate estimator. This estimator is a building block of the adaptive estimator. Under the assumption $n_I(k) \geq 100$, $n_I(k)p \leq 10$, and p approximately constant; and by virtue of the Central

[1]N. Karlsson is Vice President of R&D at Aol Platforms, 395 Page Mill Road, Palo Alto, CA 94306, USA niklas.karlsson at teamaol.com

Limit Theorem [7], $N_E(k) \sim \text{Binomial}(n_I(k), p)$ is well approximated using a Poisson distribution

$$N_E(k) \quad \sim \quad \text{Poisson}\left(n_I(k)p\right). \tag{2}$$

Consider one pair of observations n_I and n_E and leave out the time index k until later. The probability mass function of n_E based on the Poisson model is given by

$$f\left(n_E|n_I,p\right) \quad = \quad \frac{(n_I p)^{n_E} e^{-n_I p}}{n_E!}.$$

Given a prior belief function $\pi(p)$ for p, we proceed by using Bayes rule to determine a posterior $\pi(p|n_I,n_E)$ that incorporates the evidence of p encoded in n_I and n_E.

Since the conjugate prior of the Poisson probability mass function is given by the gamma probability density function, we assume the prior belief function $\pi(p)$ is given by Gamma(α_p,β_p). A standard application of Bayes rule $\pi\left(p|n_I,n_E\right) = f\left(n_E|n_I,p\right)\pi\left(p|n_I\right)/f\left(n_E|n_I\right)$, where $\pi\left(p|n_I\right) = \pi\left(p\right)$ and $f\left(n_E|n_I\right)$ is independent of p and therefore only a normalization coefficient of $\pi\left(p|n_I,n_E\right)$, shows that $p|n_I,n_E \sim \text{Gamma}\left(\alpha_p+n_E,\beta_p+n_I\right)$. This update scheme applies recursively for each new set of measurements $n_I(k), n_E(k)$, $k=1,2,\ldots$. Therefore,

$$\begin{aligned} &p|n_I(1:k),n_E(1:k) \\ &\quad\sim \quad \text{Gamma}\left(\alpha_p+\sum_{i=1}^{k} n_E(i), \beta_p+\sum_{i=1}^{k} n_I(i)\right), \end{aligned}$$

or equivalently by defining the shape and inverse scale parameters as a time-invariant Markov process with state $[\alpha_p(k),\beta_p(k)]^T$

$$\left[\begin{array}{c}\alpha_p(k)\\ \beta_p(k)\end{array}\right] \quad = \quad \left[\begin{array}{c}\alpha_p(k-1)\\ \beta_p(k-1)\end{array}\right] + \left[\begin{array}{c} n_E(k)\\ n_I(k)\end{array}\right],$$

and initialized by $[\alpha_p(0),\beta_p(0)]^T = [\alpha_{p,0},\beta_{p,0}]^T$, the posterior distribution of p at time k is given by Gamma$(\alpha_p(k),\beta_p(k))$.

However, since the unknown p may vary over time, it is necessary to discount old observations and rely more on recent observations in the estimator. We do this by introducing a memory loss specified by a forgetting factor λ_p, where $0 < \lambda_p < 1$. In particular, we adopt the revised update scheme $\alpha_p(k) = \lambda_p^k\alpha_{p,0} + \sum_{i=1}^{k}\lambda_p^{k-i}n_E(i)$ and $\beta_p(k) = \lambda_p^k\beta_{p,0} + \sum_{i=1}^{k}\lambda_p^{k-i}n_I(i)$. This leads to the posterior distribution

$$\begin{aligned} &p|n_I(1:k),n_E(1:k) \sim \\ &\quad \text{Gamma}\left(\lambda_p^k\alpha_{p,0}+\sum_{i=1}^{k}\lambda_p^{k-i}n_E(i), \lambda_p^k\beta_{p,0}+\sum_{i=1}^{k}\lambda_p^{k-i}n_I(i)\right), \end{aligned}$$

or, in recursive form:

$$\alpha_p(k) \quad = \quad \lambda_p\alpha_p(k-1)+n_E(k) \tag{3}$$

$$\beta_p(k) \quad = \quad \lambda_p\beta_p(k-1)+n_I(k) \tag{4}$$

The posterior expected loss given a squared loss function of $L(p,\hat{p}) = (p-\hat{p}(k))^2$ is defined by

$$\begin{aligned} &\text{E}^{\pi(p|n_I(1:k),n_E(1:k))}(L(p,\hat{p})) \\ &\quad = \quad \int_0^\infty (p-\hat{p}(k))^2\pi\left(p|n_I(1:k),n_E(1:k)\right)dp \end{aligned}$$

Minimizing $\text{E}^{\pi(p|n_I(1:k),n_E(1:k))}(L(p,\hat{p}))$ with respect to $\hat{p}(k)$ yields

$$\hat{p}(k) \quad = \quad \frac{\alpha_p(k)}{\beta_p(k)}. \tag{5}$$

This incorporates all available evidence into the estimate and defines the optimal Bayesian event rate estimate.

Equations (3)-(5) represents a recursive and constant computational cost estimator, but for the purpose of analysis leading up to the adaptive estimator it is useful to also express the static estimator in closed form. It is trivial to show that the estimator can be expressed as

$$\hat{p}(k) \quad = \quad \frac{\lambda_p^k\alpha_{p,0}+\sum_{i=1}^{k}\lambda_p^{k-i}n_E(i)}{\lambda_p^k\beta_{p,0}+\sum_{i=1}^{k}\lambda_p^{k-i}n_I(i)}. \tag{6}$$

Before we develop the adaptive estimator, let us in next section demonstrate the behavior of the static estimator to help motivate the need of adaptation.

IV. Example: Static Event Rate Estimation in Online Advertising

Consider four weeks' worth of simulated hourly impression and event volume data. The average hourly impression volume the first two weeks equals 80,000, and the last two weeks 40,000. However, the impression data is subject to a dramatic sinusoidal time-of-day seasonality with an amplitude equal to 80% of the average hourly impression volume. The unknown event rate equals 10^{-5} in weeks one and two, and $5\cdot10^{-4}$ in weeks three and four. Finally, the event volume is obtained by random number generation from the Binomial model in (1).

Figure 1 shows the impression and event volume time se-

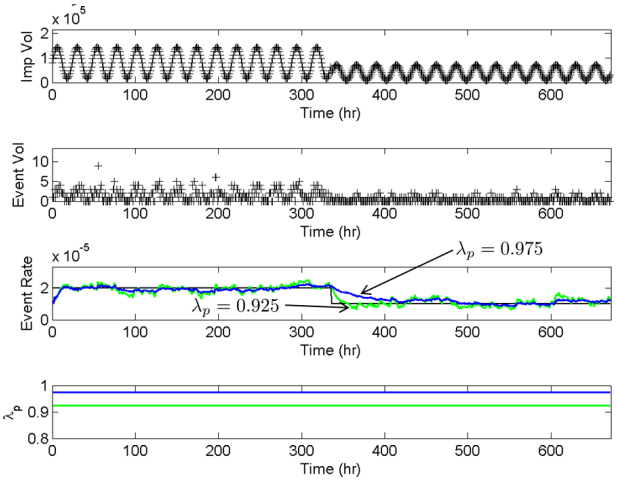

Fig. 1. Demonstration of the event rate estimation using a static Bayesian estimator (3)-(5) for two different values of λ_p.

ries data in the top and middle graphs, and demonstrates the performance of the static Bayesian event rate estimator (3)-(5) in the bottom graph. The estimator is implemented for two different values of λ_p. Based on a qualitative assessment of the performance we suggest both $\lambda_p = 0.925$ and $\lambda_p = 0.975$ lead to acceptable estimates when impression volume and event rates are large. However, the transient response after the regime change is very slow when $\lambda_p = 0.975$ and the steady state variance of the estimate is large when $\lambda_p = 0.925$. Therefore, neither 0.925 nor 0.975 is satisfactory as universal values of λ_p. We conclude there is a need to adjust the value adaptively.

A different perspective of almost the same requirement is the need for a single algorithm to work seamlessly for advertisement that is subject to any combination of high or low impression volume, and high or low event rate. It is not practical for an automated ad optimization platform to contain estimation algorithms that require manual tuning depending on properties such as impression and event rates. Moreover, relying on logic switches and/or a gain schedule to select an estimator gain λ_p is typically a fragile solution.

V. Adaptive Event Rate Estimation

A. Volatility of Event Rate Estimate

Consider now the event rate estimate as a random variable $\hat{P}(k)$ stochastic in $N_E(i)$. It follows from (6) that

$$\hat{P}(k) = \frac{\lambda_p^k \alpha_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} N_E(i)}{\lambda^k \beta_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} n_I(i)},$$

where $N_E(i) \sim \text{Binomial}(n_I(i), p)$ for $i = 1, \ldots, k$. Define $\sigma_{rel}^2(k)$ by the variance of $\hat{P}(k)$ divided by its mean; i.e.,

$$\begin{aligned} \sigma_{rel}^2(k) &:= \text{Var}\left(\frac{\hat{P}(k)}{\text{E}(\hat{P}(k))}\right) \\ &= \frac{1}{\text{E}(\hat{P}(k))^2}\text{Var}\left(\hat{P}(k)\right). \end{aligned}$$

Using known properties of the expectation and variance operators, it is straight forward to establish

$$\begin{aligned} \text{E}\left(\hat{P}(k)\right) &= \text{E}\left(\frac{\lambda_p^k \alpha_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} N_E(i)}{\lambda_p^k \beta_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} n_I(i)}\right) \\ &= \frac{\lambda_p^k \alpha_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} \text{E}\left(N_E(i)\right)}{\lambda_p^k \beta_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} n_I(i)} \\ \text{Var}\left(\hat{P}(k)\right) &= \text{Var}\left(\frac{\lambda_p^k \alpha_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} N_E(i)}{\lambda_p^k \beta_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} n_I(i)}\right) \\ &= \frac{\sum_{i=1}^k \lambda_p^{2(k-i)} \text{Var}\left(N_E(i)\right)}{\left(\lambda_p^k \beta_{p,0} + \sum_{i=1}^k \lambda_p^{k-i} n_I(i)\right)^2} \end{aligned}$$

Assume $n_I(i) = n_I = \text{constant}$. Since $N_E(i)$ is a Binomial random variable it follows that $\text{E}(N_E(i)) = n_I p$ and $\text{Var}(N_E(i)) = n_I p(1-p)$. Consequently,

$$\begin{aligned} \text{E}\left(\hat{P}(k)\right) &= \frac{\lambda_p^k \alpha_{p,0} + n_I p \sum_{i=1}^k \lambda_p^{k-i}}{\lambda_p^k \beta_{p,0} + n_I \sum_{i=1}^k \lambda_p^{k-i}} \\ \text{Var}\left(\hat{P}(k)\right) &= \frac{n_I p(1-p) \sum_{i=1}^k \lambda_p^{2(k-i)}}{\left(\lambda_p^k \beta_{p,0} + n_I \sum_{i=1}^k \lambda_p^{k-i}\right)^2} \end{aligned}$$

Recall the formula for a geometric sum given by $\sum_{j=0}^{k-1} x^j = (1-x^k)/(1-x)$, if $x \neq 1$. Using this formula, it follows that

$$\begin{aligned} \text{E}\left(\hat{P}(k)\right) &= \frac{\lambda_p^k \alpha_{p,0} + n_I p \frac{1-\lambda_p^k}{1-\lambda_p}}{\lambda_p^k \beta_{p,0} + n_I \frac{1-\lambda_p^k}{1-\lambda_p}} \\ \text{Var}\left(\hat{P}(k)\right) &= \frac{n_I p(1-p) \frac{1-\lambda_p^{2k}}{1-\lambda_p^2}}{\left(\lambda_p^k \beta_{p,0} + n_I \frac{1-\lambda_p^k}{1-\lambda_p}\right)^2} \end{aligned}$$

The steady-state behavior ($k \to \infty$) is obtained as

$$\begin{aligned} \lim_{k\to\infty} \text{E}\left(\hat{P}(k)\right) &= p \\ \lim_{k\to\infty} \text{Var}\left(\hat{P}(k)\right) &= \frac{p(1-p)(1-\lambda_p)^2}{n_I(1-\lambda_p^2)} \end{aligned}$$

Assume λ_p is close to one, and make use of the facts that $p \ll 1$ and $(1-\lambda_p^2) = (1+\lambda_p)(1-\lambda_p)$. Then,

$$\lim_{k\to\infty} \text{Var}\left(\hat{P}(k)\right) = \frac{p(1-p)(1-\lambda_p)^2}{n_I(1+\lambda_p)(1-\lambda_p)} \approx \frac{p(1-\lambda_p)}{2n_I}$$

With a slight abuse of notation, it follows

$$\sigma_{rel}^2(\infty) = \lim_{k\to\infty} \frac{1}{\text{E}(\hat{P}(k))^2}\text{Var}\left(\hat{P}(k)\right) \approx \frac{(1-\lambda_p)}{2n_I p}$$

Finally,

$$\sigma_{rel}(\infty) \approx \sqrt{\frac{(1-\lambda_p)}{2n_I p}}, \tag{7}$$

which is the relationship that shall be used for adaptation of the Bayesian estimator to be derived in Section III.

Figure 2 shows the relationship between $\sigma_{rel}(\infty)$ and

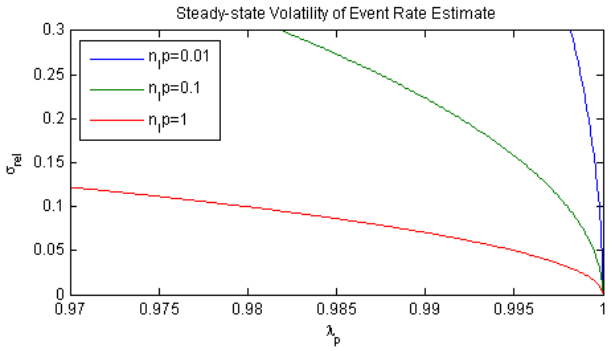

Fig. 2. The plot shows the relationship between $\sigma_{rel}(\infty)$ and λ_p for three different values of the product $n_I p$.

λ_p for three different values of the product $n_I p$. Given an estimate of $n_I p$ and a desired value of $\sigma_{rel}(\infty)$, we can determine the appropriate value of λ_p. Indeed, this is the basic idea of the adaptive event rate estimator proposed in this paper.

B. Impression Rate Estimation

The impression volume $n_I(k)$ is in practise not constant, which was an assumption in the derivation of the expression of the steady-state volatility of event rate estimate. Therefore, in order to use (7) to compute an appropriate λ_p we must compute an expected long-term estimate of the impression volume. The estimate should be free from high frequencies, but must not otherwise be of a certain accuracy since it is used only as a guidance for the adaptation.

We shall use a Bayesian framework for the impression rate estimation very similar to what was used in the derivation of the static event rate estimator. Since $n_I(k)$ represents traffic data a popular first order approximation is to assume it is generated from a Poisson process. Therefore, suppose $n_I(k)$ is a realization of a random variable $N_I(k)$, where $N_I(k) \sim$ Poisson(v), for some unknown impression rate v. Treat a possible seasonality in $n_I(k)$ as a disturbance to be filtered out.

Consider one observation n_I and leave out the time index k until later. The probability mass function of n_I based on the Poisson model is given by

$$f\left(n_I|v\right) \quad = \quad \frac{v^{n_I}e^{-v}}{n_I!}.$$

Given a prior belief function $\pi(v)$ for v, we proceed by using Bayes rule to determine a posterior $\pi(v|n_I)$ that incorporates the evidence of v encoded in n_I.

Since the conjugate prior of the Poisson probability mass function is given by the gamma probability density function, we assume the prior belief function $\pi(v)$ is given by Gamma(α_I, β_I). A standard application of Bayes rule $\pi\left(v|n_I\right) = f\left(n_I|v\right)\pi\left(v\right)/f\left(n_I\right)$, where $f\left(n_I\right)$ is independent of v, shows that $p|n_I \sim$ Gamma $\left(\alpha_I + n_I, \beta_I + 1\right)$. This update scheme applies recursively for each new measurement $n_I(k)$, $k = 1, 2, \ldots$.

Therefore

$$\begin{aligned} &v|n_I(1), n_I(2), \ldots, n_I(k) \\ &\quad \sim \quad \text{Gamma}\left(\alpha_{I,0} + \sum_{i=1}^{k} n_I(i), \beta_{I,0} + k\right) \end{aligned}$$

or equivalently by defining the shape and inverse scale parameters as a time-invariant Markov process with state $[\alpha_I(k), \beta_I(k)]^T$

$$\left[\begin{array}{c} \alpha_I(k) \\ \beta_I(k) \end{array}\right] \quad = \quad \left[\begin{array}{c} \alpha_I(k-1) \\ \beta_I(k-1) \end{array}\right] + \left[\begin{array}{c} n_I(k) \\ 1 \end{array}\right]$$

and initialized by $[\alpha_I(0), \beta_I(0)]^T = [\alpha_{I,0}, \beta_{I,0}]^T$, the posterior distribution of v at time k is given by Gamma $\left(\alpha_I(k), \beta_I(k)\right)$.

However, since the unknown v may vary over time, it is necessary to discount old observations and rely more on recent observations in the estimator. We do this by introducing a memory loss specified by a forgetting factor λ_I, where $0 < \lambda_I < 1$. In particular, we adopt the revised update scheme $\alpha_I(k) = \lambda_I^k \alpha_{I,0} + \sum_{i=1}^{k} \lambda_I^{k-i} n_I(i)$ and $\beta_I(k) = \lambda_I^k \beta_{I,0} + \sum_{i=1}^{k} \lambda_I^{k-i}$. This leads to the posterior distribution

$$\begin{aligned} &v|n_I(1:k) \sim \\ &\quad \text{Gamma}\left(\lambda_I^k \alpha_{I,0} + \sum_{i=1}^{k} \lambda_I^{k-i} n_I(i), \lambda_I^k \beta_{I,0} + \sum_{i=1}^{k} \lambda_I^{k-i}\right), \end{aligned}$$

or, in recursive form:

$$\begin{aligned} \alpha_I(k) \quad &= \quad \lambda_I \alpha_I(k-1) + n_I(k) \qquad (8) \\ \beta_I(k) \quad &= \quad \lambda_I \beta_I(k-1) + 1 \qquad (9) \end{aligned}$$

The posterior expected loss given a squared loss function of $L(v, \hat{v}) = (v - \hat{v}(k))^2$ is defined by

$$\begin{aligned} &\mathrm{E}^{\pi(v|n_I(1:k))}(L(v, \hat{v})) \\ &\quad = \quad \int_0^\infty (v - \hat{v}(k))^2 \pi\left(v|n_I(1:k)\right) dv \end{aligned}$$

Minimizing $\mathrm{E}^{\pi(v|n_I(1:k))}(L(v, \hat{v}))$ with respect to $\hat{v}(k)$ yields

$$\hat{v}(k) \quad = \quad \frac{\alpha_I(k)}{\beta_I(k)}. \qquad (10)$$

This incorporates all available evidence into the estimate and defines the optimal Bayesian impression rate estimate.

Equations (8)-(10) represents a recursive and constant computational cost estimator and is used in the adaptive estimator, but to assist in the selection of λ_I later let us explore the dynamics of this estimator in more detail.

It is trivial to show that the one-state-variable representation of the impression rate estimator is

$$\begin{aligned} \alpha_I(k) \quad &= \quad \lambda_I \alpha_I(k-1) + n_I(k) \\ \hat{v}(k) \quad &= \quad \frac{\alpha_I(k)}{\lambda_I^k \beta_{I,0} + \frac{1-\lambda_I^k}{1-\lambda_I}} \end{aligned}$$

which is a linear, time-varying dynamical system.

The steady-state dynamics, on the other hand, is

$$\begin{aligned} \alpha_I(k) \quad &= \quad \lambda_I \alpha_I(k-1) + n_I(k) \\ \hat{v}(k) \quad &= \quad (1-\lambda_I)\alpha_I(k) \end{aligned}$$

which is a linear, time-invariant (LTI) dynamical system. There is a comprehensive set of theory and tools to analyze LTI systems leading to general and very useful results. We are interested in the dynamical relationship between $n_I(k)$ and $\hat{v}(k)$.

For convenience, rewrite the two equations as a single input-output equation:

$$\begin{aligned} \hat{v}(k) \quad &= \quad (1-\lambda_I)\left(\lambda_I \alpha_I(k-1) + n_I(k)\right) \\ &= \quad \lambda_I(1-\lambda_I)\alpha_I(k-1) + (1-\lambda_I) n_I(k) \end{aligned}$$

In other words, $\hat{v}(k)$ represents a new state variable as

$$\hat{v}(k) \quad = \quad \lambda_I \hat{v}(k-1) + (1-\lambda_I) n_I(k). \qquad (11)$$

Hence, at steady-state the optimal Bayesian impression rate estimator is simply a moving average of the observed impression volume. Thanks to nice properties of LTI systems, a wealth of insights are obtained by investigating the unit step response, the unit impulse response, and the frequency response of (11).

C. Adaptive Estimation Algorithm

The proposed adaptive event rate estimator consists of three building blocks. The impression rate estimator implements (8)-(10) and produces an estimate of the average number of impressions per hour. Importantly, to operate well with the event rate estimator tuner, which is the second building block, this estimate must not contain significant high-frequent noise. The event rate estimator tuner leverages on (7), but with v in place of n_I since the volatility of the event rate estimate in Section V-A was derived under the premise of steady-state conditions and constant n_I. In other words,

$$\sigma_{rel}(\infty) \approx \sqrt{\frac{(1-\lambda_p)}{2pv}}. \quad (12)$$

Assume a desired steady-state relative standard deviation of the event rate estimate $\sigma_{rel}(\infty)$ is provided and denoted σ_{ref}. Given an impression rate estimate computed at time k and an event rate estimate computed at time $k-1$, we may use (12) to compute λ_p as

$$\lambda_p(k) = 1 - 2\hat{p}(k-1)\hat{v}(k)\sigma_{ref}^2$$

The third and final building block of the adaptive event rate estimator implements (3)-(5) and consumes $\lambda_p(k)$.

Figure 3 shows a block diagram of the integrated adaptive event rate estimator and illustrates how the subsystems interact.

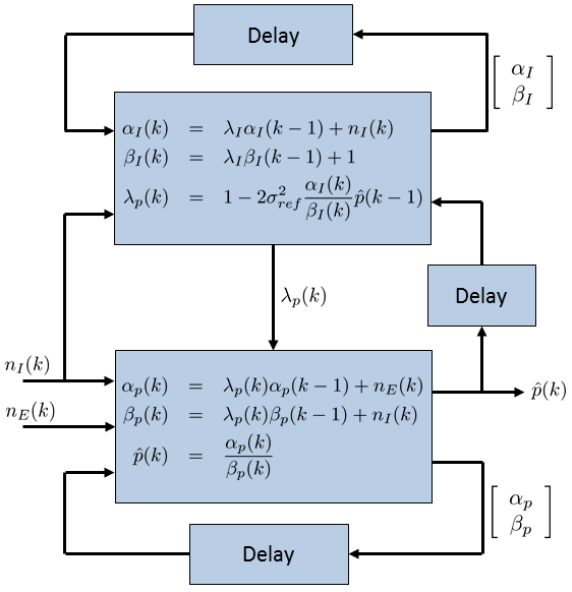

Fig. 3. Block diagram of the integrated adaptive event rate estimator.

To conclude this section, and to make it easy to implement the derived results, the final estimation algorithm is given by the following process:

Initialization: For $k = 0$, set

$$\begin{aligned}\alpha_I(0) &= \alpha_{I,0}\\ \beta_I(0) &= \beta_{I,0}\\ \alpha_p(0) &= \alpha_{p,0}\\ \beta_p(0) &= \beta_{p,0}\end{aligned}$$

Computation For $k = 1, 2, \ldots$:

State estimate propagation

$$\begin{aligned}\alpha_I(k) &= \lambda_1\alpha_I(k-1) + n_I(k)\\ \beta_I(k) &= \lambda_1\beta_I(k-1) + 1\\ \lambda_p(k) &= 1 - 2\sigma_{ref}^2\frac{\alpha_I(k)}{\beta_I(k)}\frac{\alpha_p(k-1)}{\beta_p(k-1)}\\ \alpha_p(k) &= \lambda_p(k)\alpha_p(k-1) + n_E(k)\\ \beta_p(k) &= \lambda_p(k)\beta_p(k-1) + n_I(k)\end{aligned}$$

Output calculation

$$\hat{p}(k) = \frac{\alpha_p(k)}{\beta_p(k)}$$

As a final remark, in practical implementations it is recommended to add a saturation to the computation of $\lambda_p(k)$ to prevent too sluggish and too aggressive updates of the event rate estimate (and to consider that the volatility of the event rate estimate in Section V-A was derived under the assumption λ_p is close to one). That is, instead of $\lambda_p(k) = 1 - 2\sigma_{ref}^2\hat{v}(k)\hat{p}(k-1)$, where $\hat{v}(k) = \alpha_I(k)/\beta_I(k)$ and $\hat{p}(k-1) = \alpha_p(k-1)/\beta_p(k-1)$, use

$$\begin{aligned}\lambda_p'(k) &= 1 - 2\sigma_{ref}^2(k)\hat{v}(k)\hat{p}(k-1)\\ \lambda_p(k) &= \begin{cases}\lambda_{p,min}, & \lambda_p'(k) < \lambda_{p,min}\\ \lambda_p'(k), & \lambda_{p,min} \le \lambda_p'(k) \le \lambda_{p,max}\\ \lambda_{p,max}, & \lambda_p'(k) > \lambda_{p,max}\end{cases}\end{aligned}$$

for some provided values of $\lambda_{p,min}$ and $\lambda_{p,max}$.

VI. Example: Adaptive Event Rate Estimation in Online Advertising

Consider the simulated scenario in Section IV, but let us now explore how the adaptive event rate estimator performs and how it compares with the static estimator. Besides the observed impression and event data, the adaptive estimator needs values for λ_I and σ_{ref}.

To select λ_I for the online advertising use case we should opt for a value that produces an estimate $\hat{v}(k)$ without significant high-frequent noise and time-of-day seasonality, and that is not strongly dependent on the average impression volume level in this particular example. Indeed, the estimator is most powerful if it works well for a wide range of impression volume levels. Fortunately, the steady-state dynamics of the impression rate estimator is LTI which makes it easy to analyze. Moreover, while the average impression volume and the phase of the seasonality vary a lot from one ad to another, the relative magnitude of the time-of-day seasonality

is approximately 80% of the average impression volume for the majority of ads. The linear and time-invariant dynamics defined by (11) together with the well-understood seasonality makes it easy to select λ_I systematically by trading e.g. the settling time of the response to a unit step and unit impulse input and the frequency response to harmonic input signals with relevant frequencies (most importantly the frequency corresponding to the 24 hours periodicity). Note, if the dynamics $\hat{v}(k)$ was nonlinear or time-varying it would not be a trivial task to select λ_I. After considering the unit step, impulse, and frequency responses we use $\lambda_I = 0.99$ in the remainder of this example.

The choice of σ_{ref} is mostly a matter or preference, but it is important to understand that it represents a desired steady-state value and chosen very small the estimator necessarily becomes sluggish and leads to an event rate estimate of limited value. Let us first consider $\sigma_{ref} = 0.14$.

Figure 4 compares the performance of the two static

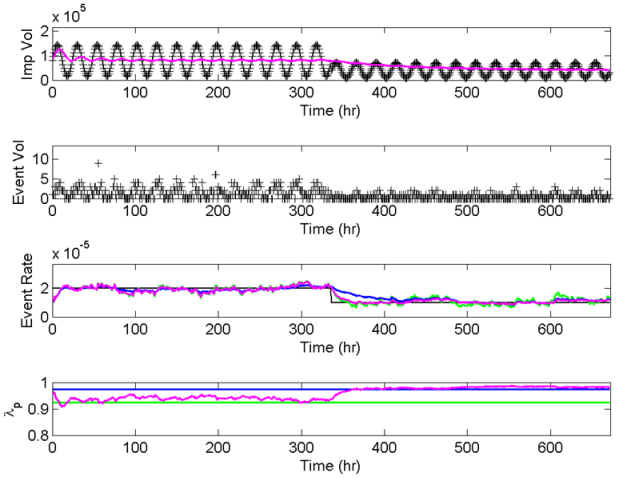

Fig. 4. Comparison between the adaptive event rate estimator with $\lambda_I = 0.99$ and $\sigma_{ref} = 0.14$ (pink), and the static event rate estimator with $\lambda_p = 0.925$ (green) and $\lambda_p = 0.975$ (blue), respectively.

estimators in Section IV with the adaptive estimator with $\lambda_I = 0.99$ and $\sigma_{ref} = 0.14$. Note how the adaptive estimator behaves approximately like the more responsive static estimator ($\lambda_p = 0.925$) when $n_I = 80,000$ and $p = 10^{-5}$ and approximately like the more insensitive estimator ($\lambda_p = 0.975$) when $n_I = 40,000$ and $p = 5 \cdot 10^{-4}$. As a result, the estimate responds rapidly to the regime change after two weeks yet stiffens up quickly to avoid a too volatile estimate in the last two weeks of the scenario.

Now, let us compare the adaptive estimator for three different values of σ_{ref} (and with $\lambda_I = 0.99$). Figure 5 shows how the event rate estimate $\hat{p}(k)$ and the computed $\lambda_p(k)$ differ when σ_{ref} is 0.08, 0.14, and 0.2. Inspect the graphs and pay attention to how the volatility and the response time at the regime shift differ among the three estimates. For $\sigma_{ref} = 0.08$ the relative volatility is small (as expected), but it takes well over hundred hours before the estimate settles down after the regime shift. On the other hand, for $\sigma_{ref} = 0.2$ the response time at the regime

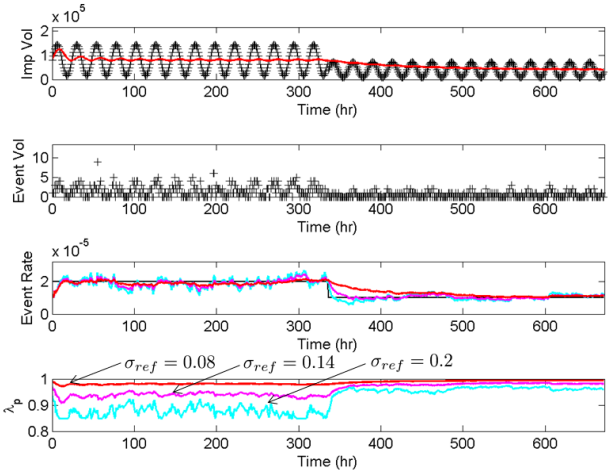

Fig. 5. Comparison of the adaptive event rate estimator for $\sigma_{ref} = 0.08$, 0.14, and 0.2, with $\lambda_I = 0.99$.

shift is very short while the the relative volatility throughout the scenario is quite large. With $\sigma_{ref} = 0.14$ the relative volatility is relatively small yet the response time to the regime shift is fast.

VII. Conclusions

We have made use of tools from Bayesian statistics and from dynamical systems to develop an adaptive event rate estimator. The work was motivated by needs in online advertising where event rates typically are very small and where the same estimation algorithms (observers) must work well across a wide range of scenarios. In some scenarios the unknown event rate and/or impression rate are order of magnitudes larger than in other scenarios.

The proposed algorithm meets the above requirements and is applicable to any problem dealing with the estimation of small event rates.

Due to space limitations, this paper only provides simulated results of the proposed estimator. However, the algorithm has been evaluated and will be evaluated further on experimental data and integrated as an adaptive observer within a closed loop feedback control system.

References

[1] Niklas Karlsson and Jianlong Zhang. Applications of feedback control in online advertising. *Proceedings of the 2013 American Control Conference*, pages 6008–6013, 2013.

[2] Niklas Karlsson. Adaptive control using Heisenberg bidding. *Proceedings of the 2014 American Control Conference*, pages 1304–1309, 2014.

[3] Niklas Karlsson. Control problems in online advertising and benefits of randomized bidding strategies. *Submitted to European Journal of Control*, 2016.

[4] Simon Haykin. *Kalman Filtering and Neural Networks*. John Wiley & Sons, Inc., 2001.

[5] James O. Berger. *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*. Springer-Verlag, 1985.

[6] Niklas Karlsson. *Systems and Methods For Adaptive Event Rate Estimation, United States Patent Application (AOLI.240495)*. USPTO, 2015.

[7] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, second edition, 2001.