Low Complexity Sequential Probability Estimation and Universal Compression for Binary Sequences with Constrained Distributions

Citation for published version (APA):
Shamir, G. I., Tjalkens, T. J., & Willems, F. M. J. (2009). Low Complexity Sequential Probability Estimation and Universal Compression for Binary Sequences with Constrained Distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium, Toronto, Ontario, Canada, 06 - 11 July 2008* (pp. 995-999). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ISIT.2008.4595136

DOI:
10.1109/ISIT.2008.4595136

Document status and date:
Published: 01/01/2009

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Low-Complexity Sequential Probability Estimation and Universal Compression for Binary Sequences with Constrained Distributions

Gil I. Shamir, Tjalling J. Tjalkens, and Frans M. J. Willems

Abstract— **Two low-complexity methods are proposed for sequential probability assignment for binary independent and identically distributed (i.i.d.) individual sequences with empirical distributions whose governing parameters are known to be bounded within a limited interval. The methods can be applied to different problems where fast accurate estimation of the maximizing sequence probability is very essential to minimizing some loss. Such applications include applications in finance, learning, channel estimation and decoding, prediction, and universal compression. The application of the new methods to universal compression is studied, and their universal coding redundancies are analyzed. One of the methods is shown to achieve the minimax redundancy within the inner region of the limited parameter interval. The other method achieves better performance on the region boundaries and is more robust numerically to outliers. Simulation results support the analysis of both methods. While non-asymptotically the gains may be significant over standard methods that maximize the probability over the complete parameter simplex, asymptotic gains are in second order. However, these gains translate to meaningful significant factor gains in other applications, such as financial ones. Moreover, the methods proposed generate estimators that are constrained within a given interval throughout the complete estimation process which are essential to applications such as sequential binary channel crossover estimation. The results for the binary case lay the foundation to studying larger alphabets.**

I. Introduction

Universal sequence probability assignment and sequence probability estimation are important in applications in finance, learning, channel estimation, prediction, universal compression, and more. The goal is to assign probability as large as possible to a sequence, whose governing parameters under a known governing statistical model are unknown in advance. Classical universal sequential probability assignment methods (see, e.g., [3], [9]) assign such a probability under the assumption that the governing parameters can be at any point in the complete parameter simplex. Averaging over the complete parameter space with some weighting *prior* gives simple add-constant estimators, such as the add-$1/2$ *Krichesvky-Trofimov* (KT) estimator [3]. Such estimators give each symbol a constant number of occurrences prior to the start of the sequence.

In many cases, there may exist some advance knowledge that indicates that the governing parameters can be only inside a subset of the parameter space. The use of such knowledge can reduce losses attained due to lack of prior knowledge of the actual governing parameters. Consider, for example, a binary independent and identically distributed (i.i.d.) sequence for which it is known that the *maximum likelihood* (ML) estimate of a 1 bit $\hat{\psi}$ is within a limited interval $[\alpha, \beta] \subset [0, 1]$. In [2], the *minimax* universal coding redundancy (for the best code and worst sequence) was derived for this case, and was shown to reduce from the standard case. Designing sequential estimators that average only over a subset of the parameter space, however, is more complicated than the standard case.

In this paper, we consider the simple binary i.i.d. case described as a basis to a more general case. We design low-complexity probability assignment methods for a sequence y^n whose unknown ML parameter $\hat{\psi}$ is known to be inside the interval $[\alpha, \beta]$. We then bound the universal compression redundancy obtained by these schemes and show the gains that can be attained over the standard methods. These gains asymptotically reduce the second order of the redundancy. However, they can be significant for shorter data blocks. Furthermore, they can accumulate to large gains with larger alphabets if the source parameters are described by *decomposing* the parameters into binary trees. When compressing sources with memory with an algorithm such as the *context tree weighting* (CTW) [9], the statistics in each state of the source are of an i.i.d. source. If gains are achieved for each state, they can accumulate to large overall gains in practice.

Gains may extend well beyond compression to applications in prediction, estimation, universal investment portfolios [1], and more. While the loss in compression is logarithmic in the ratio between the maximizing probability and the assigned one (i.e., the attenuation of the maximizing probability by the estimator), other loss functions may be linear in this attenuation. A single bit gain in compression reflects a factor of 2 gain in this ratio. Consider a process constantly selecting reinvestment between two investment types. With some probability one investment will double, while the other will be lost. With the remaining probability, the opposite outcome will take place. Universal compression redundancy gain of 3 bits is equivalent to an increase in wealth here by a factor of 8.

Unlike the standard KT estimator, the initial estimates of the new estimators are already biased in the proper direction, leading to earlier convergence to the maximizing probability and to the gain in performance. Some applications, such as crossover probability estimation of a *binary symmetric channel*

[1] G. I. Shamir is with ECE Department, University of Utah, Salt Lake City, UT 84112, U.S.A., e-mail: gshamir@ece.utah.edu. T. J. Tjalkens and F. M. J. Willems are with the Eindhoven University of Technology, Electrical Engineering Department, 5600 MB Eindhoven, The Netherlands, e-mails: T.J.Tjalkens@tue.nl, F.M.J.Willems@tue.nl. The work of the first author was partially supported by NSF Grant CCF-0347969.

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

(BSC), cannot tolerate estimators outside some known interval, which may lead to catastrophic performance.

Two methods are proposed for the sequential estimator. The first directly mixes over the limited parameter space with a normalized truncated Dirichlet-$1/2$ prior. Over the complete interval, this prior gives the KT estimator. The second addresses the bounded parameter interval as that of a parameter that results from passing a sequence x^n generated with a parameter $\theta \in [0, 1]$ through a noisy binary channel to generate y^n (see, e.g., [6], [7]). The estimator attempts to estimate the parameter of the "clean" sequence and transform it to the noisy sequence.

II. Notation and Preliminaries

Let $y^n \stackrel{\triangle}{=} (y_1, y_2, \ldots, y_n)$ be a sequence of n i.i.d. bits, consisting of $n_1(y^n)$ 1 bits and $n_0(y^n)$ 0 bits. Its ML estimate of the probability of 1 is $\hat{\psi} = n_1(y^n)/n$. It is assumed that $\hat{\psi} \in [\alpha, \beta]$, $0 \le \alpha < \beta \le 1$, where α and β are known in advance. The ML probability of y^n is given by

$$P_{\hat{\psi}}(y^n) = \hat{\psi}^{n_1(y^n)} \cdot (1-\hat{\psi})^{n_0(y^n)}. \tag{1}$$

The *individual* sequence redundancy of a code that assigns probability $Q(y^n)$ to y^n is given by [1]

$$R_n(Q, y^n) \stackrel{\triangle}{=} \log P_{\hat{\psi}}(y^n) - \log Q(y^n). \tag{2}$$

The individual *minimax* redundancy of a class Λ is that of the best code for the worst sequence that can be produced by the class. The minimax redundancy for the class $\mathcal{B}_{\alpha,\beta}$ of binary i.i.d. sequences whose governing parameter is constrained to the interval $[\alpha, \beta]$ was computed in [2], and shown to be [2]

$$R_n(\mathcal{B}_{\alpha,\beta}) = \frac{1}{2}\log n + \log C_{\alpha,\beta} - \frac{1}{2}\log\frac{\pi}{2} + O\left(n^{-1/2}\right), \tag{3}$$

where

$$C_{\alpha,\beta} = \int_\alpha^\beta \frac{0.5dx}{\sqrt{x(1-x)}} = \left(\sin^{-1}\sqrt{\beta} - \sin^{-1}\sqrt{\alpha}\right). \tag{4}$$

The minimax redundancy derivation allows for sequences y^n for which $\hat{\psi} \notin [\alpha, \beta]$. The ML estimator $\hat{\psi}^*$ for such sequences must still be constrained such that $\hat{\psi}^* \in [\alpha, \beta]$. Thus if $\hat{\psi} \le \alpha$ then $\hat{\psi}^* = \alpha$ and if $\hat{\psi} > \beta$ then $\hat{\psi}^* = \beta$. Here, we only consider $\hat{\psi} \in [\alpha, \beta]$.

In the special case of $[\alpha, \beta] = [0, 1]$, $C_{0,1} = \pi/2$, yielding

$$R_n(\mathcal{B}_{0,1}) = \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + O\left(n^{-1/2}\right). \tag{5}$$

Practical probability assignments for this case can be obtained by *mixing* (averaging) the sequence probability over the complete parameter space with some prior $\omega(\psi)$ that integrates to 1 over this space. This gives a sequence y^n probability

$$Q(y^n) = \int_0^1 \omega(\psi)\psi^{n_1(y^n)}(1-\psi)^{n_0(y^n)}d\psi. \tag{6}$$

A uniform prior gives the well-known add-1 Laplace estimator. While this estimator attains good redundancy in the inner part of the interval, it fails to perform well in the boundaries (around 0 and 1). A *Dirichlet*-$1/2$ (beta) prior, given by

$$\omega(\psi) = \frac{1}{\pi\sqrt{\psi(1-\psi)}} \tag{7}$$

gives the well-known add-$1/2$ KT estimator [3], which can be assigned to y^n sequentially. The KT estimator is initialized to $Q(y^0) = 1$, and is updated by

$$Q\left(y^{t+1}\right) = Q\left(y^t\right) \cdot \frac{n_{y_{t+1}}\left(y^t\right) + 0.5}{t+1} \tag{8}$$

where $n_{y_{t+1}}\left(y^t\right)$ is the occurrence count of bit y_{t+1} in the prefix sequence y^t.

The KT estimator performs more uniformly over the interval $[0, 1]$, but is yet not minimax optimal (see, e.g., [11]) in second order due to losses that still occur in the boundaries. Specifically, in the binary case, it achieves asymptotic redundancy

$$R_n(Q_{KT}, y^n) \le \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + o(1) \tag{9}$$

if $\hat{\psi} \in (n^{-\varepsilon}, 1 - n^{-\varepsilon})$ for an arbitrarily small $\varepsilon > 0$. Otherwise,

$$R_n(Q_{KT}, y^n) \le \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + \frac{1}{12}\log e + O\left(\frac{1}{n}\right) \tag{10}$$

as long as $0 < \hat{\psi} < 1$. Finally,

$$R_n(Q_{KT}, y^n) \le \frac{1}{2}\log n + \frac{1}{2}\log\pi + O\left(\frac{1}{n}\right) \tag{11}$$

for $\hat{\psi} = 0$ or $\hat{\psi} = 1$. In [9], it was shown that even for small n, $R_n(Q_{KT}, y^n)$ is guaranteed not to exceed $0.5\log n + 1$.

III. Method I: Scaled Cut off Dirichlet-$1/2$ Prior

To derive a sequential probability estimate within $[\alpha, \beta]$, we can cut off the Dirichlet-$1/2$ prior to the interval $[\alpha, \beta]$ and scale the resulting prior. This leads to

$$Q(y^n) = \int_\alpha^\beta \frac{1}{2C_{\alpha,\beta}\sqrt{\psi(1-\psi)}}\psi^{n_1(y^n)}(1-\psi)^{n_0(y^n)}d\psi. \tag{12}$$

The constant $C_{\alpha,\beta}$ results from the scaling. It is given in (4) and guarantees that the prior integrates to 1 over $[\alpha, \beta]$.

Theorem 1: The probability assigned to y^n in (12) can be computed sequentially by an initialization step $Q\left(y^0\right) = 1$, and an update step,

$$\begin{aligned} Q\left(y^{t+1}\right) = Q\left(y^t\right) \cdot \frac{n_{y_{t+1}}\left(y^t\right) + 0.5}{t+1} &+ \\ (2y_{t+1} - 1) \cdot \frac{\alpha^{n_1(y^t)+0.5}(1-\alpha)^{n_0(y^t)+0.5}}{2C_{\alpha,\beta} \cdot (t+1)} &- \\ (2y_{t+1} - 1) \cdot \frac{\beta^{n_1(y^t)+0.5}(1-\beta)^{n_0(y^t)+0.5}}{2C_{\alpha,\beta} \cdot (t+1)} &. \end{aligned} \tag{13}$$

Note that the KT estimator is a special case of the above sequential assignment with $[\alpha = 0, \beta = 1]$. Specifically, in that case, $C_{\alpha,\beta} = \pi/2$, and (13) reduces to the binary form of

[1] The logarithm function is taken to the base of 2. We ignore integer length constraints, and treat $-\log Q(y^n)$ as the code length.

[2] For two functions $f(n)$ and $g(n)$, $f(n) = o(g(n))$ if $\forall c, \exists n_0$, such that, $\forall n > n_0$, $|f(n)| < c\,|g(n)|$; $f(n) = O(g(n))$ if $\exists c, n_0$, such that, $\forall n > n_0$, $|f(n)| \le c\,|g(n)|$.

the KT estimator in (8). The proof of Theorem 1 is presented in [6] and [8] and is based on integration by parts and the fact that $Q\left(y^t\right) = Q\left(y^t 0\right) + Q\left(y^t 1\right)$, where the latter two denote concatenation of 0 and 1, respectively, to the string y^t.

Theorem 1 derives a limited interval version of the KT estimator. A similar approach can be taken with a uniform prior, yielding a limited interval version of the Laplace estimator.

Theorem 2: Fix ε arbitrarily small, and let n be sufficiently large. Let $\hat{\psi} = n_1(y^n)/n \in [\alpha, \beta]$ be the ML estimator of a sequence y^n. Define $\varepsilon_1 \stackrel{\triangle}{=} 1/\sqrt{n}^{1-\varepsilon}$. Then,

$$R_n\left(Q, y^n\right) \leq \frac{1}{2}\log n + \log C_{\alpha,\beta} - \frac{1}{2}\log\frac{\pi}{2} + o(1) \qquad (14)$$

for $\hat{\psi} \in [\alpha + \varepsilon_1, \beta - \varepsilon_1]$. Second,

$$R_n\left(Q, y^n\right) \leq \frac{1}{2}\log n + \log C_{\alpha,\beta} - \frac{1}{2}\log\frac{\pi}{8} + o(1) \qquad (15)$$

for $\hat{\psi} \in [\alpha, \alpha + \varepsilon_1]$ or $\hat{\psi} \in [\beta - \varepsilon_1, \beta]$ where in both cases $1/n^{1-2\varepsilon} \leq \hat{\psi} \leq 1 - 1/n^{1-2\varepsilon}$. Finally,

$$R_n\left(Q, y^n\right) \leq \frac{1}{2}\log n + \log C_{\alpha,\beta} - \frac{1}{2}\log\frac{\pi}{4} + o(1) \qquad (16)$$

for $\hat{\psi} \notin \left[1/n^{1-2\varepsilon}, 1 - 1/n^{1-2\varepsilon}\right]$.

Theorem 2 shows that the sequential estimator of Theorem 1 asymptotically achieves the minimax redundancy in (5) in the inner part of the interval $[\alpha, \beta]$. At the boundaries of the interval, there is a penalty of 1 bit, unless the interval boundary is close to either 0 or 1. In the latter case, a lower penalty above the minimax redundancy in (5) of $1/2$ bit is obtained.

The bounds of (14) and (16) reduce to the respective asymptotic bounds of the KT estimator for $\hat{\psi} \in [0, 1]$. The new estimator gains (a reduction of) $\log(\pi/2) - \log C_{\alpha,\beta}$ bits over the KT estimator. The gain is reduced in inner boundaries because the mixture does not include the other side of the boundary. The universal gains over standard KT encoding shown in Theorem 2 are in second order performance. As shown in the numerical results in Section V, these gains are essential for moderate to short block sizes. However, the universal compression gains can translate in other applications to significant factors of probability estimator attenuation gains.

The proof of Theorem 2 is rather complicated and is presented in [8]. The idea is to compute the redundancy as a difference of logarithms, and insert $P_{\hat{\psi}}\left(y^n\right)$ into the kernel integral. Then, the integration interval is reduced, such that any point within the integral is asymptotically in the vicinity of $\hat{\psi}$. This allows approximations that bring the integral into one over a Gaussian distribution. The integration interval is carefully designed, so that the integral approaches 1 for the inner part of the interval, and $1/2$ at the boundaries. Adjusting constants, the redundancy bounds are obtained. Boundary bounds plotted in Section V are more precise than (15). A different approach is taken for $\alpha = 0$ or $\beta = 1$.

IV. Method II: Transformed Dirichlet-1/2 Prior

The sequential estimator in Theorem 1 appears to be the generalization of the KT estimator for a limited parameter interval, and has similar properties with respect to minimax performance in its parameter space. It thus looses in performance at the boundaries. For specific values of α, β, and n, it may be possible to obtain more uniform performance with a different estimator.

A bigger problem of the estimator in Theorem 1 is its numerical robustness. Unlike sequential estimators based on the standard approach (see, e.g., [3], [4], [5], [9], [10]) which may generate several probability estimators and add them to provide $Q\left(y^t\right)$, the estimator of Theorem 1 adds but may also subtract a bias from a quantity updated sequentially. The sign of the bias depends on the actual bits in y^n. Subtraction of very small biases from very small probabilities can lead to lack of numerical stability, resulting in inaccurate probability estimators, including negative estimates. This problem is enhanced when the actual $\hat{\psi}$ is outside the assumed interval $[\alpha, \beta]$. This leads to the necessity of a more standard approach estimator.

As shown in [6], [7], one can view a sequence y^n governed by $\psi \in [\alpha, \beta]$ as a noisy version of a "clean" sequence x^n governed by $\theta \in [0, 1]$. The clean sequence is transformed through a binary channel with $P(Y = 1|X = 0) = p$ and $P(Y = 0|X = 1) = q$ to produce the noisy one, where capital letters denote random variables. This setting implies that

$$\psi = (1-\theta)p + \theta(1-q) \;\Leftrightarrow\; \theta = \frac{\psi - p}{1 - q - p} \qquad (17)$$

The relation between α, β and p, q is $\alpha = p$ and $\beta = 1 - q$. Using (17), a Dirichlet-1/2 prior over θ transforms to

$$\omega(\psi) = \frac{1}{\pi\sqrt{(\psi - p)(1 - q - \psi)}} = \frac{1}{\pi\sqrt{(\psi - \alpha)(\beta - \psi)}}. \qquad (18)$$

Alternatively, a probability can be assigned to y^n by assigning it first to x^n and transforming x^n over the channel. Due to the stochastic nature of the channel, however, a sequence y^n can result from all possible sequences x^n with the proper bits inverted. Hence, the assignment of $Q\left(y^n\right)$ is a sum of mixtures. For every possible x^n, a mixture over the parameter θ is performed. Then, assignments over all possible x^n are summed together with proper weights. Each $Q\left(x^n\right)$ is weighted by the probability that x^n transforms to the given y^n. For simplicity, let $a = n_0(y^n)$ and $b = n_1(y^n)$. For a specific pair x^n and y^n, use $\nu_{00} = \nu_{00}(x^n, y^n)$, $\nu_{01} = \nu_{01}(x^n, y^n)$, $\nu_{10} = \nu_{10}(x^n, y^n)$, and $\nu_{11} = \nu_{11}(x^n, y^n)$ to denote the joint occurrence count of the subscript pair in (x^n, y^n). The conditional probability that y^n is produced at the output of the channel with input x^n is given by

$$P\left(y^n|x^n\right) = (1-p)^{\nu_{00}} p^{\nu_{01}} (1-q)^{\nu_{11}} q^{\nu_{10}}. \qquad (19)$$

With prior $\omega(\theta)$,

$$Q\left(x^n\right) = \int_0^1 \omega(\theta)(1-\theta)^{\nu_{00}+\nu_{01}} \theta^{\nu_{10}+\nu_{11}} d\theta \qquad (20)$$

and the probability assigned to y^n is given by

$$Q\left(y^n\right) = \sum_{x^n} Q\left(x^n\right) P\left(y^n|x^n\right). \qquad (21)$$

Theorem 3: Let $\omega(\theta)$ be the Dirichlet-1/2 prior over $[0,1]$ given in (7). Then, the assignment in (21) satisfies

$$Q\left(y^n\right) = \int_\alpha^\beta \frac{\psi^{n_1(y^n)}(1-\psi)^{n_0(y^n)}}{\pi\sqrt{(\psi-\alpha)(\beta-\psi)}} \cdot d\psi. \tag{22}$$

Theorem 3 shows that mixing the probability assigned to x^n over θ and transforming x^n to y^n is identical to directly mixing the probability assigned to y^n using the prior over ψ in (18) that results from mapping θ to ψ.

Proof: Observing that $\nu_{00}+\nu_{10}=a$ and $\nu_{01}+\nu_{11}=b$ and that for a given sequence y^n, there are precisely $\binom{a}{\nu_{00}}\binom{b}{\nu_{01}}$ sequences x^n that together with y^n have the joint composition $(\nu_{00},\nu_{01},\nu_{10},\nu_{11})$, it follows that

$$\begin{aligned} Q\left(y^n\right) &= \sum_{x^n}\int_0^1 \omega(\theta)\{(1-\theta)(1-p)\}^{\nu_{00}}\{(1-\theta)p\}^{\nu_{01}} \\ &\quad \cdot\{\theta q\}^{\nu_{10}}\{\theta(1-q)\}^{\nu_{11}}d\theta \\ &= \int_0^1 \omega(\theta)\sum_{v=0}^{a}\sum_{w=0}^{b}\binom{a}{v}\binom{b}{w}\{(1-\theta)(1-p)\}^v \\ &\quad \cdot\{(1-\theta)p\}^w\{\theta q\}^{a-v}\{\theta(1-q)\}^{b-w}d\theta \\ &= \int_0^1 \omega(\theta)\{(1-\theta)(1-p)+\theta q\}^a \\ &\quad \cdot\{(1-\theta)p+\theta(1-q)\}^b d\theta \end{aligned} \tag{23}$$

Substituting the Dirichlet-1/2 prior to $\omega(\theta)$, changing variables following (17), recalling that $a=n_0(y^n)$, $b=n_1(y^n)$, $\alpha=p$, and $\beta=1-q$, (23) yields (22). ■

It remains to show how (21) can be implemented with a low-complexity sequential algorithm. This can be done using a state transition diagram which resembles those proposed in [4], [5], [10]. A state s at time t represents the composite (type) of all sequences x^t with equal empirical distributions. It will be denoted by $n_1(x^t)$ for all x^t leading to s. Therefore, there are $t+1$ states $s=0,1,\ldots,t$ at time t. Each state is assigned a weight

$$G_t\left(s,y^t\right) = \sum_{x^t:n_1(x^t)=s} Q\left(x^t\right)P\left(y^t \mid x^t\right) \tag{24}$$

that is the contribution of its type to $Q\left(y^t\right)$. Then,

$$Q\left(y^t\right) = \sum_{s=0}^{t} G_t\left(s,y^t\right). \tag{25}$$

State weights are updates sequentially. Initially, only $s=0$ exists, and its weight is initialized by $G_0\left(s,y^0\right)=1$. At any t, $G_t\left(s,y^t\right)=0$ for all $s<0$ or $s>t$, by definition. Then, for every $s=0,1,\ldots,t$, the following update is performed at time t,

$$\begin{aligned} &G_t\left(s,y^t\right) \\ &= \left[(1-p)(1-y_t)+py_t\right]\cdot\frac{t-s-0.5}{t}\cdot G_{t-1}\left(s,y^{t-1}\right)+ \\ &\quad \left[(1-q)y_t+q(1-y_t)\right]\cdot\frac{s-0.5}{t}\cdot G_{t-1}\left(s-1,y^{t-1}\right). \end{aligned} \tag{26}$$

After updating all existing states at time t, (25) is used to update $Q\left(y^t\right)$. The idea is that regardless of y_t, each state s, $0<s<t$, can be entered either from itself, by $x_t=0$, or from $s-1$ if $x_t=1$. State $s=0$ can only be entered from itself with $x_t=0$, and state $s=t$ only from $t-1$ by $x_t=1$. The first term in each component of the sum in (26) gives $P(y_t|x_t)$ for the proper state transition (either from s to s or from $s-1$ to s). The second term is the KT probability of x_t, which implements the mixture over θ. Figure 1 illustrates a transition diagram. The updates of the first terms in the products in (26) are denoted on the transitions.

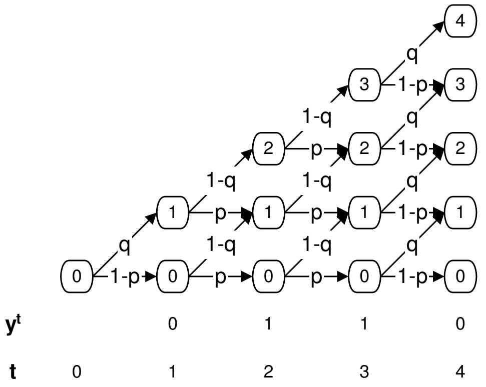

Fig. 1: State transition diagram for the probability assignment in (25)-(26) for the sequence $y^4=0110$.

Unlike the fixed per-symbol complexity assignment of Theorem 1, the method in (25)-(26) has linear per-symbol complexity (quadratic overall). However, on the other hand, it is numerically more robust, because no subtractions are performed. It is possible to lower the complexity by keeping only a small fraction of *surviving* states in the diagram, consisting of $s=n_1(x^t)$, for which $|n_1(x^t)/t-\theta|\leq\delta$, where θ is the transformed value of $\psi=n_1(y^t)/t$ in (17). The reduction of complexity using this method is beyond the scope of this paper, but is studied in future work.

The asymptotic redundancy achieved by the probability assignment in (25)-(26) is summarized below

Theorem 4: Fix ε arbitrarily small, and let n be sufficiently large. Let $\hat{\psi}=n_1(y^n)/n\in[\alpha,\beta]$ be the ML estimator of a sequence y^n. Define $\varepsilon_1 \overset{\triangle}{=} 1/\sqrt{n}^{1-\varepsilon}$. Then,

$$\begin{aligned} R_n\left(Q,y^n\right) &\leq \frac{1}{2}\log n+\frac{1}{2}\log\frac{\pi}{2}+\frac{1}{2}\log\left(1-\frac{\alpha}{\hat{\psi}}\right) \\ &\quad +\frac{1}{2}\log\left(1-\frac{1-\beta}{1-\hat{\psi}}\right)+o(1) \end{aligned} \tag{27}$$

for $\hat{\psi}\in[\alpha+\varepsilon_1,\beta-\varepsilon_1]$. For $\hat{\psi}\to\alpha>0$,

$$R_n\left(Q,y^n\right)\leq\frac{1+\varepsilon}{4}\log n+\frac{1}{2}\log(2\pi)+\frac{1}{2}\log\frac{\beta-\alpha}{\sqrt{\alpha(1-\alpha)}}+o(1) \tag{28}$$

and for $\hat{\psi}\to\beta<1$,

$$R_n\left(Q,y^n\right)\leq\frac{1+\varepsilon}{4}\log n+\frac{1}{2}\log(2\pi)+\frac{1}{2}\log\frac{\beta-\alpha}{\sqrt{\beta(1-\beta)}}+o(1). \tag{29}$$

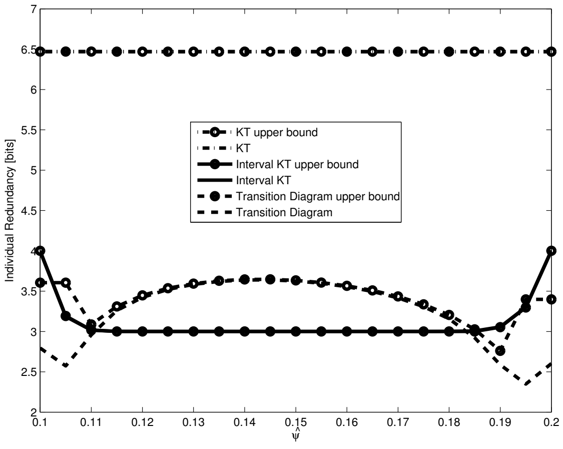

Fig. 2: Individual sequence redundancy for the KT estimate and the two sequential estimators for bounded intervals for $n = 5000$, $\psi \in [0.1, 0.2]$ and the same range of $\hat{\psi}$.

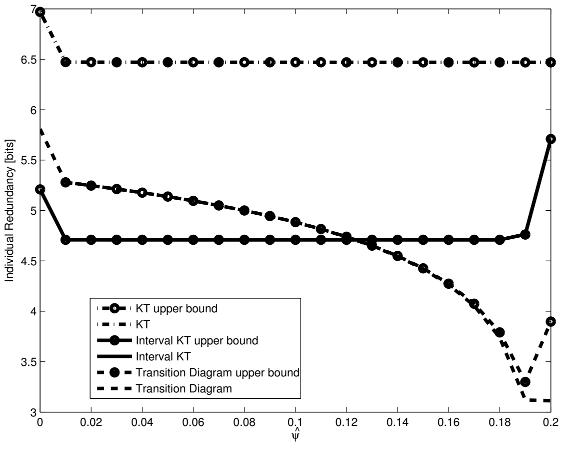

Fig. 3: Individual sequence redundancy for the KT estimate and the two sequential estimators for bounded intervals for $n = 5000$, $\psi \in [0, 0.2]$ and the same range of $\hat{\psi}$.

Theorem 4 shows that the redundancy of this scheme depends on the value of $\hat{\psi}$. The redundancy in the first region can be uniformly bounded by

$$R_n\left(Q, y^n\right) < \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + \log\frac{\beta-\alpha}{\sqrt{\beta(1-\alpha)}} + o(1). \tag{30}$$

Unlike the method in Theorem 1, the method here gains in first order in the region boundaries, reducing the first order redundancy term by a factor of 2. The proof of Theorem 4 appears in [8], and applies similar techniques of the proof of Theorem 2, although somewhat differently.

V. Numerical Results

Figures 2 and 3 show redundancy obtained for the KT estimator and the two bounded probability interval estimators proposed. Each figure shows 5000 bits coded with parameter within a different interval. The gains of the new methods over the KT estimator are clear and are significant even for 5000 bits. The performance of the estimators in the simulations matches the bounds in Theorems 2 and 4. The performance of the first estimator of Theorem 1 is shown to be better and almost uniform in the inner part of the interval, while the second estimator is better around non-extreme boundaries.

VI. Summary and Conclusions

Two low-complexity sequential estimators were proposed for probability assignment to binary sequences whose empirical parameter is known to be confined within an interval $[\alpha, \beta]$ with $\alpha \geq 0$, and $\beta \leq 1$. The redundancy performances of universal compression codes that use the estimators were bounded. Due to the use of the confined interval, the estimators were shown to gain on standard methods as the KT estimator. One estimator, based on cutting off and scaling the standard Dirichlet-$1/2$ for the interval $[\alpha, \beta]$, was shown to perform rather uniformly in the inner part of the interval. The other was stronger in non-extreme boundaries. The methods can be used for many applications, including applications in which losses are linearly proportional to the ratio between assigned probability and the maximizing probability, such as financial applications. The gains over standard methods then become even more significant. Finally, the methods proposed in this work lay the foundation to the more general non-binary case, in which the parameters governing a sequence are possibly confined to only a small subspace of the parameter space.

Acknowledgments

We thank W. Szpankowski for information about [2].

References

[1] T. M. Cover, "Universal portfolios," *Math. Finance*, vol. 1, no. 1, pp. 1-29, Jan. 1991.

[2] M. Drmota, and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2686-2707, Nov. 2004.

[3] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, pp. 199-207, Mar. 1981.

[4] G. I. Shamir and N. Merhav, "Low complexity sequential lossless coding for piecewise stationary memoryless sources," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1498-1519, Jul. 1999.

[5] G. I. Shamir and D. J. Costello, Jr., "Asymptotically optimal low complexity sequential lossless coding for piecewise stationary memoryless sources - Part I: The regular case," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2444-2467, Nov. 2000.

[6] G. I. Shamir, T. J. Tjalkens, and F. M. J. Willems, "Universal noiseless compression for noisy data", *ITA*, San Diego, Cal. 2007.

[7] G. I. Shamir, T. J. Tjalkens, and F. M. J. Willems, "Universal noiseless compression for discrete noisy sequences," in preparations.

[8] G. I. Shamir, T. J. Tjalkens, and F. M. J. Willems, "Low-complexity sequential probability estimation and universal compression for binary sequences with constrained distributions," in preparations.

[9] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The Context-Tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, pp. 653-664, May 1995.

[10] F. M. J. Willems, "Coding for a binary Independent Piecewise-Identically-Distributed source," *IEEE Trans. Inf. Theory*, vol. 42, pp. 2210-2217, Nov. 1996.

[11] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, pp. 431-445, Mar. 2000.