Piecewise Constant Prediction

Erik Ordentlich
Information Theory Research
Hewlett-Packard Laboratories
Palo Alto, CA 94304
Email: erik.ordentlich@hp.com

Marcelo J. Weinberger
Information Theory Research
Hewlett-Packard Laboratories
Palo Alto, CA 94304
Email: marcelo.weinberger@hp.com

Yihong Wu*
Statistics Department, the Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA
Email: yihongwu@wharton.upenn.edu

***Abstract*—Minimax prediction of binary sequences is investigated for cases in which the predictor is forced to issue a piecewise constant prediction. The minimax strategy is characterized for Hamming loss whereas, for logarithmic loss, an asymptotically minimax strategy, in the sense that the main asymptotic redundancy term equals the main asymptotic term of the minimax redundancy, is proposed. The average redundancy case is also analyzed for i.i.d. distributions.**

I. Introduction

Consider a game in which, as a predictor observes a binary sequence $x^n = x_1 x_2 \cdots x_n$, it makes causal predictions on each bit x_{t+1}, $t = 0, 1, \cdots, n-1$, based on the observed prefix x^t. These predictions take the form of probability assignments $p_{t+1}(a|x^t)$, $a \in \{0, 1\}$. Once x_{t+1} is revealed, the predictor incurs a loss, e.g., $-\log p_{t+1}(x_{t+1}|x^t)$ in the data compression problem, or $1 - p_{t+1}(x_{t+1}|x^t)$ for (expected) Hamming loss, which accumulates over time. The goal of the predictor is to approach the cumulative loss of the best *constant* predictor for x^n (determined in hindsight, with full knowledge of x^n, termed *Bayes envelope*), which is the empirical entropy of x^n in the data compression problem, or $\min(n_0(x^n), n_1(x^n))$ for Hamming loss, where $n_a(x^n)$ denotes the number of occurrences of $a \in \{0, 1\}$ in x^n. In one version of the game, the goodness of the predictor is assessed by its excess loss over the Bayes envelope (termed *regret*, or redundancy in the data compression case) for the *worst case* sequence (maximum regret), the best strategy is thus termed *minimax*, and the corresponding maximum regret is the *minimax regret*. Notice that the minimax strategy (and the minimax regret) may depend on the horizon n of the game.

Now, imagine a situation in which the predictor is forced to "freeze" its prediction for a number of prediction rounds. In the simplest such scenario, for a given block length T, the probability assignments $p_{iT+1}, p_{iT+2}, \cdots, p_{(i+1)T}$, $i = 0, 1, \cdots, m-1$, must all coincide, where we assume that $n = mT$ for some positive integer m. Thus, p_{iT+j}, $j = 1, 2, \cdots, T$, can only depend on x^{iT} and, in particular, must be independent of $x_{iT+1}, x_{iT+2}, \cdots, x_{iT+j-1}$. The question arises: How badly can the minimax regret be affected by the constraint of *piecewise constant* prediction? This question is of practical importance in scenarios in which the benefits of a small minimax regret are offset by the cost of computing the minimax strategy for each round of the game. For example, as argued in [1], in an energy-constrained environment in which the role of data compression is to save storage or transmission power, the assessment of the benefit of data compression should take into account the implementation cost of the data compression algorithm. Savings in this cost obtained by "freezing" the adaptation of the algorithm within a block may thus be beneficial despite the corresponding compression loss due to piecewise constant restriction.

The binary prediction problem was first studied in the framework of the *sequential decision problem* [3]. The minimax strategy for Hamming loss was devised by Cover [4], whereas for data compression it is given by the Normalized Maximum-Likelihood (NML) code, due to Shtarkov [5]. Cover's minimax scheme yields the same regret over all sequences, its main asymptotic term being $\sqrt{n/(2\pi)}$. For data compression, the main asymptotic term of the redundancy of the NML code is $(1/2)\log n$ in the binary case. A horizon-independent, simpler to implement approximation of this minimax strategy, which achieves the leading term of the asymptotic minimax redundancy, is given by the Krichevskii-Trofimov (KT) probability assignment [6].

A variant of this problem which, as we shall see, is quite related to the piecewise constant setting, was proposed in [2]. In this variant, the predictor has access to a *delayed* version of the sequence, or is forced to make inferences on the observations a number of instants in advance. Such situations may arise when the application of the prediction is delayed relative to the observed sequence due to, e.g., computational constraints. The delay d, which is assumed known, affects the prediction strategy in that p_{t+1} is now based on $x_1 x_2 \cdots x_{t-d}$ only. It is shown in [2] that, in the delayed prediction setting, the minimax strategy consists of sub-sampling the data at a $1/(d+1)$ rate, and applying the (non-delayed) minimax strategy to each of the $d+1$ sub-sampled sub-sequences. Since the sum of the Bayes envelopes corresponding to each subsequence is not larger than the Bayes envelope of the entire sequence, it is easy to see that the regret of this strategy is upper-bounded by $d+1$ times the minimax regret for sequences of length $n/(d+1)$. The proof that this regret is indeed the (delayed) minimax regret uses an auxiliary piecewise constant strategy (constructed from any given delayed strategy) and, for the data compression case, a convexity argument.

One can consider more general prediction games, in which

* The work of Yihong Wu was essentially done while visiting Hewlett-Packard Laboratories, Palo Alto, CA.

the sequence of observations belongs to some finite alphabet $\mathcal{A}$, and the player causally assigns probability distributions to a corresponding sequence of *actions* $b_1 b_2 \cdots b_n$, taken from an action space $\mathcal{B}$. The observation and action incur an expected loss. Notice that the piecewise constant binary prediction game with block length T and Hamming loss can be cast as an unconstrained game over a sequence of length n/T, in which the observation alphabet is $\mathcal{A} = \{0,1\}^T$, the action space is $\mathcal{B} = \{0^T, 1^T\}$, and the loss function is Hamming. Thus, a general minimax solution for the unconstrained game would include a solution to our problem. Unfortunately, minimax strategies for the general game cannot be characterized as easily as Cover's scheme for the binary case with Hamming loss [7].

In this paper, we start by characterizing, in Section II, the piecewise constant minimax strategy for Hamming loss, under the assumption that $n = mT$ for a given block length T. The (piecewise constant) minimax regret turns out to be T times the minimax regret for horizon m. Since the minimax regret grows as the square root of the sequence length, the asymptotic penalization due to the piecewise constant constraint is a multiplicative factor $\sqrt{T}$. In contrast to the delayed prediction setting, the fact that the above quantity is a lower bound on the piecewise constant minimax regret is easy to see, whereas the characterization of the minimax strategy is far more complex. Indeed, since there is an obvious one-to-one correspondence between piecewise constant prediction strategies on sequences of length n, and unconstrained prediction strategies on sequences of length m, the lower bound follows easily from considering piecewise constant sequences x^n. The upper bound, on the other hand, requires to recursively build up the minimax strategy, whereas in the delayed prediction setting it suffices to simply sub-sample the unconstrained minimax strategy.

In Section III we study the piecewise constant binary data compression problem. While the argument leading to a lower bound extends to this case, yielding a multiplicative factor T as asymptotic penalization, we are unable to give a complete characterization of the minimax strategy. However, we prove that a simple variant of the KT probability assignment [6], in which the estimate is obtained by adding $T/2$ (instead of $1/2$ as in the usual KT estimate) to the counts of 0's and 1's, is asymptotically piecewise constant minimax. Thus, the main asymptotic term of the piecewise constant minimax redundancy takes the form $(T/2)\log n$. Again, the analysis of the upper bound is significantly more involved than in the delayed data compression setting.

In Section IV we study the piecewise constant data compression problem in a probabilistic setting, in which the observations are assumed to be drawn from an (unknown) i.i.d. distribution, where again, for simplicity, we assume binary sequences. The goal here is to minimize the average (rather than the maximum) redundancy, for which lower and upper bounds with main term $(1/2)\log n$ are well known for the unconstrained case (i.e., respectively, Rissanen's lower bound [8], holding for all distributions except for a set of measure zero, and the KT probability assignment, which is asymptotically optimal for every distribution). The question then is: Does the piecewise constant constraint affect the achievable average redundancy? It is interesting to notice that the answer to the counterpart question in the delayed data compression setting is straightforward. Indeed, for i.i.d. distributions, the expected loss incurred at time t for a delayed strategy (with delay d) is the same as the expected loss that the predictor would incur, without delay, at time $t-d$. Therefore, ignoring the delay and assigning at time t the same probability that a non-delayed compression strategy would have assigned at time $t-d$, we incur for the sequence x_{d+1}^n the same loss as, without delay, for the sequence x^{n-d}. Hence, in contrast to the minimax scenario, the delay is asymptotically inconsequential.

The conclusion in the piecewise constant scenario turns out to be similar, but the analysis is far less straightforward. By a nonstandard application of Rissanen's lower bound, we show that applying *any* asymptotically optimal (unconstrained) strategy so that the probability assigned at time $iT+1$ is "frozen" for the entire i-th block, the average redundancy achieves Rissanen's lower bound for all distributions, except possibly for a set of vanishing volume. In fact, for such a general result, an exception set is unavoidable: indeed, we also demonstrate the existence of asymptotically optimal (unconstrained) strategies for which this approach cannot yield optimal average redundancy for *every* distribution. However, we further show that if we specialize this approach to the (asymptotically optimal) KT probability assignment, then no exception set is needed, and the frozen scheme is asymptotically optimal for all distributions.

Throughout the sequel, for fixed positive integers T (block size) and m (number of blocks), we shall formally define a piecewise constant probability assignment $\hat{p}$ with block size T and horizon $n = mT$ as a probability distribution on binary sequences x^n whose conditional probabilities satisfy $\hat{p}_{t+1}(\cdot|x^t) = \hat{p}_{T\lfloor t/T \rfloor + 1}(\cdot|x^{T\lfloor t/T \rfloor})$ for all t. Thus, the conditional probabilities are constant over blocks of size T, and, therefore, such an assignment is completely determined by the conditional probabilities $\hat{p}_{iT+1}(\cdot|x^{iT})$, for $i = 0, \ldots, m-1$. The set of all piecewise constant probabilities with block size T and horizon $n(= mT)$ shall be denoted by $\mathcal{P}_{T,n}$.

II. Piecewise constant minimax prediction

In this section, we study piecewise constant minimax prediction of binary sequences under Hamming loss. The (expected) Hamming loss of a piecewise constant predictor corresponding to a $\hat{p} \in \mathcal{P}_{T,m'T}$ on a sequence $x^{m'T} \in \{0,1\}^{m'T}$ is given by $\mathcal{L}(\hat{p}, x) = \sum_{i=1}^{m'T}(1 - \hat{p}_t(x_t|x^{t-1}))$. Fixing throughout the horizon $n = mT$, let $B(\ell) \triangleq \min(\ell, n-\ell)$ denote the Bayes envelope for a sequence of length n containing ℓ ones. For a *given* sequence x^{kT}, $0 \le k \le m$ and subset $\mathcal{Y} \subseteq \{0,1\}^{(m-k)T}$, define

$$R_{k,T}(x^{kT}, \mathcal{Y}) = \min_{\hat{p} \in \mathcal{P}_{T,(m-k)T}} \max_{\mathbf{y} \in \mathcal{Y}} \left\{ \mathcal{L}(\hat{p}, \mathbf{y}) - B(n_1(x^{kT}\mathbf{y})) \right\}$$

where $x^{kT}\mathbf{y}$ denotes the concatenation of the sequences x^{kT} and $\mathbf{y}$. We are interested in the cases $R_{k,T}(x^{kT}) \triangleq R_{k,T}(x^{kT}, \{0,1\}^{(m-k)T})$ and $R'_{k,T}(x^{kT}) \triangleq R_{k,T}(x^{kT}, \{0^T, 1^T\}^{m-k})$. Notice that the difference between $R'_{k,T}$ and $R_{k,T}$ is in the inner maximization: in the former, $\mathbf{y}$ is constrained to be piecewise constant, while in the latter it is unconstrained. The domains of $R_{k,T}$ and $R'_{k,T}$ are both *unconstrained* binary sequences of length kT. Notice also that these quantities depend on x^{kT} only through the Bayes envelope of the concatenation of x^{kT} and $\mathbf{y}$ (which in turn affects the maximizing sequence $\mathbf{y}$ and the minimizing strategy $\hat{p}^{(m-k)}$). By definition, $R_{0,T}$ is the minimax regret for piecewise constant strategies (the object of the study in this section), whereas $R'_{0,T}$ is T times the minimax regret for unconstrained strategies with horizon m.

Theorem 1. *For all k, $0 \le k \le m$, and $x^{kT} \in \{0,1\}^{kT}$, $R_{k,T}(x^{kT}) = R'_{k,T}(x^{kT})$.*

Corollary 1. *The minimax regret $R_T(mT)$ for the piecewise constant prediction problem with Hamming loss, block length T, and horizon mT, satisfies $R_T(mT) = TR_1(m)$, where $R_1(m)$ denotes Cover's (unconstrained) minimax prediction regret with horizon m.*

The fact that $R_T(mT) \ge TR_1(m)$ is straightforward for more general prediction settings, including data compression. Indeed, the loss of a piecewise constant strategy over a piecewise constant sequence equals T times the loss of a corresponding (unconstrained) strategy on sequences of length m, obtained by sub-sampling the original sequences. By definition of minimax regret, there exists a sequence x^m for which the latter loss is at least $R_1(m)$ plus the Bayes envelope of x^m. The desired inequality then follows by noticing that a T-fold replication of x^m increases its Bayes envelope by a factor of T. As claimed in Corollary 1, the case of binary prediction with Hamming loss is special in that the lower bound holds with equality.

Proof of Theorem 1: The proof will proceed by backward induction, with base case $k = m$. Each induction step will also include a proof of the following auxiliary claims: $R'_{k,T}(x^{kT})$ depends on x^{kT} only through $N = n_1(x^{kT})$, is a convex function of (integer) N, and for all nonnegative integers N, with a slight abuse of notation, $|R'_{k,T}(N+1) - R'_{k,T}(N)| \le 1$.

For the base case, from the above definitions, it is immediate that $R'_{m,T}(x^{mT}) = R_{m,T}(x^{mT}) = -B(n_1(x^{mT}))$, and from the definition of B, it is immediate that the claimed (auxiliary) properties of $R'_{m,T}$ hold. For the induction step, we have

$$R_{k-1,T}(x^{(k-1)T}) = \min_{\hat{p}\in\mathcal{P}_{T,(m-k+1)T}} \max_{\substack{\mathbf{z}\in\{0,1\}^T \\ \mathbf{y}\in\{0,1\}^{(m-k)T}}} \left\{ \mathcal{L}(\hat{p}, \mathbf{zy}) - B(n_1(x^{(k-1)T}\mathbf{zy})) \right\}.$$

Notice that a causal strategy $\hat{p}\in\mathcal{P}_{T,(m-k+1)T}$ can be viewed as a fixed predictor $\hat{p}_1 = p$ for the first block $\mathbf{z}$, which incurs a loss $p(T-n_1(\mathbf{z}))+\overline{p}n_1(\mathbf{z})$ (where $\overline{p}=1-p$), followed by a causal strategy $\hat{p}\in\mathcal{P}_{T,(m-k)T}$, the choice of which *can depend* on $\mathbf{z}$, for the remaining blocks $\mathbf{y}$, incurring a loss $\mathcal{L}(\hat{p}, \mathbf{y})$. Clearly, this dependency can be expressed by switching the maximum on $\mathbf{z}$ with the minimum on $\hat{p}\in\mathcal{P}_{T,(m-k)T}$, yielding

$$\begin{aligned} R_{k-1,T}(x^{(k-1)T}) &= \min_{p\in[0,1]} \max_{\mathbf{z}\in\{0,1\}^T} \left\{ p(T-n_1(\mathbf{z}))+\overline{p}n_1(\mathbf{z})+ \right. \\ &\left. \min_{\hat{p}\in\mathcal{P}_{T,(m-k)T}} \max_{\mathbf{y}\in\{0,1\}^{(m-k)T}} \left\{ \mathcal{L}(\hat{p}, \mathbf{y}) - B(n_1(x^{(k-1)T}\mathbf{zy})) \right\} \right\} \\ &= \min_{p\in[0,1]} \max_{\mathbf{z}\in\{0,1\}^T} \left\{ p(T-n_1(\mathbf{z}))+\overline{p}n_1(\mathbf{z}) + R_{k,T}(x^{(k-1)T}\mathbf{z}) \right\} \qquad (1) \\ &= \min_{p\in[0,1]} \max_{\mathbf{z}\in\{0,1\}^T} \left\{ p(T-n_1(\mathbf{z}))+\overline{p}n_1(\mathbf{z}) + R'_{k,T}(x^{(k-1)T}\mathbf{z}) \right\} \end{aligned}$$

where the last equality follows by the induction hypothesis that $R_{k,T} = R'_{k,T}$. By the induction hypothesis that $R'_{k,T}$ depends on $x^{(k-1)T}\mathbf{z}$ through $n_1(x^{(k-1)T})+n_1(\mathbf{z})$, we further have

$$\begin{aligned} R_{k-1,T}(x^{(k-1)T}) &= \min_{p\in[0,1]} \max_{\mathbf{z}\in\{0,1\}^T} \left\{ p(T-n_1(\mathbf{z}))+\overline{p}n_1(\mathbf{z}) \right. \\ &\qquad \left. + R'_{k,T}(n_1(x^{(k-1)T})+n_1(\mathbf{z})) \right\} \qquad (2) \\ &= \min_{p\in[0,1]} \max \left\{ pT + R'_{k,T}(n_1(x^{(k-1)T})), \right. \\ &\qquad \left. \overline{p}T + R'_{k,T}(n_1(x^{(k-1)T})+T) \right\} \qquad (3) \end{aligned}$$

where (3) follows by the convexity (in $n_1(\mathbf{z})$) of the expression in the maximization in (2) (implying that the maximum occurs at one of the extremes), which, in turn, follows from the auxiliary induction hypothesis that $R'_{k,T}(N)$ is convex in N. Now, starting from $R'_{k-1,T}(x^{(k-1)T})$, the same chain of equalities leading to (1) can be followed, with the maximization constrained to piecewise constant sequences. Clearly, the result would be the expression in the right-hand side of (3). Therefore, by (3), $R'_{k-1,T}(x^{(k-1)T}) = R_{k-1,T}(x^{(k-1)T})$, establishing the induction step for the main claim, as well as the fact that $R'_{k-1,T}$ depends on $x^{(k-1)T}$ through $n_1(x^{(k-1)T})$.

To prove the remaining auxiliary induction steps, we note that (3) can be solved (by equating the two terms in the maximum) to yield that the minimizing p takes the form

$$p^* = \frac{1}{2} + \frac{R'_{k,T}(n_1(x^{(k-1)T})+T) - R'_{k,T}(n_1(x^{(k-1)T}))}{2T} \qquad (4)$$

which satisfies $p^*\in[0,1]$ by the Lipschitz condition induction hypothesis that $|R'_{k,T}(N+1) - R'_{k,T}(N)| \le 1$. Substituting (4) into (3) yields

$$R'_{k-1,T}(n_1(x^{(k-1)T})) = \frac{T}{2} + \frac{R'_{k,T}(n_1(x^{(k-1)T})+T) + R'_{k,T}(n_1(x^{(k-1)T}))}{2} \qquad (5)$$

from which the convexity and Lipschitz condition induction steps both readily follow from the respective induction hypotheses. □

We notice that (4) and (5) explicitly characterize the piecewise constant minimax strategy for Hamming loss.

III. Piecewise constant minimax data compression

In this section, we consider the piecewise constant binary data compression problem in a minimax setting. For simplicity, we maintain the assumption that $n{=}mT$, although we notice that the results hold in more generality. As observed in Section II, the minimax piecewise constant redundancy, defined as

$$\tilde{R}_T(n) = \min_{\hat{p}\in\mathcal{P}_{T,n}} \max_{x^n\in\{0,1\}^n} \left[\log \frac{1}{\hat{p}(x^n)} - n\hat{H}(x^n)\right] \tag{6}$$

where $\hat{H}(x^n)$ denotes the empirical entropy of x^n, satisfies

$$\tilde{R}_T(n) \geq T\tilde{R}_1(m)\,. \tag{7}$$

As discussed in Section I, $\tilde{R}_1(n)$, attained by the NML code [5] for each n, satisfies the following asymptotics

$$\tilde{R}_1(n) = \frac{1}{2}\log n + O(1), \quad n\to\infty, \tag{8}$$

where the leading term in (8) is achieved by the KT probability assignment [6]. The main result of this section is given by the following theorem.

Theorem 2. *The minimax piecewise constant redundancy satisfies* $\tilde{R}_T(n) = \frac{T}{2}\log n + O(1)$,[1] *the right-hand side of which is attained by the piecewise constant probability assignment specified by*

$$\hat{p}_{\mathrm{U},iT+1}(1|x^{iT}) \triangleq \hat{p}_{iT+1}(1|x^{iT}) = \frac{n_1(x^{iT})+T/2}{iT+T} \tag{9}$$

$0\leq i<m$ *(add-*$\frac{T}{2}$ *estimator).*

Proof: In view of the lower bound (7) and the definition of minimax redundancy (6), it suffices to show that, for every sequence x^{mT}, we have

$$\hat{p}_{\mathrm{U}}(x^{mT}) \geq \frac{1}{2}(mT)^{-\frac{T}{2}} 2^{-mT\hat{H}(x^{mT})}\,. \tag{10}$$

To this end, we show that, for any x^{mT}, with $n_1(x^{mT}) = a$ and $n_0(x^{mT}) = b = mT - a$,

$$\hat{p}_{\mathrm{U}}(x^{mT}) \geq \frac{1}{2}(mT)^{-\frac{T}{2}} \left(\frac{a}{mT}\right)^a \left(\frac{b}{mT}\right)^b. \tag{11}$$

We proceed by induction on m. For $m=1$, $\hat{p}_{\mathrm{U}}(x^T) = 2^{-T} \geq \frac{1}{2}T^{-\frac{T}{2}}$ for all $T\geq 1$. Next, assume that (11) holds for m. Let $n_1(x^{(m+1)T}) = a+c$ and $n_0(x^{(m+1)T}) = b+d$, with $c+d=T, 0\leq c,d\leq T$. By the probability assignment in (9),

$$\hat{p}_{\mathrm{U}}(x^{(m+1)T}) = \hat{p}_{\mathrm{U}}(x^{mT}) \left(\frac{a+\frac{T}{2}}{(m+1)T}\right)^c \left(\frac{b+\frac{T}{2}}{(m+1)T}\right)^d.$$

[1] We suppress a possible dependence on T in the $O(1)$ term.

Thus, by the induction hypothesis,

$$\begin{aligned}\hat{p}_{\mathrm{U}}(x^{(m+1)T}) &\geq \frac{1}{2}(mT)^{-\frac{T}{2}} \left(\frac{a}{mT}\right)^a \left(\frac{b}{mT}\right)^b \\ &\quad\times \left(\frac{a+\frac{T}{2}}{(m+1)T}\right)^c \left(\frac{b+\frac{T}{2}}{(m+1)T}\right)^d \\ &= \frac{1}{2}((m+1)T)^{-\frac{T}{2}} \left(\frac{a+c}{(m+1)T}\right)^{a+c} \left(\frac{b+d}{(m+1)T}\right)^{b+d} \\ &\quad\times \left(1+\frac{1}{m}\right)^{(m+\frac{1}{2})T} \left(\frac{a+\frac{T}{2}}{a+c}\right)^c \left(\frac{a}{a+c}\right)^a \left(\frac{b+\frac{T}{2}}{b+d}\right)^d \left(\frac{b}{b+d}\right)^b \end{aligned} \tag{12}$$

where the equality follows by a simple reordering of the terms. Now, let $G(m) \triangleq \left(1+\frac{1}{m}\right)^{m+\frac{1}{2}}$. It can be shown that G is strictly decreasing on $(0,\infty)$, so that $G(m)\geq G(\infty) = \mathrm{e}$ (we omit the proof due to space limitations). Also, let $F(a,c) \triangleq \left(\frac{a+\frac{T}{2}}{a+c}\right)^c \left(\frac{a}{a+c}\right)^a$, which can be shown to be a decreasing function of a on $(0,\infty)$ (again, we omit the proof). Therefore, $F(a,c)\geq F(\infty,c) = \mathrm{e}^{-c}$ and, similarly, $F(b,d)\geq \mathrm{e}^{-d}$. It then follows from (12) that

$$\begin{aligned}\hat{p}_{\mathrm{U}}(x^{(m+1)T}) &\geq \frac{1}{2}((m+1)T)^{-\frac{T}{2}} \\ &\quad\times \left(\frac{a+c}{(m+1)T}\right)^{a+c} \left(\frac{b+d}{(m+1)T}\right)^{b+d} \mathrm{e}^T\mathrm{e}^{-c}\mathrm{e}^{-d} \\ &= \frac{1}{2}((m+1)T)^{-\frac{T}{2}} \left(\frac{a+c}{(m+1)T}\right)^{a+c} \left(\frac{b+d}{(m+1)T}\right)^{b+d}\end{aligned}$$

which completes the induction step. □

IV. Piecewise constant universal compression: Average case redundancy

In this section, we consider the piecewise constant universal compression problem from an average redundancy point of view. Recall that, given a compressor expressed as a probability assignment $\hat{p}(x^n)$ on sequences of length n, the average case redundancy with respect to an i.i.d. source with marginal p is defined as $\tilde{R}_{avg}(\hat{p},p) = D(p^{\otimes n}||\hat{p})$, where $p^{\otimes n}$ denotes the i.i.d. distribution on length n sequences with marginal p. As discussed in Section I, it was observed in [2] that the difference between the average code length of any compression scheme and its delayed counterpart is $O(1)$. Thus, the average redundancy of the delayed version of any compression scheme with average redundancy meeting Rissanen's lower bound of $(1/2)\log n + o(\log n)$ (for binary sources), also meets Rissanen's lower bound. Theorem 3 below is a counterpart of this result for the piecewise constant setting.

Given any probability assignment $\hat{p}$ with $\tilde{R}_{avg}(\hat{p},p) = (1/2)\log n + o(\log n)$ for *all* p, and block length T, let $\hat{p}^{(T)}$ denote the T-piecewise constant version of $\hat{p}$, defined via the conditional probability assignments

$$\hat{p}^{(T)}_{t+1}(\cdot|x^t) = \hat{p}_{T\lfloor t/T\rfloor+1}(\cdot|x^{T\lfloor t/T\rfloor}).$$

Thus, $\hat{p}^{(T)}_{t+1}(\cdot|x^t)$ is fixed throughout each block of indices $iT, iT{+}1, \ldots, iT{+}T{-}1$, with i being a nonnegative integer,

to be the conditional probability that $\hat{p}$ would have induced at the start of the block.

Theorem 3. *Given any probability assignment $\hat{p}$ with $\max_p \tilde{R}_{avg}(\hat{p}, p) = (1/2)\log n + o(\log n)$, the T-piecewise constant version $\hat{p}^{(T)}$ satisfies $\tilde{R}_{avg}(\hat{p}^{(T)}, p) = (1/2)\log n + o(\log n)$ for Lebesgue almost all p.*

Proof: In analogy to $\hat{p}^{(T)}$, for $\delta = 0, 1, \ldots, T-1$, we can define the (T, δ)-piecewise constant versions of $\hat{p}$ as

$$\hat{p}_{t+1}^{(T,\delta)}(\cdot|x^t) = \hat{p}_{T\lfloor(t-\delta)/T\rfloor+\delta+1}(\cdot|x^{T\lfloor(t-\delta)/T\rfloor+\delta})$$

with $\hat{p}_t(\cdot|\cdot) = 1/2$ for $t \leq \delta$. Thus, $\hat{p}_{t+1}^{(T,\delta)}(\cdot|x^t)$ is fixed over blocks as above, but the blocks are offset by δ with respect to those defining $\hat{p}^{(T)}$. Note that $\hat{p}^{(T,0)}$ coincides with $\hat{p}^{(T)}$.

We claim that

$$\frac{1}{T}\sum_{\delta=0}^{T-1} \tilde{R}_{avg}(\hat{p}^{(T,\delta)}, p) = \tilde{R}_{avg}(\hat{p}, p) + O(1). \qquad (13)$$

Indeed, by the i.i.d. nature of the source and the definition of $\hat{p}^{(T,\delta)}$, we can rewrite the left-hand side of (13) in terms of conditional probabilities and regroup terms, with the $O(1)$ error term arising from a few terms at the boundaries of the sequence. It follows from (13) that, if $\tilde{R}_{avg}(\hat{p}, p) = (1/2)\log n + o(\log n)$ and, for some $\epsilon > 0$ and infinitely many values of n, $\tilde{R}_{avg}(\hat{p}^{(T)}, p) > (1/2)\log n + \epsilon \log n$, then $\tilde{R}_{avg}(\hat{p}^{(T,\delta)}, p) < (1/2)\log n - \epsilon \log n + o(\log n)$, for some δ and infinitely many values of n. However, by Rissanen's lower bound [8], this can only happen for p in a set of Lebesgue measure zero. Since δ is finite-valued (countable would suffice), it follows, therefore, that for $\hat{p}$ with the properties assumed in the theorem, $\tilde{R}_{avg}(\hat{p}^{(T)}, p) = (1/2)\log n + o(\log n)$, except possibly for p in a set of Lebesgue measure zero. □

The following example (for $T = 2$) shows that the measure zero exception set in Theorem 3 can arise. Consider $\hat{p}$ defined as follows. For odd t, let

$$\hat{p}_t(1|x^{t-1}) = \frac{n_1(x^{t-1})+1}{t+1} \qquad (14)$$

whereas, for even t, let

$$\hat{p}_t(1|x^{t-1}) = \frac{\hat{p}(x^{t-1}1)}{\hat{p}(x^{t-1})} = \frac{\left[(t+1)\binom{t}{n_1(x^{t-1})+1}\right]^{-1} + 1(x^{t-1}=\mathbf{1})}{\left[t\binom{t-1}{n_1(x^{t-1})}\right]^{-1} + 1(x^{t-1}=\mathbf{1}) + 1(x^{t-1}=\mathbf{0})} \qquad (15)$$

where $1(x^{t-1} = \mathbf{1})$ (resp. $1(x^{t-1} = \mathbf{0})$) is 1 (resp. 0) if x^{t-1} is all 1's (resp. 0's) and 0 otherwise. Notice that (14) corresponds to Laplace's estimator while (15) corresponds to the conditional probability distribution induced by an equal mixture of the Laplace probability assignment (corresponding to the Dirichlet-1 mixture of product distributions) and the respective deterministic distributions of all 1's and all 0's.

For $p \neq 0, 1$, the above $\hat{p}$ coincides with the Laplace estimator after the first occurrence of a bit differing from x_1. The probability of this event taking i symbols to occur falls off exponentially in i, and the code length difference between both estimators is $O(\log i)$ in the worst case. Therefore, the excess redundancy over the Laplace estimator is $O(1)$. Since the latter is well known to attain an average redundancy of $(1/2)\log n + o(\log n)$, it follows that $\hat{p}$ does so as well (for $p \neq 0, 1$). For $p = 0, 1$, on the other hand, a simple calculation shows that the redundancy arising from odd samples is $(1/2)\log n + o(\log n)$, while the redundancy arising from even samples is $o(\log n)$. Thus, $\hat{p}$ meets the conditions of Theorem 3. As for the corresponding $\hat{p}^{(2)}$, it corresponds to using the Laplace estimator conditional probabilities (14) over each block of size 2 and again a simple calculation shows that the average redundancy is $\log n + o(\log n)$, for $p = 0, 1$. This fact is inherently due to the known analogous redundancy for the underlying Laplace estimator for these deterministic cases.

Our final result for piecewise universal compression under average redundancy concerns the T-piecewise constant version of the KT probability assignment, denoted by $\hat{p}_{\mathsf{KT}}$. In this case, we can prove the following result.

Proposition 1. *For all $p \in [0, 1]$ and all T, $\tilde{R}_{avg}(\hat{p}_{\mathsf{KT}}^{(T)}, p) = \frac{1}{2}\log n + O(1)$.*

Thus, in the context of the previous theorem, the T-piecewise constant version of the KT probability assignment turns out to attain Rissanen's lower bound for *all* p. We omit a complete proof of Proposition 1. The key idea is to show that for any $r > 0$, with overwhelming probability,

$$E(\log \hat{p}_{\mathsf{KT},t}(X_t|X^{t-1})|X^{T\lfloor t/T\rfloor}) - E(\log \hat{p}_{\mathsf{KT},T\lfloor t/T\rfloor+1}(X_t|X^{T\lfloor t/T\rfloor})|X^{T\lfloor t/T\rfloor}) < 1/t^{2-r}. \qquad (16)$$

The proof is completed by accounting also for the contribution to the average redundancy in the low-probability event that (16) does not hold and then summing the result over t.

References

[1] Y. Wu, E. Ordentlich, and M. J. Weinberger, "Energy-Optimized Lossless Compression: Rate-Variability Tradeoff," Proceedings of the 2011 IEEE International Symposium on Information Theory (ISIT'11), St. Petersburg, Russia, Aug. 2011.

[2] M.J. Weinberger and E. Ordentlich, "On Delayed Prediction of Individual Sequences," *IEEE Trans. Inform. Theory,* Vol. IT-48, No. 7, pp. 1959–1976, July 2002.

[3] J. F. Hannan, "Approximation to Bayes risk in repeated plays," in *Contributions to the Theory of Games, Volume III, Ann. Math. Studies*, vol. 3, pp. 97–139, Princeton, NJ, 1957.

[4] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes*, (Prague), pp. 263–272, Publishing House of the Czechoslovak Academy of Sciences, 1967.

[5] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Inform. Trans.*, vol. 23, pp. 175–186, July 1987.

[6] R. E. Krichevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[7] T. H. Chung, *Minimax Learning in Iterated Games via Distributional Majorization*. PhD thesis, Department of Electrical Engineering, Stanford University, 1994.

[8] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.