Coding of Sources with Two-Sided Geometric Distributions and Unknown Parameters*

Neri Merhav†
Electrical Engineering Department
Technion
Haifa 32000, Israel

Gadiel Seroussi and Marcelo J. Weinberger
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304, USA.

Abstract

Lossless compression is studied for a countably infinite alphabet source with an unknown, off-centered, two-sided geometric (TSG) distribution, which is a commonly used statistical model for image prediction residuals. In this paper, we demonstrate that arithmetic coding based on a simple strategy of model adaptation, essentially attains the theoretical lower bound to the universal coding redundancy associated with this model. We then focus on more practical codes for the TSG model, that operate on a symbol-by-symbol basis, and study the problem of adaptively selecting a code from a given discrete family. By taking advantage of the structure of the optimum Huffman tree for a known TSG distribution, which enables simple calculation of the codeword of every given source symbol, an efficient adaptive strategy is derived.

Index Terms: Lossless image compression, infinite alphabet, geometric distribution, exponential distribution, Golomb codes, prediction residual, universal coding, sequential coding, universal modeling.

---

*Parts of this paper were presented in the 1996 International Conference on Image Processing, Lausanne, Switzerland, and in the 1997 International Symposium on Information Theory, Ulm, Germany.

†This work was done while the author was on sabbatical leave at Hewlett-Packard Laboratories, Palo Alto, California. The author is also with Hewlett-Packard Laboratories–Israel in Haifa, Israel.

1 Introduction

A traditional paradigm in data compression is that sequential lossless coding can be viewed as the following inductive statistical inference problem. At each time instant t, after having observed past source symbols $x^t = (x_1, x_2, \cdots, x_t)$, but before observing x_{t+1}, one assigns a conditional probability $p(\cdot|x^t)$ to the next symbol x_{t+1}, and accumulates a loss (i.e., code length) $\sum_t -\log p(x_{t+1}|x^t)$, to be minimized in the long run. In contrast to non-sequential (multi-pass) methods, in the sequential setting, the conditional distribution $p(\cdot|x^t)$ is learned solely from the past x^t, and so, the above code length can be implemented sequentially by arithmetic coding. The sequential decoder, which instantaneously has access to the previously decoded data x^t, can determine $p(\cdot|x^t)$ as well, and hence can also decode x_{t+1}.

In universal coding for a parametric class of sources, the above probability assignment is designed to simultaneously best match every possible source within this class. For example, the *context* (or finite-memory) model [1, 2] has been successfully applied to lossless image compression [3, 4, 5, 6], an application which serves as the main motivation for this work. According to this model, the conditional probability of each symbol, given the entire past, depends only on a bounded, but possibly varying number of the most recent past symbols, referred to as "context." In this case, the conditional symbol probabilities given each possible context are natural parameters.

A fundamental limit to the performance of universal coding is given by Rissanen's lower bound [7, Theorem 1] on the universal coding redundancy for a parametric class of sources. This lower bound is described as follows. Let $\{P_\psi,\ \psi \in \Psi\}$ be a parametric class of information sources indexed by a K-dimensional parameter vector ψ, which takes on values in a bounded subset $\Psi \subset \mathbb{R}^K$. Assume that there exists a $\sqrt{n}$-consistent estimator $\hat{\psi}_n = \hat{\psi}_n(x^n)$ for ψ in the sense that $\lim_{n\to\infty} P_\psi\{x^n :\ \sqrt{n}||\hat{\psi}_n - \psi|| > c\}$ exists for fixed c and is upper bounded by a function $\sigma(c)$ that is independent of ψ and tends to zero as $c \to \infty$. Let $Q(\cdot)$ be an arbitrary probability distribution on the space of source n-tuples, which is independent of the unknown value of ψ. Then, for every $\epsilon > 0$ and every ψ, except for a subset of Ψ with vanishing Lebesgue measure as a function of n,

$$D(P_\psi||Q) \triangleq E_\psi \log \frac{P_\psi(X^n)}{Q(X^n)} \geq (1-\epsilon)\frac{K}{2}\log n, \tag{1}$$

where E_ψ denotes expectation w.r.t. P_ψ, $X^n = (X_1, ..., X_n)$ is a random source vector drawn by P_ψ, and logarithms here and throughout the sequel are taken to the base 2. The left-hand side of (1) represents the unnormalized coding redundancy associated with lossless coding according to Q while the underlying source is P_ψ. The right-hand side represents the unavoidable cost of universality when the code is not allowed to depend on ψ. This inequality tells us that if Q is chosen under a pessimistic assumption of an overly large K, then each unnecessary degree of freedom would cost essentially $0.5 \log n$ extra bits beyond the necessary model cost. Thus, the choice of K plays a fundamental role in modeling problems. By (1), it is important to keep it at the minimum necessary level whenever possible, by use of available prior information on the data to be modeled, so as to avoid overfitting. In the above example of the context model, K is given by the product of the number of contexts and the number of parameters per context. Thus, reducing the latter (e.g., by utilizing prior knowledge on the structure of images to be compressed) allows for a larger number of contexts without penalty in overall model cost.

The discussion thus far applies to general parametric classes of information sources. Motivated by the application of lossless image compression, in which prediction [8] is a very useful tool to capture expected relations (e.g., smoothness) between adjacent pixels, our focus henceforth will be confined to the class of integer-valued sources with a distribution given by the *two-sided geometric* (TSG) model. It has been observed [9] that prediction errors are well-modeled by the TSG distribution (TSGD) centered at zero, henceforth referred to as *centered TSGD*. According to this distribution, the probability of an integer value x of the prediction error ($x = 0, \pm1, \pm2, ...$), is proportional to $\theta^{|x|}$, where $\theta \in (0,1)$ controls the two-sided exponential decay rate. When combined with a context model as in [4, 5], the TSG model is attractive also because there is only one parameter (θ) per context, although the alphabet is in principle infinite (and in practice finite but quite large, e.g., 8 bits per pixel). This allows for a modeling strategy based on a fairly large number of contexts at a reasonable model cost.

Motivated by the objective of providing a theoretical framework for recently developed lossless image compression algorithms (e.g., [5], see also [10])[1], we shall study lossless compression for a model that is somewhat more general than the centered TSG in that it includes also a shift parameter d for each context. This parameter reflects a DC offset typically present in the prediction residual signal of context-based schemes, due to integer-value constraints and possible bias in the estimation part. Non-integer values of d are also useful for better capturing the two adjacent modes often observed in empirical context-dependent histograms of prediction errors. The more general model is defined next. First, notice that the outcomes of a source are conditionally independent given their contexts. Therefore, according to the context model, one can view the subsequence of symbols that follow any given fixed context, as if it emerged from a memoryless source, whose TSGD parameters correspond to this context. Thus, although the TSG model in image compression is well-motivated [4, 5] when combined with the context model, for the sake of simplicity, we shall consider the parametric class of memoryless sources $\{P_\psi\}$, $\psi = (\theta, d)$ (hence $K = 2$), given by

$$P_\psi(x) = P_{(\theta,d)}(x) = C(\theta, d)\theta^{|x+d|}, \qquad x = 0, \pm 1, \pm 2, ..., \tag{2}$$

where $0 < \theta < 1$ as above, $0 \leq d < 1$, and

$$C(\theta, d) = (1-\theta)/(\theta^{1-d} + \theta^d) \tag{3}$$

is a normalization factor. This limited range of d, which corresponds to distribution modes at 0 and -1, can be obtained by a suitable adaptive predictor with an error feedback loop [5, 6]. The centered TSGD corresponds to $d = 0$, and, when $d = \frac{1}{2}$, $P_{(\theta,d)}$ is a bi-modal distribution with equal peaks at -1 and 0. (The preference of -1 over $+1$ here is arbitrary).

In general, the TSG model (2) is used without prior knowledge of the parameters (θ, d). Thus, a coding strategy based on arithmetic coding requires a sequential probability assignment scheme. As discussed in Section 2, the bound (1) applies (with $K = 2$), so the goal of a universal probability assignment for the TSG model is to achieve a coding redundancy of $(\log n)/n$ bits per symbol, simultaneously for all models in the class. One such simple strategy of model

[1] The algorithm in [5] has recently been adopted as the baseline for the lossless image compression standard JPEG-LS [11].

adaptation, derived by the method of mixtures, is demonstrated in Section 2. To this end, the parametric family $\{P_{(\theta,d)}\}$ is modified so as to make probability assignments given by mixture integrals have closed form expressions that are implementable in a sequential manner.

In many situations, however, symbol-by-symbol coding is more attractive than arithmetic coding from a practical point of view [5], despite incurring larger redundancy. This approach is especially appealing when the Huffman codes for the targeted class of sources (for *known* parameters) form a structured family, which enables simple calculation of the codeword of every given source symbol. Based on the observed sequence x^t, one can select a code in the family sequentially, and use this code to encode x_{t+1}. Unlike in Section 2, the set of available coding strategies for each symbol is discrete, and the adaptation approach is inherently "plug-in." The performance of this on-line algorithm is measured by its average code length (under the unknown model parameters), and the objective is to perform essentially as well as the best fixed strategy in the family for the unknown parameter values. A structured family of codes relaxes the need of dynamically updating code tables due to possible variations in the estimated parameter ψ (see, e.g., [12]).

The analogy between the TSG distribution, and the *one-sided* geometric (OSG) distribution of nonnegative integers, for which the well-structured *Golomb codes* [13] are optimal [14], suggested *ad hoc* approaches to adaptive symbol-by-symbol coding of centered TSG distributions [15, 16]. The complete characterization of minimum expected-length prefix codes for the TSG sources in (2) for known values of θ and d, presented in the companion paper [17], makes it possible to approach in a more comprehensive way the design of low complexity adaptive strategies for encoding TSG models. In Section 3, we provide optimal adaptation criteria (in a well-defined sense) for a further simplified, sub-optimal family of codes used in practice [15, 5] and analyzed in [17].

2 Universal Probability Assignment for TSG models

Consider the class of sources defined in (2), where $\psi = (\theta, d)$ is unknown a-priori. Since Rissanen's lower bound on the universal coding redundancy (1) applies (as will be shown in the sequel), and since $K = 2$, this redundancy essentially cannot fall below $(\log n)/n$ bits per symbol, simultaneously for most sources in $\Psi = (0,1) \times [0,1)$.

In view of this, our goal is to devise a universal probability assignment strategy $\hat{Q}$ that essentially achieves this lower bound. Moreover, we would like to avoid the dependence of the per-symbol probability assignment at each time instant t on future data as well as on the horizon n of the problem, which may not be specified in advance.

It is well known that for certain parametric classes of sources, e.g., finite-alphabet memoryless sources parametrized by the letter probabilities, these objectives can be achieved by the method of mixtures (see, e.g., [18, 19, 20]). The idea behind this method is to assign a certain prior $w(\psi)$ on the parameter set Ψ, and to define the probability assignment as

$$\hat{Q}(x^n) = \int_{\Psi} dw(\psi) P_{\psi}(x^n) \tag{4}$$

where $\{P_{\psi}\}$ is the targeted parametric class of sources. Since $\hat{Q}(x^t) = \sum_{x_{t+1}} \hat{Q}(x^{t+1})$ and $\hat{Q}(x_{t+1}|x^t) = \hat{Q}(x^{t+1})/\hat{Q}(x^t)$, it is guaranteed that instantaneous probability assignments do not depend on future outcomes. If, in addition, w does not depend on n, then neither do the probability assignments $\hat{Q}(x_{t+1}|x^t)$ for $t < n$. In this respect, the method of mixtures has a clear advantage over two-pass methods that are based on explicit batch estimation of ψ, where these sequentiality properties do not hold in general. The goal of attaining Rissanen's lower bound can be also achieved for certain choices of the prior w. In some cases (see, e.g., [21]), there is a certain choice of w for which the lower bound is essentially attained not only on the average, but moreover, *pointwise* for every x^n. In other words,

$$\log \frac{P_{\psi}(x^n)}{\hat{Q}(x^n)} \leq \frac{K}{2} \log n + O(1)$$

for every x^n and every $\psi \in \Psi$, where $O(1)$ designates a term that is upper bounded by a constant uniformly for every sequence.

Unfortunately, in contrast to the well-studied finite-alphabet case, where there is a closed-form expression for the mixture integral (4) for every x^n, and the instantaneous probability assignments are easy to derive, the TSG model does not directly lend itself to this technique. The simple reason is that there is no apparent closed-form expression for mixtures of the parametric family $\{P_\psi\}$ in (2). Nevertheless, it turns out that after a slight modification of the TSG model, which gives a somewhat larger class of distributions, the method of mixtures becomes easily applicable without essentially affecting the redundancy. Specifically, the idea is the following. Let us re-define the parametric family as $\{Q_\varphi\}$, where now $\varphi = (\theta, \rho)$ and

$$Q_\varphi(x) \triangleq Q_{(\theta,\rho)}(x) = \begin{cases} \rho(1-\theta)\theta^x & x = 0, 1, 2, ... \\ (1-\rho)(1-\theta)\theta^{-x-1} & x = -1, -2, ... \end{cases} \tag{5}$$

with $\theta \in (0,1)$ as above, and $\rho \in [0,1]$. Clearly, the new parameter ρ designates the probability that a random variable drawn according to the distribution (5) be nonnegative. By the relations $Q_\varphi(x+1) = \theta Q_\varphi(x)$, $x \geq 0$, and $Q_\varphi(x-1) = \theta Q_\varphi(x)$, $x < 0$, every source in the original definition of the TSG model (2) corresponds to some source in the modified TSG model (5), with the same value for the parameter θ and with the parameter ρ given by

$$\rho = \frac{\theta^d}{\theta^{1-d} + \theta^d} \quad . \tag{6}$$

However, while the original TSG model allows only for $\rho \in (\theta/(1+\theta), 1/(1+\theta)]$ for a given θ, the model (5) permits any $\rho \in [0,1]$. It follows that the modified TSG model (5) is strictly richer than the original model (2), but without increasing the dimension K of the parameter space, and hence without extra model cost penalty. Therefore, it will be sufficient to devise a universal probability assignment $\hat{Q}$ for the modified TSG model.

We will also use the modified TSG model to prove the existence of a $\sqrt{n}$-consistent estimator and hence the applicability of Rissanen's lower bound. This is valid because of the following consideration: Since the Lebesgue measure occupied by the set of sources that correspond to the original TSG model is a fixed fraction (larger than 25%) of the set of sources in the modified model (5), then a lower bound that holds for "most" sources (Lebesgue) in the modified class, still holds for "most" sources (Lebesgue) in the original class. Thus, it will be sufficient to prove $\sqrt{n}$-consistency of a certain estimator for the modified model.

In order to construct a universal probability assignment for the modified TSG model, we will consider the representation of an arbitrary integer x as a pair (y, z), where

$$y = y(x) \triangleq \begin{cases} 0 & x \geq 0 \\ 1 & x < 0 \end{cases} \tag{7}$$

and

$$z = z(x) \stackrel{\triangle}{=} |x| - y(x). \tag{8}$$

Since the relation between x and (y, z) is one-to-one, no information is lost by this representation. The key observation now is that if X is a random variable drawn under distribution (5), then $Y = y(X)$ and $Z = z(X)$ are independent, where Y is binary $\{0, 1\}$ with parameter $\rho \stackrel{\triangle}{=} Q_\rho^Y(0) \stackrel{\triangle}{=} \Pr\{Y = 0\}$, and Z is OSG with parameter θ, that is,

$$\Pr\{Z = z\} \stackrel{\triangle}{=} Q_\theta^Z(z) \stackrel{\triangle}{=} Q_\varphi(z) + Q_\varphi(-z-1) = (1-\theta)\theta^z, \qquad z = 0, 1, 2, \ldots . \tag{9}$$

Accordingly, given a memoryless source $X_1, X_2, \ldots$ with a distribution given by (5), one creates, using $y(\cdot)$ and $z(\cdot)$, two independent memoryless processes, $Y_1, Y_2, \ldots \sim Q_\rho^Y$ and $Z_1, Z_2, \ldots \sim Q_\theta^Z$, where the former is Bernoulli with parameter ρ, and the latter is OSG with parameter θ.

The independence between $\{Y_t\}$ and $\{Z_t\}$ and the fact that each one of these processes is parametrized by a different component of the parameter vector, significantly facilitate the universal probability assignment (and hence also universal arithmetic coding) for this model class, since these processes can be encoded separately without loss of optimality. To encode $y_{t+1} = y(x_{t+1})$, we use the probability assignment [19]

$$\hat{Q}^Y\{y_{t+1} = 1 | y^t\} = \frac{N_t + 1/2}{t+1} \tag{10}$$

where

$$N_t = \sum_{i=1}^{t} y_i \tag{11}$$

and for $t = 0$, $y^t = y^0$ is interpreted as the null string with $N_0 \stackrel{\triangle}{=} 0$. This probability assignment is induced by a mixture of type (4) using the Dirichlet(1/2) prior on ρ, that is, the prior which is inversely proportional to $\sqrt{\rho(1-\rho)}$. Similarly, the probability assignment for z_{t+1} given z^t is the result of a Dirichlet(1/2) mixture over θ, which gives

$$\hat{Q}^Z(z_{t+1}|z^t) = \frac{t + 1/2}{S_t + z_{t+1} + 1/2} \cdot \prod_{j=0}^{z_{t+1}} \frac{S_t + j + 1/2}{S_t + t + j + 1} \tag{12}$$

where

$$S_t = \sum_{i=1}^{t} z_i \tag{13}$$

and $S_0 \triangleq 0$ (cf. derivation in Equation (22) below). Finally, the sequential probability assignment associated with x^n is defined as

$$\hat{Q}(x^n) = \prod_{t=0}^{n-1} \hat{Q}(x_{t+1}|x^t) \tag{14}$$

where

$$\hat{Q}(x^{t+1}|x^t) = \hat{Q}^Y(y^{t+1}|y^t)\hat{Q}^Z(z^{t+1}|z^t). \tag{15}$$

Our main result in this section is summarized in the direct part of the following theorem.

Theorem 1 *Let $Q_{(\theta,\rho)}(x^n) = \prod_{t=1}^{n} Q_{(\theta,\rho)}(x_t)$.*

(a) (Converse part): Let $Q(x^n)$ be an arbitrary probability assignment. Then, for every $\epsilon > 0$,

$$E_{(\theta,\rho)} \log \frac{Q_{(\theta,\rho)}(X^n)}{Q(X^n)} \geq (1-\epsilon)\log n$$

for every $(\theta,\rho) \in (0,1) \times [0,1]$ except for points in a subset whose Lebesgue measure tends to zero as $n \to \infty$.

(b) (Direct part): Let $\hat{Q}(x^n)$ be defined as in equations (10)-(15). Then, for every $(\theta,\rho) \in (0,1) \times [0,1]$, and for every n-vector of integers x^n,

$$\log \frac{Q_{(\theta,\rho)}(x^n)}{\hat{Q}(x^n)} \leq \log n + \frac{1}{2}\log\left(\frac{S_n}{n} + 1\right) + C$$

where C is a constant that does not depend on n or x^n.

Discussion. Several comments regarding Theorem 1 are in order.

Lower bound. To show the applicability of Rissanen's lower bound [7, Theorem 1] for the off-centered TSG model, which corresponds to the converse part of the theorem, we reduce the problem to the well-known Bernoulli case, a special case in, e.g., [22, Theorem 1]. However, since [22, Theorem 1] requires that the parameters range in an interval that is bounded away from 0 and 1, for the sake of completeness we provide an independent proof. Furthermore, one can use the same tools to show the applicability of the bound in [23, Theorem 1], namely

$$\liminf_{n\to\infty} \frac{E_{(\theta,\rho)} \log[Q_{(\theta,\rho)}(X^n)/Q(X^n)]}{\log n} \geq 1$$

for all $(\theta, \rho) \in (0,1) \times [0,1]$ except in a set of Lebesgue measure zero.

Pointwise redundancy and expected redundancy. Strictly speaking, the minimum pointwise redundancy is not attained uniformly in x^n since S_n/n is arbitrarily large for some sequences. However, if the data actually has *finite* alphabet (which is practically the case in image compression), then S_n/n is uniformly bounded by a constant, and the minimum pointwise redundancy (w.r.t. the best model in the *infinite* alphabet class) is essentially attained. In any case, even if the alphabet is infinite, as assumed by the TSG model, the minimum *expected* redundancy is always attained since the expectation with respect to θ of $\log(S_n/n+1)$ is bounded by

$$E_\theta \log\left(\frac{S_n}{n}+1\right) \le \log\left(\frac{E_\theta S_n}{n}+1\right) = \log\left(\frac{1}{1-\theta}\right),$$

which is a constant.

Maximum likelihood estimation and the plug-in approach. For the class of finite-alphabet memoryless sources, parametrized by the letter probabilities, it is well-known that the mixture approach admits a direct "plug-in" implementation, where at each time instant, the parameter vector is first estimated by (a biased version of) the maximum likelihood (ML) estimator and then used to assign a probability distribution to the next outcome (see, e.g., the assignment (10)). It is interesting to observe that this plug-in interpretation does not exist with the OSG class, where the ML estimator for θ at time t, as well as for model (5), is given by

$$\tilde{\theta}_t = \frac{S_t}{S_t + t} \tag{16}$$

for sequences such that $S_t \neq 0$ (when $S_t = 0$ there is no ML estimator of θ in the range $(0,1)$). Nonetheless, an indirect plug-in mechanism is valid here: since the expression in (9) can be interpreted as the probability of a run of z zeros followed by a one under a Bernoulli process with parameter θ, then encoding an OSG source is equivalent to encoding the corresponding binary sequence of runs. In universal coding, while the biased ML estimator of [19] is used to update the estimate of θ after *every* bit, a direct, naive plug-in approach would correspond to updating the estimate of θ *only after occurrences of ones*, and hence may not perform as well.

To summarize, optimal encoding of x_{t+1} as per Theorem 1 and the ensuing discussion, can be realized with a sequence of $|x_{t+1}| - y(x_{t+1}) + 2$ binary encodings. First, we encode y_{t+1}, which determines whether x_{t+1} is negative. Then, we encode $|x_{t+1}| - y_{t+1}$ by first testing whether it is zero; in case it is positive, we proceed by inquiring whether it is one, and so forth. The corresponding probability estimates are based on S_t and N_t, which serve as sufficient statistics for the distribution (5).

The remaining part of this section is devoted to the proof of Theorem 1.

Proof of Theorem 1.

We begin with part (a). According to Rissanen's lower bound [7], it is sufficient to prove the existence of $\sqrt{n}$-consistent estimators $\hat{\rho}$ and $\hat{\theta}$ for ρ and for θ, respectively, such that the probabilities of the events $\{\sqrt{n}|\hat{\rho} - \rho| > c\}$ and $\{\sqrt{n}|\hat{\theta} - \theta| > c\}$ are both upper bounded by a function $\sigma(c)$ for all $n \geq n_c$, where $\sigma(c)$ and n_c do not depend on either θ or ρ, and $\sigma(c)$ tends to zero as $c \to \infty$.

For the parameter ρ, consider the estimator $\hat{\rho} = 1 - N_n/n$, calculated from the n observations of the Bernoulli process $y_1, ..., y_n$. Using the fact [24] that for $\alpha, \beta \in [0,1]$,

$$D_B(\alpha||\beta) \stackrel{\Delta}{=} \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta} \geq \frac{2(\alpha-\beta)^2}{\ln 2}\,, \tag{17}$$

the Chernoff bounding technique gives

$$\begin{aligned} \Pr\{\sqrt{n}|\hat{\rho} - \rho| \geq c\} &\leq \exp\{-n \ln 2 \min_{|\rho' - \rho| \geq c/\sqrt{n}} D_B(\rho'||\rho)\} \\ &\leq \exp\{-2n \min_{|\rho' - \rho| \geq c/\sqrt{n}} (\rho' - \rho)^2\} = \exp(-2c^2). \end{aligned} \tag{18}$$

As for the parameter θ, consider the estimator $\hat{\theta} = 1 - M_n/n$, where M_n is the number of zeros in $z_1, ..., z_n$.[2] Since the random variable given by the indicator function $1_{\{z_t = 0\}}$ is Bernoulli with parameter θ, then similarly to the derivations in (17) and (18), we again obtain $\Pr\{\sqrt{n}|\hat{\theta} - \theta| > c\} \leq \exp(-2c^2)$. Thus, $\sigma(c) = \exp(-2c^2)$, independently of ρ and θ, in this case. This completes the proof of part (a).

Turning now to part (b), we shall use the following relation, which confines [19, Equation (2.3)] to the binary alphabet case. For the Dirichlet(1/2) prior given by

$$w(\alpha) = \frac{[\alpha(1-\alpha)]^{-1/2}}{\Gamma(\alpha)\Gamma(1-\alpha)}, \qquad \alpha \in (0,1)$$

[2]Notice that this is not the ML estimator for θ.

and for nonnegative integers j and J ($j \le J$) we have:

$$\int_0^1 w(\alpha)\alpha^j(1-\alpha)^{J-j}d\alpha = \frac{\Gamma(j+\frac{1}{2})\Gamma(J-j+\frac{1}{2})}{\pi J!}. \tag{19}$$

Applying Stirling's formula, one obtains

$$-\log\int_0^1 w(\alpha)\alpha^j(1-\alpha)^{J-j}d\alpha \le Jh\left(\frac{j}{J}\right)+\frac{1}{2}\log J+\frac{C}{2} \tag{20}$$

where C is a constant that does not depend on j and J, and $h(u) \stackrel{\Delta}{=} -u\log u-(1-u)\log(1-u)$ is the binary entropy function.

Consider, first, universal coding of a binary string y^n using the Dirichlet(1/2) mixture over the class of Bernoulli sources with parameter ρ. Then, according to Equation (19) the mixture distribution is given by

$$\hat{Q}^Y(y^n) = \int_0^1 w(\rho)\rho^{n-N_n}(1-\rho)^{N_n}d\rho = \frac{\Gamma(N_n+\frac{1}{2})\Gamma(n-N_n+\frac{1}{2})}{\pi n!},$$

which can be written in a product form as $\prod_t \hat{Q}^Y(y_{t+1}|y^t)$, where each term is given as in Equation (10). According to Equation (20),

$$-\log\hat{Q}^Y(y^n) \le nh\left(\frac{N_n}{n}\right)+\frac{1}{2}\log n+\frac{C}{2}. \tag{21}$$

Consider, next, universal coding of z^n using the Dirichlet(1/2) mixture over the class of OSG distributions with parameter θ, that is,

$$\hat{Q}^Z(z^n) = \int_0^1 w(\theta)(1-\theta)^n\theta^{S_n}d\theta = \frac{\Gamma(n+\frac{1}{2})\Gamma(S_n+\frac{1}{2})}{\pi(S_n+n)!}, \tag{22}$$

which can be written in a product form as $\prod_t \hat{Q}^Z(z_{t+1}|z^t)$, where each term is given as in Equation (12). Again, (20) implies

$$\begin{aligned}-\log\hat{Q}^Z(z^n) &\le (S_n+n)h\left(\frac{S_n}{S_n+n}\right)+\frac{1}{2}\log(S_n+n)+\frac{C}{2}\\ &= (S_n+n)h\left(\frac{S_n}{S_n+n}\right)+\frac{1}{2}\log n+\frac{1}{2}\log\left(\frac{S_n}{n}+1\right)+\frac{C}{2}.\end{aligned} \tag{23}$$

On the other hand, for every (θ,ρ),

$$\begin{aligned}-\log Q_{(\theta,\rho)}(x^n) &\ge -\log\sup_{\theta,\rho} Q_{(\theta,\rho)}(x^n)\\ &= -\log\max_{\rho} Q^Y_\rho(y^n)-\log\sup_\theta Q^Z_\theta(z^n)\\ &= nh\left(\frac{N_n}{n}\right)+(S_n+n)h\left(\frac{S_n}{S_n+n}\right),\end{aligned} \tag{24}$$

where the last step follows from plugging the ML estimator (16) in the OSG distribution (9), with the equality holding trivially for $S_n = 0$. Combining equations (21), (23), and (24), we get

$$\begin{aligned} -\log \hat{Q}(x^n) &= -\log \hat{Q}^Y(y^n) - \log \hat{Q}^Z(z^n) \\ &\leq -\log Q_{(\theta,\rho)}(x^n) + \log n + \frac{1}{2}\log\left(\frac{S_n}{n}+1\right) + C. \end{aligned}$$

for any x^n and (θ, ρ). This completes the proof of Theorem 1. □

3 Low complexity adaptive codes

In Section 2, we presented an optimal strategy for encoding integers modeled by the extended TSGD (5). This strategy is also optimal for the TSG model (2), and requires arithmetic coding. In this section, we consider adaptive coding of the distribution (2) on a symbol-by-symbol basis, which normally incurs larger redundancy but is attractive from a practical point of view, e.g., in image-coding applications.[3] Even though, in general, adaptive strategies are easier to implement with arithmetic codes, the structured family of Huffman codes for TSG sources with known parameters introduced in the companion paper [17] provides an appealing alternative for low complexity adaptive coding of TSG models.

More generally, for a countable family of symbol-wise codes $\mathcal{C} = \{C^{(1)}, C^{(2)}, \cdots, C^{(j)}, \cdots\}$, consider an on-line algorithm that encodes x_{t+1} by selecting a code $C_t \in \mathcal{C}$, based on x^t. The performance of this on-line algorithm is measured by its average code length (under the unknown model parameters), and the objective is to perform essentially as well as the best fixed strategy in $\mathcal{C}$ for the unknown parameter values. This setting is akin to the sequential probability assignment problem studied in Section 2. However, unlike in Section 2, here the set of available coding strategies for each symbol is discrete, and the approach is inherently "plug-in." For a fixed code $C^* \in \mathcal{C}$, let $\Delta\overline{\lambda}(j)$ denote the expected per-symbol code length

[3]The use of symbol-by-symbol coding in low-complexity image compression systems is plausible, since contexts with very low entropy distributions, for which the optimal prefix code could be severely mismatched, are uncommon in photographic images. For other types of images, the redundancy of pixel-based prefix codes is addressed in [5] by embedding an alphabet extension into the conditioning model in "flat" regions that tend to present very peaked distributions.

difference between $C^{(j)}$ and C^*. Then, it can readily be verified that the expectation of the code length difference $\Delta\Lambda(x^n)$ over the entire sequence x^n is given by

$$E[\Delta\Lambda(X^n)] = \sum_{j=1}^{\infty} \Delta\overline{\lambda}(j) \sum_{t=1}^{n} \Pr\{C_t = C^{(j)}\}\,. \tag{25}$$

A particular plug-in approach, which relies on parameter estimation, is based on partitioning the parameter space into classes and assigning to each class a code in $\mathcal{C}$. Given a parameter estimate based on x^t, C_t is chosen as the code assigned to the corresponding class. Thus, (25) motivates the following on-line strategy for adaptive coding of a TSG model: Given an estimate of θ and d based on the sufficient statistics S_t and N_t (as defined in equations (11) and (13)), select C_t as the corresponding optimal prefix code prescribed by [17]. In this case, $\mathcal{C}$ is the family of Huffman codes from [17], the classes are the optimal decision regions for codes in $\mathcal{C}$ for given parameter values, and hence C^* in (25) is chosen as the actual Huffman code for the unknown parameters. If the probability $\Pr\{C_t = C^{(j)}\}$ decays rapidly enough for $C^{(j)} \neq C^*$ as the estimates converge to the true parameter values, and the average code length differences $\Delta\overline{\lambda}(j)$ are suitably bounded, then the per-symbol expected code length loss will be $O(1/n)$. An advantage of this strategy is that it depends only on S_t and N_t, as opposed to the popular plug-in approach of selecting the code that would have performed best on x^t. The latter approach was used in [25] to encode OSG distributions.

Code family. The family of optimal prefix codes from [17] is based on *Golomb codes* [13], whose structure enables simple calculation of the codeword of every given source symbol, without recourse to the storage of code tables, as would be the case with unstructured, generic Huffman codes. In an adaptive mode, a structured family of codes further relaxes the need of dynamically updating code tables due to possible variations in the estimated parameters (see, e.g., [12]). In [17], the parameter space (θ, d) is partitioned, and a different optimal prefix code corresponds to each class in the partition ($d \leq \frac{1}{2}$ is assumed, since the case $d > \frac{1}{2}$ can be reduced to the former by means of the reflection/shift transformation $x \to -(x+1)$ on the TSG-distributed variable x). Each class is associated with a Golomb code G_m, for which the "Golomb parameter" m is given by a many-to-one function of θ and d. Depending on the class, an integer x is encoded either by applying a class-dependent modification of G_m to a function of $|x|$, followed by a sign bit whenever $x \neq 0$, or as $G_m(M(x))$, where

$$M(x) = 2|x| - y(x) \tag{26}$$

is a one-to-one mapping onto the nonnegative integers (the indicator function $y(x)$ is defined in (7)). The mapping (26) gives the index of an integer in the interleaved sequence $0, -1, 1, -2, 2, \ldots$ and was first employed in [15]. Under the assumption $d \leq 1/2$, $M(x)$ orders integers by probability. For $d > \frac{1}{2}$, the relevant mapping is $M'(x) = M(-x-1)$. Notice that the codes $G_m(M(x))$ are *asymmetric*, in that x and $-x$ yield different code lengths for some integers x. In contrast, the codes based on $|x|$ and a possible sign bit are symmetric.

The on-line strategy suggested by (25) can be demonstrated with this family. However, even though arithmetic coding is avoided, both the region determination in order to find the optimal code for the estimated pair (θ, d), and the encoding procedure, may be too complex in some applications. For that reason, [17] considers a sub-family of codes used in practical lossless image compression schemes such as LOCO-I [5], which is based on Golomb codes for which the code parameter is a power of 2. Given an integer parameter $r \geq 0$, the code G_{2^r} encodes a nonnegative integer z in two parts: the r least significant bits of z, followed by the number formed by the remaining higher order bits of z, in *unary* representation. Furthermore, this sub-family uses only the asymmetric codes based on the mappings $M(\cdot)$ and $M'(\cdot)$, for which we denote $G_{2^r}(M(x)) \stackrel{\Delta}{=} \mathcal{G}_r(x)$. The mapping $M'(\cdot)$ is relevant only for $r = 0$, since $\mathcal{G}_r(x) = \mathcal{G}_r(-x-1)$ for every integer x and $r > 0$. We further denote $\mathcal{G}'_0(x) = G_1(M'(x))$. It is shown in [17, Corollary 1] that the average code length with the best code in $\mathcal{C}$ is within less than 5% of the optimal prefix codes for TSG distributions, with largest deterioration in the very low entropy region.[4]

Adaptive coding. In this section we follow the above practical compromise, and we consider adaptive coding for the reduced family of codes $\mathcal{C} = \{\mathcal{G}_r\} \bigcup \mathcal{G}'_0$. Similar derivations are possible with other sub-families, e.g., the one in [26] or the entire family of Huffman codes from [17]. In [26], some of the symmetric codes are included, leading to a more complex

[4]Despite the sub-optimality of this sub-family of codes, tests performed over a broad set of images used to develop the new ISO/IEC standard JPEG-LS [11] reveal that LOCO-I is within about 4% of the best available compression ratios (given by [6]) at a running time complexity close to an order of magnitude lower.

analysis for class determination. Theorem 2 below states that, in a probabilistic setting, an on-line strategy based on ML parameter estimation for the distribution (5) and a partition of the parameter space into optimal decision regions corresponding to codes in $\mathcal{C}$, performs essentially as well as the best code in $\mathcal{C}$. As in Theorem 1, the result is proved for the extended class (5), although the family $\mathcal{C}$ is motivated by the optimal prefix codes for the model class (2). The code selection for x_{t+1} is based on the sufficient statistics S_t and N_t.

Theorem 2 *Let $\phi \triangleq (\sqrt{5}+1)/2$. Encode x_{t+1}, $0 \le t < n$, according to the following decision rules:*

a. If $S_t \le \phi t$, compare S_t, $t-N_t$, and N_t. If S_t is largest, choose code $\mathcal{G}_1$. Otherwise, if $t-N_t$ is largest, choose $\mathcal{G}_0$. Otherwise, choose $\mathcal{G}'_0$.

b. If $S_t > \phi t$, choose code $\mathcal{G}_{r+1}$, $r \ge 1$ provided that

$$\frac{1}{\phi^{(2^{-r+1})}-1} < \frac{S_t}{t} \le \frac{1}{\phi^{(2^{-r})}-1}\,. \tag{27}$$

Let $\Lambda(x^n)$ denote the code length resulting from applying this adaptation strategy to the sequence x^n. Let $\Lambda^*(\theta,\rho)$ denote the minimum expected per-symbol codeword length over codes in $\mathcal{C}$ for the (unknown) parameters θ and ρ. Then,

$$\frac{1}{n}E_{(\theta,\rho)}[\Lambda(X^n)] \le \Lambda^*(\theta,\rho) + O\left(\frac{1}{n}\right)\,.$$

Discussion.

Other code families. Theorem 2 involves the decision regions derived in [17, Lemma 4] for $\mathcal{C}$ in the case of known parameters, substituting the estimates S_t/t and N_t/t for the parameters $S \triangleq \theta/(1-\theta)$ and $1-\rho$, respectively. However, a similar result would hold for any code family and partition of the parameter space, provided that mild regularity conditions on the difference between the expected code lengths for any pair of codes are satisfied.

Relation to prior work. A result analogous to Theorem 2 is proved in [25] for the alternative plug-in strategy of encoding x_{t+1} with the code that would have performed best on x^t, under an OSG distribution. There, the deviation from optimality is bounded as $O(1/\sqrt{n})$. Moreover, this alternative approach was analyzed *for individual data sequences* (as opposed to the probabilistic setting adopted here and in [25]) in the broader context of the sequential decision problem [27]. Specifically, this problem is about on-line selection of a certain strategy b_t, at each time instant t, depending on past observations x^t, so as to minimize a cumulative loss function $\sum_t l(b_t, x_{t+1})$ in the long run, for an arbitrary individual sequence x^n. It was shown in [27], that by allowing *randomized* selection of $\{b_t\}$, it is possible to approach optimum performance (in the expected value sense) within $O(1/\sqrt{n})$, uniformly for *every sequence* provided that the alphabet is finite. Here, adaptive coding is clearly a special case of the sequential decision problem, where the alphabet is in, practice, finite, b_t is a code C_t in the family $\mathcal{C}$, and $l(b_t, x_{t+1})$ is the corresponding code length for x_{t+1}. In this context, randomization would be applicable under the assumption that both encoder and decoder have access to a common random sequence. (A similar assumption is imposed in lossy compression schemes based on dithered quantization.) It should be pointed out that, in our case, there is indeed a difference between the two plug-in strategies, i.e., the one in [25] and the one proposed herein. For example, for the sequence $x^6 = 022222$, $S_6/6 = 10/6 > \phi$, so the approach based on ML estimation encodes x_7 with the code $\mathcal{G}_2$, whereas direct inspection reveals that the best code for x^6 is $\mathcal{G}_1$. In addition, notice that data compression as presented in Section 2 is clearly also a special case of the sequential decision problem, where b_t is a conditional probability assignment $p(\cdot|x^t)$ for x_{t+1} and $l(b_t, x_{t+1}) = -\log p(x_{t+1}|x^t)$. The sequential probability assignment problem differs from the adaptive coding problem treated in this section in that the set of available strategies is not discrete, and, hence, the results in [27] do not apply.

Low complexity approximation. The decision region boundaries (27) admit a low complexity approximation, for which it is useful to define the functions $S(r)$ and $\gamma(r)$, $r > 0$, by

$$S(r) \triangleq \frac{1}{\phi^{(2^{-r+1})} - 1} \triangleq \frac{2^{r-1}}{\ln \phi} - \frac{1}{2} + \gamma(r)\,. \tag{28}$$

It can be shown that $\gamma(r)$ is a *decreasing* function of r, that ranges between $\phi + \frac{1}{2} - (1/\ln\phi) \approx 0.04$ ($r = 1$), and 0 ($r \to \infty$). Since $\phi \approx 1.618$ and $1/\ln\phi \approx 2.078$, (28) implies that $S(r)$ is within 4% of $2^r - \frac{1}{2} + \frac{1}{8}$ for every $r > 0$. Thus, using approximate values of $S(r)$ and $S(r+1)$ in lieu of the bounds in (27), a good approximation to the decision rule of Theorem 2 for encoding x_{t+1} is:

Let $S'_t = S_t + (t/2) - (t/8)$.

a. If $S'_t \leq 2t$, compare S_t, N_t, and $t - N_t$. If S_t is larger, choose code $\mathcal{G}_1$. Otherwise, if $t - N_t$ is larger, choose $\mathcal{G}_0$. Otherwise, choose $\mathcal{G}'_0$.

b. If $S'_t > 2t$, choose code $\mathcal{G}_{r+1}$, $r \geq 1$ provided that

$$t2^r \leq S'_t < t2^{r+1}\,.$$

This simplified rule is used in LOCO-I [5] and it can be implemented with a few shift and add operations.

Proof of Theorem 2.

We apply (25) to the family of codes $\mathcal{C}$, with C_t chosen by the proposed on-line selection rule, and C^* denoting a code with minimum expected code length $\Lambda^*(\theta,\rho)$ over $\mathcal{C}$ for the (unknown) parameters θ and ρ. It suffices to prove that the right-hand side of (25) is upper-bounded by a constant as $n \rightarrow \infty$. The choice of optimal codes $C^* \in \mathcal{C}$ is given in [17, Lemma 4], where the decision regions for the parameters S and $1-\rho$ are the same as those in Theorem 2 for their estimates S_t/t and N_t/t. Although [17, Lemma 4] is shown for the model class (2), the proof is also valid for the extended class (5). Let $r \geq 0$ denote the index of a code in $\mathcal{C} = \{\mathcal{G}_r\} \bigcup \mathcal{G}'_0$, with both $\mathcal{G}_0$ and $\mathcal{G}'_0$ indexed by $r = 0$. Consider the increasing function $S(r)$ of r, $r \geq -1$, with $S(r)$ defined as in (28), $r > 0$, $S(0) \triangleq \max\{\rho, 1-\rho\}$, and $S(-1) \triangleq 0$. Let $\overline{r}^*$ and $\underline{r}^*$ denote the minimum and maximum integers, respectively, satisfying

$$S(\underline{r}^* - 1) < S < S(\overline{r}^*)\,. \tag{29}$$

By [17, Lemma 4], the optimal codes for (θ,ρ) are indexed by $\overline{r}^*$ and $\underline{r}^*$, where either $\underline{r}^* = \overline{r}^* - 1$ and (θ,ρ) lies on a code selection boundary, or $\overline{r}^* = \underline{r}^*$ is the only possible index for C^*. We divide the outer sum on the right-hand side of (25) in two parts, one corresponding to codes $\mathcal{G}_r$ such that $r > \overline{r}^*$, which yields a sum Δ_1, and one for the other non-optimal codes in $\mathcal{C}$ (codes $\mathcal{G}_r$ such that $r < \underline{r}^*$, and, if not optimal, $\mathcal{G}'_0$), which yields a sum Δ_2. Thus, (25) takes the form

$$E_{(\theta,\rho)}[\Lambda(x^n)] = n\Lambda^*(\theta,\rho) + \Delta_1 + \Delta_2\,. \tag{30}$$

We first upper-bound Δ_1. Clearly, if $r > \overline{r}^*$ then the code length difference between $\mathcal{G}_r$ and a code indexed by $\overline{r}^*$ can be at most $r - \overline{r}^*$ bits per encoding, due to a longer binary part using $\mathcal{G}_r$ ($r > \overline{r}^*$ cannot increase the unary part). Thus,

$$\Delta_1 \leq \sum_{r=\overline{r}^*+1}^{\infty} (r - \overline{r}^*) \sum_{t=1}^{n} \Pr\{C_t = \mathcal{G}_r\} = \sum_{r=\overline{r}^*}^{\infty} \sum_{t=1}^{n} \Pr\{r(t) > r\} \tag{31}$$

where $r(t)$ satisfies $\mathcal{G}_{r(t)} = C_t$. With $N_t^* \triangleq \max\{N_t, t-N_t\}$, the proposed on-line selection rule is such that

$$\begin{aligned} \Pr\{r(t) > r\} &= \Pr\{S_t > tS(r)\}\,, r > 0\,, \\ \Pr\{r(t) > 0\} &= \Pr\{S_t > N_t^*\}\,. \end{aligned} \tag{32}$$

First, assume $r > 0$, and define

$$\theta(r) \triangleq \frac{S(r)}{1 + S(r)} = (\phi - 1)^{(2^{-r+1})}\,, \tag{33}$$

which is also an increasing function of r. By (29), we have $1 > \theta(r) > \theta$ for all $r \geq \overline{r}^*$. In addition, the process $\{z_i\}$ defining S_t in Equation (13), Section 2, is distributed OSG (Equation (8)). It can be then seen that the Chernoff bounding technique gives

$$\Pr\{S_t > tS(r)\} \leq 2^{-tD(\theta(r)||\theta)} \tag{34}$$

where $D(\theta(r)||\theta)$ denotes the informational divergence between OSG sources with parameters $\theta(r)$ and θ, respectively, which is positive for $r \geq \overline{r}^*$.

Next, for $r = \overline{r}^* = 0$ and any real number S', we have

$$\Pr\{S_t > N_t^*\} \leq \Pr\{S_t > tS'\} + \Pr\{N_t^* < tS'\}\,. \tag{35}$$

By (29), $S < S(0)$. If $S \geq \frac{1}{2}$, choose any S' satisfying $S < S' < S(0)$; otherwise, let $S' = \frac{1}{2}$. Define $\theta(0) \triangleq S'/(1 + S')$, so that $1 > \theta(0) > \theta$. Clearly, (34) applies also for $r = 0$, but substituting S' for $S(0)$ on the left-hand side, thus bounding the first probability on the right-hand side of (35). Since $N_t^* \geq t/2$, the second probability is zero in case $S < \frac{1}{2}$. Otherwise, if $S \geq \frac{1}{2}$, since the process $\{y_i\}$ defining N_t in Equation (11) is Bernoulli with $\Pr\{Y = 0\} = \rho$ (see Section 2), the Chernoff bounding technique further yields

$$\Pr\{N_t^* < tS'\} = \Pr\{1 - S' < \frac{N_t}{t} < S'\} \leq 2^{-tD_B(S'||S(0))} \tag{36}$$

where the informational divergence $D_B(\cdot||\cdot)$ *for Bernoulli processes* is defined in Equation (17), Section 2.

It then follows from (31) through (36) that for all $\overline{r}^*$

$$\Delta_1 \leq \sum_{r=\overline{r}^*}^{\infty} \frac{1}{2^{D(\theta(r)||\theta)}-1} + \mathcal{F}(S,\rho) \tag{37}$$

where

$$\mathcal{F}(S,\rho) = \sum_{t=1}^{\infty} 2^{-tD_B(S'||S(0))} = \frac{1}{2^{D_B(S'||S(0))}-1}$$

for $\frac{1}{2} \leq S < S(0)$, and $\mathcal{F}(S,\rho)$=0 otherwise. Thus, in order to upper-bound Δ_1 with a constant (that depends only on the actual parameters S and ρ), it suffices to prove the convergence of the series $\sum_{r=\overline{r}^*}^{\infty} 2^{-D(\theta(r)||\theta)}$. It can readily be verified that

$$D(\theta(r)||\theta) = \frac{D_B(\theta(r)||\theta)}{1-\theta(r)} \geq \frac{2(\theta(r)-\theta)^2}{(1-\theta(r))\ln 2} \geq \frac{2(\theta(\overline{r}^*+1)-\theta)^2}{(1-\theta(r))\ln 2} \triangleq \frac{\kappa(\theta,\rho)}{1-\theta(r)} \tag{38}$$

where the first inequality follows from (17), the second holds for every $r > \overline{r}^*$, and $\kappa(\theta,\rho)$ is a positive constant that depends only on θ and ρ. In addition, it follows from (33) and (28) that for all $r \geq 1$

$$\frac{1}{1-\theta(r)} = S(r)+1 > \frac{2^{r-1}}{\ln\phi}\,. \tag{39}$$

Clearly, (38) and (39) imply

$$\Delta_1 < \infty\,. \tag{40}$$

As for the sum Δ_2, we consider two cases: $\underline{r}^* > 0$ and $\underline{r}^* = 0$. In the first case, codes indexed by r, $0 \leq r < \underline{r}^*$, encode an integer x with at most $M(x)$ bits more than $\mathcal{G}_{\underline{r}^*}$, due to a longer unary part (the binary part decreases at least by one). Thus, the expected code length increase in (25) is uniformly upper-bounded for all codes that contribute to Δ_2, implying

$$\Delta_2 \leq E_{(\theta,\rho)}[M(x)] \sum_{t=1}^{n} \Pr\{r(t) < \underline{r}^*\}$$

where $r(t)$ is the index of C_t. Since $M(x) = 2z+y$, with z defined in Equation (8) and distributed OSG with parameter θ, and y defined in Equation (7) and Bernoulli with $\Pr\{Y=0\} = \rho$ (see Section 2), we have $E_{(\theta,\rho)}[M(x)] = 2S+1-\rho$. For $\underline{r}^* > 1$, $r(t) < \underline{r}^*$ if and only if $S_t \leq tS(\underline{r}^*-1)$.

Since, by (29), we have $\theta > \theta(\underline{r}^* - 1)$, using again the Chernoff bounding technique we obtain

$$\Delta_2 \leq \left[\frac{1+\theta}{1-\theta} - \rho\right] \sum_{t=1}^{\infty} 2^{-tD(\theta(\underline{r}^*-1)||\theta)} = \left[\frac{1+\theta}{1-\theta} - \rho\right] \frac{1}{2^{D(\theta(\underline{r}^*-1)||\theta)} - 1} . \tag{41}$$

For $\underline{r}^* = 1$, the case $r(t) = 0$ arises if and only if $S_t \leq N_t^*$. Since (29) implies $S > S(0)$, we can choose S' such that $S(0) < S' < S$, and define $\theta(0) \triangleq S'/(1+S')$, to obtain

$$\begin{aligned} \Pr\{S_t \leq N_t^*\} &\leq \Pr\{S_t \leq tS'\} + \Pr\{N_t \geq tS'\} + \Pr\{N_t \leq t(1-S')\} \\ &\leq 2^{-tD(\theta(0)||\theta)} + 2^{-tD_B(S'||\rho)} + 2^{-tD_B(S'||1-\rho)} \end{aligned}$$

which again yields a constant upper bound on Δ_2. Finally, in the case $\underline{r}^* = 0$, the magnitude of the average code length discrepancy between $\mathcal{G}_0$ and $\mathcal{G}'_0$ is $2S(0) - 1$. In addition, the decision between the two codes is governed by N_t, implying

$$\Delta_2 \leq (2S(0) - 1) \sum_{t=1}^{\infty} 2^{-tD_B(\frac{1}{2}||\rho)} = \frac{2S(0) - 1}{2^{D_B(\frac{1}{2}||\rho)} - 1} \tag{42}$$

for $\rho \neq \frac{1}{2}$, and $\Delta_2 = 0$ otherwise. Theorem 2 follows from equations (30) and (40) through (42). □

References

[1] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–664, Sept. 1983.

[2] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 643–652, May 1995.

[3] S. Todd, G. G. Langdon, Jr., and J. Rissanen, "Parameter reduction and context selection for compression of the gray-scale images," *IBM Jl. Res. Develop.*, vol. 29 (2), pp. 188–193, Mar. 1985.

[4] M. J. Weinberger, J. Rissanen, and R. Arps, "Applications of universal context modeling to lossless compression of gray-scale images," *IEEE Trans. Image Processing*, vol. 5, pp. 575–586, Apr. 1996.

[5] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," 1998. Submitted to *IEEE Trans. Image Proc.* Available as Hewlett-Packard Laboratories Technical Report.

[6] X. Wu and N. D. Memon, "Context-based, adaptive, lossless image coding," *IEEE Trans. Commun.*, vol. 45 (4), pp. 437–444, Apr. 1997.

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

[8] A. Netravali and J. O. Limb, "Picture coding: A review," *Proc. IEEE*, vol. 68, pp. 366–406, 1980.

[9] J. O'Neal, "Predictive quantizing differential pulse code modulation for the transmission of television signals," *Bell Syst. Tech. J.*, vol. 45, pp. 689–722, May 1966.

[10] M. J. Weinberger, G. Seroussi, and G. Sapiro, "LOCO-I: A low complexity, context-based, lossless image compression algorithm," in *Proc. 1996 Data Compression Conference*, (Snowbird, Utah, USA), pp. 140–149, Mar. 1996.

[11] ISO/IEC JTC1/SC29 WG1 (JPEG/JBIG), "Information technology - Lossless and near-lossless compression of continuous-tone still images," 1998. Final Draft International Standard FDIS14495-1 (JPEG-LS). Also, ITU Recommendation T.87.

[12] D. E. Knuth, "Dynamic Huffman coding," *J. Algorithms*, vol. 6, pp. 163–180, 1985.

[13] S. W. Golomb, "Run-length encodings," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 399–401, July 1966.

[14] R. Gallager and D. V. Voorhis, "Optimal source codes for geometrically distributed integer alphabets," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 228–230, Mar. 1975.

[15] R. F. Rice, "Some practical universal noiseless coding techniques - parts I-III," Tech. Rep. JPL-79-22, JPL-83-17, and JPL-91-3, Jet Propulsion Laboratory, Pasadena, CA, Mar. 1979, Mar. 1983, Nov. 1991.

[16] K.-M. Cheung and P. Smyth, "A high-speed distortionless predictive image compression scheme," in *Proc. of the 1990 Int'l Symposium on Information Theory and its Applications*, (Honolulu, Hawaii, USA), pp. 467–470, Nov. 1990.

[17] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for two-sided geometric distributions," 1998. Submitted to *IEEE Trans. Inform. Theory*. Available as Technical Report No. HPL-94-111, Apr. 1998, Hewlett-Packard Laboratories.

[18] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

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

[20] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 714–722, May 1995.

[21] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 384–396, Mar. 1994.

[22] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, July 1986.

[23] J. Rissanen, "Stochastic complexity and modeling," *Annals of Statistics*, vol. 14, pp. 1080–1100, Sept. 1986.

[24] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[25] P. G. Howard and J. S. Vitter, "Fast and efficient lossless image compression," in *Proc. 1993 Data Compression Conference*, (Snowbird, Utah, USA), pp. 351–360, Mar. 1993.

[26] G. Seroussi and M. J. Weinberger, "On adaptive strategies for an extended family of Golomb-type codes," in *Proc. 1997 Data Compression Conference*, (Snowbird, Utah, USA), pp. 131–140, Mar. 1997.

[27] J. F. Hannan, "Approximation to Bayes risk in repeated plays," *Contributions to the Theory of Games, Vol. III, Annals of Mathematics Studies*, pp. 97–139, Princeton, NJ, 1957.