Lossless Compression for Sources with Two-Sided Geometric Distributions *

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 off-centered, two-sided geometric (TSG) distribution, which is a commonly used statistical model for image prediction residuals. In the first part of 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. In the second part, we focus on more practical codes for the TSG, that operate on a symbol-by-symbol basis. Specifically, we present a complete characterization of minimum expected-length prefix codes for TSG sources. The family of optimal codes introduced here is an extension of the Golomb codes, which are optimal for one-sided geometric distributions of nonnegative integers. As in the one-sided case, the resulting optimum Huffman tree has a structure that enables simple calculation of the codeword of every given source symbol. Our characterization avoids the heuristic approximations frequently used when modifying Golomb codes so as to apply to two-sided sources. Finally, we provide adaptation criteria for a further simplified, sub-optimal family of codes used in practice.

Index Terms: Lossless image compression, Huffman code, 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 mass function (PMF) $p(\cdot|x^t)$ 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 instantaneous conditional PMF $p(\cdot|x^t)$ 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 tends to zero as $c \to \infty$ uniformly in Ψ. 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,

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

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 PMF given by the *two-sided geometric* (TSG) distribution model. It has been observed [9] that prediction errors are well-modeled by the TSG distribution centered at zero, henceforth referred to as *centered TSG*. According to this distribution, the probability of an integer value x of the prediction error ($x = 0, \pm 1, \pm 2, ...$), 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 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])[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. It is also useful for better capturing the two adjacent modes often observed in empirical context-dependent histograms of prediction errors. More precisely, the more general model is described next. First, since the outcomes of a source are conditionally independent given their contexts, 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 TSG distribution 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 $\{P_\psi\}$ defined as follows. Let $\psi = (\theta, d)$ (hence $K = 2$), and let the marginal PMF of a symbol be 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

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

is a normalization factor, $0 < \theta < 1$ as above, and $0 \le d \le 1$. This limited range of d, which corresponds to PMF modes at 0 and -1, can be obtained by a suitable error feedback loop [5, 6]. The centered TSG distribution corresponds to $d = 0$, and, when $d = \frac{1}{2}$, $P_{(\theta,d)}$ is a

[1] The algorithm in [5] has recently been adopted as the baseline for the lossless image compression standard JPEG-LS [10].

bi-modal distribution with equal peaks at -1 and 0. (The preference of -1 over $+1$ here is arbitrary). The case $d = -1$ is essentially identical to the centered TSG.

This paper consists of three parts. Section 2 focuses on universal probability assignment for the TSG model (2), for which the bound (1) applies (with $K = 2$) as discussed in that section. This assignment is used as a preamble for arithmetic coding. Specifically, it is demonstrated that arithmetic coding based on a simple strategy of model adaptation, essentially attains the optimal universal coding redundancy prescribed by the lower bound (1). This strategy is derived by the method of mixtures. 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.

The remaining sections are devoted to Huffman coding on a symbol-by-symbol basis, which normally incurs larger redundancy, but is more attractive from a practical point of view. It can be readily verified that the TSG distribution (2) has a finite entropy, given by

$$H(\theta, d) = \frac{h(\theta)}{1-\theta} + h(\rho)\,,$$

where $h(u) \stackrel{\Delta}{=} -u \log u - (1-u)\log(1-u)$ is the binary entropy function and

$$\rho = \frac{\theta^d}{\theta^{1-d} + \theta^d} \tag{4}$$

is the probability that a random variable drawn according to the distribution (2) be non-negative. By [11], this guarantees that a minimum expected-length prefix code exists and can be obtained by a sequence of Huffman-like procedures (however, this general result is non-constructive). Infinite entropy distributions are addressed in [12]. We first develop (Section 3) a complete characterization of minimum expected-length prefix codes for the TSG sources in (2) assuming known values of θ and d. The family of optimal prefix codes introduced here is an extension of the Golomb codes [13], which are optimal for *one-sided geometric* (OSG) distributions of nonnegative integers [14]. As in the one-sided case, the structure of the resulting optimum Huffman tree enables simple calculation of the codeword of every given source symbol, without recourse to the storage of code tables for large alphabets. The complexity of this calculation is essentially the same as that of Golomb codes. Previous approaches to the problem have focused mainly on the case $d = 0$. A popular approach [15] is to encode an integer by applying a Golomb code to its index in the sequence $0, -1, +1, -2, +2, -3, +3, ...$ When $d \leq 0.5$, this "folding" of the negative values into the positive ones ranks the integers according to non-increasing probabilities. A different heuristic approach, based on encoding the absolute value with a Golomb code and appending a sign bit for nonzero values, was proposed in [16]. As shown in Section 3, these strategies are sub-optimal for some ranges of the parameters (θ, d), even when restricted to the line $d = 0$. Some partial answers to the question of optimal codes for $d = 0$ can also be found in [17].

Finally, in Section 4, we relax the assumption that θ and d are known a-priori, in the framework of symbol-by-symbol coding. Unlike in Section 2, here the set of available

coding strategies for each symbol is discrete, and the adaptation approach is inherently "plug-in." We provide optimal adaptation criteria (in a well-defined sense) for a further simplified, sub-optimal family of codes used in practice [15, 5]. It should be pointed out that in the adaptive mode, 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., [18]).

2 Universal Probability Assignment for TSG's

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., [19, 20, 21]). 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

$$\hat{Q}(x^n) = \int_{\Psi} dw(\psi) P_{\psi}(x^n) \tag{5}$$

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., [22]), 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 (5) for every x^n, and the instantaneous probability assignments are easy to derive, the TSG distribution 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 distribution model, which gives a somewhat larger class of PMF's, 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

$$Q_\varphi(x) \stackrel{\Delta}{=} Q_{(\theta,\rho)}(x) = \begin{cases} \rho(1-\theta)\theta^x & x = 0, 1, 2, ... \\ (1-\rho)(1-\theta)\theta^{-x-1} & x = -1, -2, ... \end{cases} \tag{6}$$

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 (6) 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 (6), with the same value for the parameter θ and with the parameter ρ given by the relation (4). However, while the original TSG model allows only for $\rho \in [\theta/(1+\theta), 1/(1+\theta)]$ for a given θ, the model (6) permits any $\rho \in [0, 1]$. It follows that the modified TSG model (6) 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 (6), 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

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

and

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

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 (6), then $Y = y(X)$ and $Z = z(X)$ are independent, where Y is binary $\{0, 1\}$ with parameter $\rho \stackrel{\Delta}{=} Q_\rho^Y(0) \stackrel{\Delta}{=} \Pr\{Y = 0\}$, and Z is OSG with parameter θ, that is,

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

Accordingly, given a memoryless source $X_1, X_2, ...$ with marginal PMF given by (6), one creates, using $y(\cdot)$ and $z(\cdot)$, two independent memoryless processes, $Y_1, Y_2, ... \sim Q_\rho^Y$ and $Z_1, Z_2, ... \sim Q_\theta^Z$, 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 [20]

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

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 $N_0 \stackrel{\Delta}{=} 0$. This probability assignment is induced by a mixture of type (5) 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 $S_0 \stackrel{\Delta}{=} 0$ (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., [23, Theorem 1]. However, since [23, 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 [24, 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

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

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 process, where the ML estimator for θ at time t, as well as for model (6), 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 [20] 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 $|x_{t+1}| - f(x_{t+1}) + 2$ 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 (6).

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 [25] that for $\alpha, \beta \in [0,1]$,

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

the Chernoff bounding technique gives

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

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 [20, Equation

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

(2.3)] to the binary alphabet case. For the Dirichlet(1/2) prior given by

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

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.
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's 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),

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

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 Optimal prefix codes for TSG's

In Section 2, we presented an optimal strategy for encoding integers modeled by the modified TSG distribution (6). This strategy is also optimal for the TSG model (2), and requires arithmetic coding. In this section, we consider Huffman 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] Throughout this section the values of the parameters θ and d are assumed known. Moreover, the offset parameter d is assumed to be in the range $0 \leq d \leq \frac{1}{2}$. Clearly, the case $d > \frac{1}{2}$ can be reduced to $d < \frac{1}{2}$ by means of the reflection/shift transformation $x \to -(x+1)$ on the TSG-distributed variable x. The parameter θ, in turn, belongs to the open interval $(0, 1)$. Next, we develop a complete characterization of minimum expected-length prefix codes for the TSG models (2). To this end, we will partition the parameter space (θ, d) into regions, each region corresponding to a variant of a basic code construction device. In the next few definitions and lemmas, we define the partition and some of its basic properties.

For a given d, define $\delta \stackrel{\Delta}{=} \min\{d, \frac{1}{2} - d\}$. Clearly, $\delta \leq \frac{1}{4}$. For every positive integer ℓ and every pair of model parameters (θ, d), define the functions

$$r_0(\ell, \theta, d) = \theta^{2\ell-1}(1 + \theta^{-2\delta}) + \theta^{\ell-1} - 1, \tag{25}$$

$$r_1(\ell, \theta, d) = \theta^{2\ell-1}(1 + \theta^{2\delta}) + \theta^{\ell} - 1, \tag{26}$$

$$r_2(\ell, \theta, d) = \theta^{\ell}(1 + \theta^{-2\delta}) - 1, \tag{27}$$

and

$$r_3(\ell, \theta, d) = \theta^{\ell}(1 + \theta^{2\delta}) - 1. \tag{28}$$

Lemma 1 *(i) Given $\ell > 1$ and d, r_0 has a unique root $\theta_0(\ell, d) \in (0, 1)$. Similarly, for $\ell \geq 1$, r_1, r_2, and r_3 have unique roots in $(0, 1)$, denoted respectively $\theta_1(\ell, d), \theta_2(\ell, d)$, and $\theta_3(\ell, d)$.*

(ii) For $\theta \in (0, 1)$ and $0 \leq i \leq 3$, $\theta \leq \theta_i(\ell, d)$ if and only if $r_i(\ell, \theta, d) \leq 0$.

[3] The use of symbol-by-symbol coding in low-complexity image compression systems is plausible, since contexts with very skewed prediction error 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 context conditioning for contexts representing flat regions, which tend to present very skewed distributions.

(iii) For $\ell \geq 1$,

$$\theta_0(\ell, d) < \theta_1(\ell, d) \leq \theta_2(\ell, d) \leq \theta_3(\ell, d) \leq \theta_0(\ell + 1, d)\,,$$

where we define $\theta_0(1, d) = 0$. Moreover, equality between $\theta_1(\ell, d)$ and $\theta_2(\ell, d)$, and between $\theta_3(\ell, d)$ and $\theta_0(\ell + 1, d)$ occurs only at $d = \frac{1}{4}$, while equality between $\theta_2(\ell, d)$ and $\theta_3(\ell, d)$ occurs only at $d \in \{0, \frac{1}{2}\}$. Therefore, $\theta_1(\ell, d) < \theta_0(\ell + 1, d)$.

Proof. (i) The existence and uniqueness of a root $\theta_i(\ell, d) \in (0, 1)$ of r_i, $0 \leq i \leq 3$, is established by observing that, for fixed ℓ and d in the appropriate ranges, r_i is a continuous function of θ in $(0, 1)$, $r_i(\ell, \theta, d) \to -1$ as $\theta \to 0^+$, $r_i(\ell, \theta, d)$ has a positive limit as $\theta \to 1^-$, and $\partial r_i/\partial\theta > 0$, $\theta \in (0, 1)$. The monotonicity of r_i also yields part (ii) of the lemma. Notice that $r_0(1, \theta, d) \to 0$ as $\theta \to 0^+$, justifying the definition of $\theta_0(1, d) = 0$.
As for part (iii), we first observe that

$$r_1(\ell, \theta, d) - r_0(\ell, \theta, d) = \theta^{2\ell-1}(\theta^{2\delta} - \theta^{-2\delta}) + (\theta^{\ell} - \theta^{\ell-1}) < 0,$$

where the last inequality follows from $\theta < 1$, $\delta \geq 0$, and $\ell \geq 1$. Thus, due to the strict monotonicity of r_0 and r_1, we must have $\theta_0(\ell, d) < \theta_1(\ell, d)$. We now compare $\theta_2(\ell, d)$ with $\theta_1(\ell, d)$. For clutter reduction, we omit the arguments (ℓ, d) of the θ_i when they are clear from the context. It follows from the definition of θ_2 that

$$\theta_2^{\ell} = \frac{1}{1 + \theta_2^{-2\delta}}.$$

Substituting θ_2 for θ in definition (26), we obtain

$$r_1(\ell, \theta_2, d) = \frac{\theta_2^{-1}}{(1 + \theta_2^{-2\delta})^2}(1 + \theta_2^{2\delta}) + \frac{1}{1 + \theta_2^{-2\delta}} - 1 = \frac{\theta_2^{4\delta-1} - 1}{1 + \theta_2^{2\delta}} \geq 0,$$

where the last inequality follows from $\delta \leq \frac{1}{4}$. Thus, by part (ii), $\theta_2(\ell, d) \geq \theta_1(\ell, d)$, with equality occurring at $d = \delta = \frac{1}{4}$.
Next, definitions (27) and (28) imply $r_2(\ell, \theta, d) \geq r_3(\ell, \theta, d)$ for $\theta \in (0, 1)$. Thus, we must have $\theta_2 \leq \theta_3$ by parts (i) and (ii) of the lemma. Equality occurs at $d \in \{0, \frac{1}{2}\}$, in which case $\delta = 0$, and $r_2(\ell, \theta, d) = r_3(\ell, \theta, d)$. Also, we have

$$\theta_3^{\ell} = \frac{1}{1 + \theta_3^{2\delta}}.$$

Substituting θ_3 for θ in the expression for $r_0(\ell + 1, \theta, d)$ derived from definition (25), we obtain

$$r_0(\ell + 1, \theta_3, d) = \frac{\theta_3}{(1 + \theta_3^{2\delta})^2}(1 + \theta_3^{-2\delta}) + \frac{1}{1 + \theta_3^{2\delta}} - 1 = \frac{\theta_3^{1-2\delta} - \theta_3^{2\delta}}{1 + \theta_3^{2\delta}} \leq 0.$$

Thus, $\theta_3(\ell, d) \leq \theta_0(\ell + 1, d)$, with equality at $d = \delta = \frac{1}{4}$. □

It follows from Lemma 1 that, for a given value of d, the r_i define a partition of the interval $(0,1)$ into sub-intervals, with boundaries given by the values $\theta_i(\ell,d)$ in lexicographically increasing order of (ℓ,i). Namely, we have

$$\begin{aligned} 0=\theta_0(1,d)<\theta_1(1,d) &\leq \theta_2(1,d)\leq\theta_3(1,d)\leq\theta_0(2,d)<\theta_1(2,d)\leq\ \cdots \\ \cdots &\leq \theta_0(\ell,d)<\theta_1(\ell,d)\leq\theta_2(\ell,d)\leq\theta_3(\ell,d)\leq\ \cdots\ <\ 1. \end{aligned} \tag{29}$$

Moreover, it is easy to see from definition (25) that $\theta_0(\ell,d)\to 1$ as $\ell\to\infty$.
The different intervals defined by the boundaries θ_i become two-dimensional regions once the dependence on d is taken into account. Each pair of model parameters (θ,d) falls in a region characterized by an integer parameter $\ell(\theta,d)$, and by a sub-interval corresponding to ℓ. By Lemma 1, part (ii), the parameter $\ell(\theta,d)$ is given by

$$\ell(\theta,d)=\max_{\ell\in\mathbf{Z}^+}\{\ell\,|\,r_0(\ell,\theta,d)>0\}. \tag{30}$$

Since $\lim_{\ell\to\infty}\theta_0(\ell,d)=1$, $\ell(\theta,d)$ is well defined for all θ and d in the range of interest. In fact, $\ell(\theta,d)$ can be explicitly computed by setting $z=\theta^\ell$ and solving the quadratic equation

$$z^2(1+\theta^{-2\delta})+z-\theta=0,$$

which has a unique solution z_0 in the open interval $(0,1)$. Then,

$$\ell(\theta,d)=\left\lfloor\frac{\log z_0}{\log\theta}\right\rfloor.$$

For ease of reference, we label the regions defined by the partition in (29) as follows:

Region I: $\theta_0(\ell,d)<\theta\leq\theta_1(\ell,d)$,

Region II: $\theta_1(\ell,d)<\theta\leq\theta_2(\ell,d)$, $d\leq\frac{1}{4}$,

Region II': $\theta_1(\ell,d)<\theta\leq\theta_2(\ell,d)$, $d>\frac{1}{4}$,

Region III': $\theta_2(\ell,d)<\theta\leq\theta_3(\ell,d)$,

Region IV: $\theta_3(\ell,d)<\theta\leq\theta_0(\ell+1,d)$, $d\leq\frac{1}{4}$,

Region IV': $\theta_3(\ell,d)<\theta\leq\theta_0(\ell+1,d)$, $d>\frac{1}{4}$.

We define *Region III* as the union of regions II', III' and IV'. The different two-dimensional regions for $\ell=1,2$ are illustrated in Figure 1. Notice the symmetry around $d=\frac{1}{4}$.
We now turn to the basic devices of our code construction. For any integer x, define

$$M(x)=\begin{cases} 2x & x\geq 0,\\ 2|x|-1 & x<0. \end{cases} \tag{31}$$

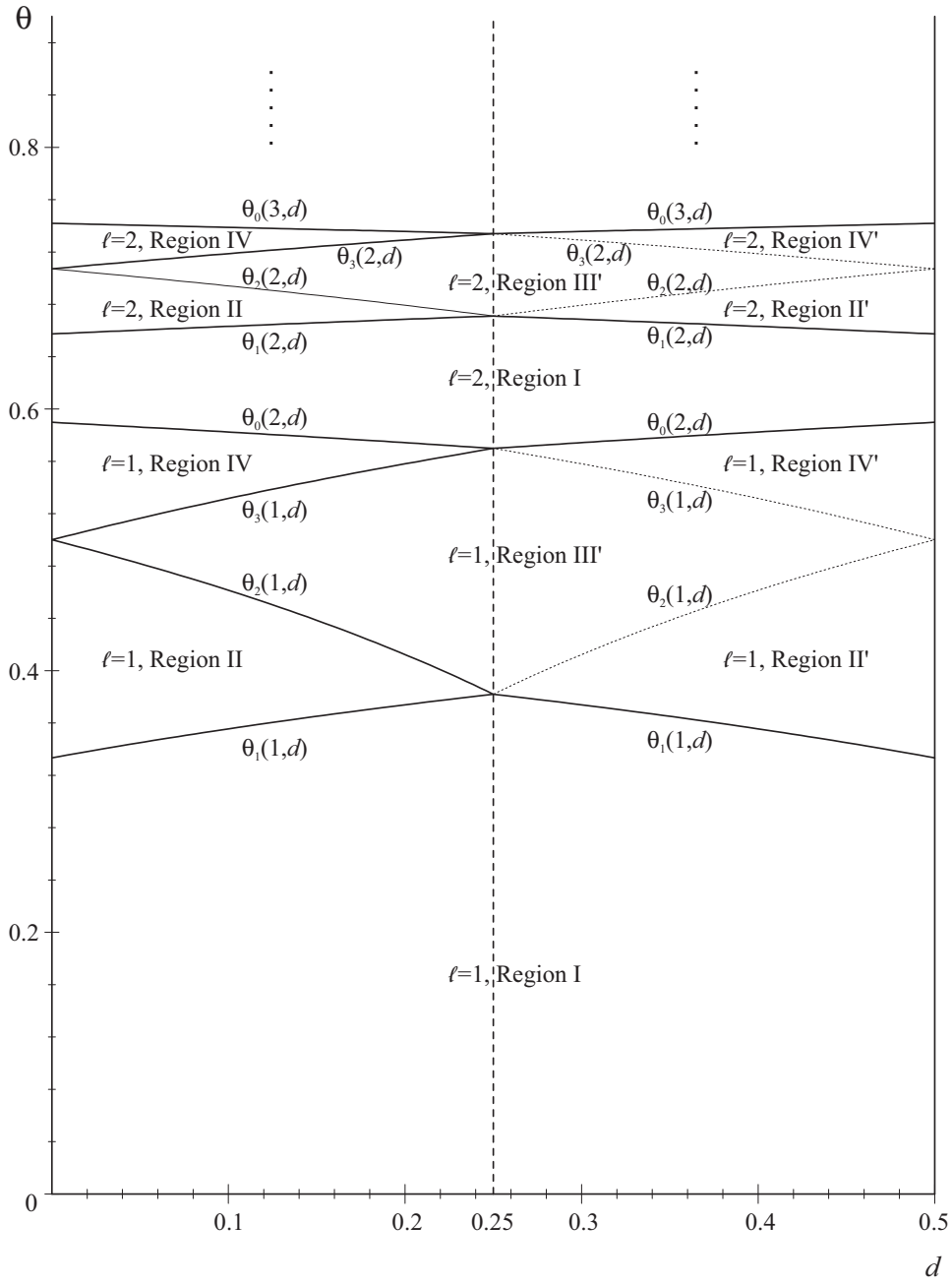

Figure 1: Parameter regions. Region III is defined as the union of regions II', III', and IV'.

For nonnegative integers i, the inverse function $\mu(i)$ of M is given by

$$\mu(i) = \begin{cases} i/2 & i \text{ even}, \\ -(i+1)/2 & i \text{ odd}. \end{cases}$$

Since $0 \le d \le \frac{1}{2}$, the integers are ranked in decreasing probability order by

$$P_{(\theta,d)}(0) \ge P_{(\theta,d)}(-1) \ge P_{(\theta,d)}(1) \ge P_{(\theta,d)}(-2) \ge P_{(\theta,d)}(2) \ge \cdots \tag{32}$$

Thus, $M(x)$ is the index of x in the probability ranking, starting with index 0 and with ties, if any, broken according to the order given in (32). Conversely, $\mu(i)$ is the symbol with the ith highest probability.

For any positive integer L, let G_L denote the Golomb code [13] of order L, which encodes a non-negative integer y in two parts: (a) an *adjusted binary* representation of $y' = y \bmod L$,[4] using $\lfloor \log L \rfloor$ bits if $y' < 2^{\lceil \log L \rceil} - L$, $\lceil \log L \rceil$ bits otherwise, and (b) a *unary* representation of $q = \lfloor y/L \rfloor$, using $q+1$ bits.

We are now ready to state the main result of this section.

Theorem 2 *Let x denote an integer-valued random variable distributed according to (2) and (3) for a given pair of model parameters (θ, d), $0 < \theta < 1$, $0 \le d \le \frac{1}{2}$, and let $\ell = \ell(\theta, d)$ as defined in (30). Then, an optimal prefix code for x is constructed as follows:*

(**Region I**) *If $\theta_0(\ell, d) < \theta \le \theta_1(\ell, d)$, encode x using $G_{2\ell-1}(M(x))$.*

(**Region II**) *If $d \le \frac{1}{4}$ and $\theta_1(\ell, d) < \theta \le \theta_2(\ell, d)$, encode $|x|$ using the code $G_\ell(\chi_\ell(|x|))$, where the mapping $\chi_\ell(|x|)$ is defined below, and append a sign bit whenever $x \ne 0$. Let r be the integer satisfying $2^{r-1} \le \ell < 2^r$, and let $s = 2^r - \ell$. Define*

$$\chi_\ell(|x|) = \begin{cases} s, & |x| = 0 \text{ and } s \ne \ell, \\ 0, & |x| = s \text{ and } s \ne \ell, \\ |x|, & \text{otherwise}. \end{cases}$$

(**Region III**) *If $d \le \frac{1}{4}$ and $\theta_2(\ell, d) < \theta \le \theta_3(\ell, d)$, or $d > \frac{1}{4}$ and $\theta_1(\ell, d) < \theta \le \theta_0(\ell+1, d)$, encode x using $G_{2\ell}(M(x))$.*

(**Region IV**) *If $d \le \frac{1}{4}$ and $\theta_3(\ell, d) < \theta \le \theta_0(\ell+1, d)$, define s as in Region II, encode $|x|$ using $J_\ell(|x|)$ defined below, and append a sign bit whenever $x \ne 0$.*

$$J_\ell(|x|) = \begin{cases} G_\ell(|x|-1), & |x| > s, \\ G_\ell(|x|), & 1 \le |x| < s, \\ G_\ell(0)0, & x = 0, \\ G_\ell(0)1, & |x| = s. \end{cases}$$

[4] $a \bmod b$ denotes the least nonnegative residue of a modulo b

Discussion.

Relation to prior work. Theorem 2 includes the main result of [14] as a special case when $d = \frac{1}{4}$. In this case, the distribution (2), after reordering of the integers in decreasing probability order, is equivalent to an OSG distribution with parameter $\phi = \sqrt{\theta}$. As shown in [14], the optimality transition for such a distribution between the L-th order Golomb code and the $L+1$st, $L \geq 1$, occurs at the (unique) value $\phi \in (0,1)$ such that $g_L(\phi) = \phi^L + \phi^{L-1} - 1 = 0$. It can be readily verified that $r_0(\ell, \phi^2, \frac{1}{4}) = 0$ if and only if $g_{2\ell-1}(\phi) = 0$, and $r_1(\ell, \phi^2, \frac{1}{4}) = 0$ if and only if $g_{2\ell}(\phi) = 0$, $\ell \geq 1$.

Notice that the optimal codes for regions I and III are *asymmetric*, in that they assign different code lengths to x and $-x$ for some values of x. In contrast, the codes for regions II and IV are symmetric. The mapping (31) was first employed in [15] to apply Golomb codes to alphabets that include both positive and negative numbers. Theorem 2 shows that this strategy (which was also used in [5] and always produces asymmetric codes) is optimal for values of θ and d corresponding to regions I and III, but is not so for regions II and IV. In fact, both [15] and [5] actually use a *sub-family* of the Golomb codes, for which the code parameter is a power of 2, making the encoding and decoding procedures extremely simple. This sub-family is further investigated in Section 4 in conjunction with adaptation strategies for the code parameters, in case θ and d are unknown a-priori. A different heuristic approach, based on encoding the absolute value with a Golomb code and appending a sign bit for nonzero values, was proposed in [16]. Theorem 2 shows that this heuristic (which always produces symmetric codes) is optimal only in Region II, and then only when ℓ is a power of two.

Method of the proof. In the proof of Theorem 2, we will borrow from [14] the concept of a *reduced source*. This concept is generalized in [11] and shown to be applicable to all finite entropy distributions of the integers, albeit in a non-constructive fashion. Here, for each of the regions defined for (θ, d), and each integer $m \geq 0$, we will define a finite mth order reduced source $\mathcal{R}_{L,m}$ as a multiset containing the first $2m - b$ probabilities in the ranking (32), where $b \in \{0, 1\}$ depends on the region, and a finite set of *super-symbol* probabilities, some of which represent infinite "tails" of the remaining integers. The index L also expresses region dependence, and it satisfies $L = 2\ell - 1$ for Region I and $L = 2\ell$ otherwise.

We will use Huffman's algorithm to construct an optimal prefix code for $\mathcal{R}_{L,m}$, and will then let m tend to infinity, thus obtaining a code for the integers. The code length assigned by our construction to an arbitrary integer x will be the one assigned by the optimal prefix code for $\mathcal{R}_{L,m}$, for the least m such that $2m - b \geq M(x)$. By the nature of the construction, this code length will remain unchanged for larger values of m. The argument validating the limiting step, and why it yields an optimal prefix code for the original infinite source, is given in [14] and it carries to our construction. The exact definition of the reduced sources used, and the way the Huffman construction on a reduced source proceeds, will vary according to the region the parameter pair (θ, d) falls into, thus leading to different code structures for the different regions. It turns out that the two-sided nature of the

distribution adds surprising complexity to the characterization as compared to the one-sided case (this holds even in the simpler case $d = 0$). This is evidenced by the variety of regions and codes in Theorem 2. In addition to the problems solved in the one-sided case, the characterization of the two-sided case includes finding various types of tails and reduced sources, the transitions between them, and the fine structure of the codes in regions II and IV.

We now offer some insight into how the various parameter regions (and hence the above mentioned complexity) arise. The functions $r_0(\ell, \theta, d)$ and $r_1(\ell, \theta, d)$ control the positive integer parameter L characterizing a basic property of Golomb-type codes: Starting from some codeword length Λ, the code contains exactly L codewords of length $\Lambda+i$ for all $i \geq 0$ (for the codes of Theorem 2, Λ is at most $\lambda + 2$, where λ is the minimal codeword length). The lines $r_1(\ell, \theta, d) = 0$ mark the transition from regions with $L = 2\ell - 1$ to regions with $L = 2\ell$, $\ell \geq 1$, while the lines $r_0(\ell, \theta, d) = 0$ mark the transition from $L = 2\ell$ to $L = 2\ell + 1$. The lines $r_2(\ell, \theta, d) = 0$ and $r_3(\ell, \theta, d) = 0$, in turn, determine how the optimal code construction handles "natural pairs" of symbols in regions with $L = 2\ell$. These are pairs of symbols that are close in probability, i.e., $\{x, -x\}$ for $d \leq \frac{1}{4}$ and $\{x-1, -x\}$ for $d > \frac{1}{4}$, where x is a positive integer. Focusing on the case $d \leq \frac{1}{4}$, and assuming x is sufficiently large, we observe that if the optimal code tree construction merges x and $-x$ (i.e., makes them sibling leaves), then by the constraints imposed by r_0 and r_1 in determining the value of L, the resulting probability $\pi = P_{(\theta,d)}(x) + P_{(\theta,d)}(-x)$ must fall in the proximity of the interval $[P_{(\theta,d)}(x - \ell), P_{(\theta,d)}(-(x - \ell))]$ on the real line. It turns out that the regions for $L = 2\ell$ are determined by whether π is to the *left* (Region II), *inside* (Region III'), or to the *right* (Region IV) of the interval. When π falls inside the interval, merging of x and $-x$ in the optimal tree construction would prevent the translated natural pair $\{x - \ell, -(x - \ell)\}$ from merging. Because of the self-similar character of the distribution, this condition applies to all x, and it results, in general, in a construction that does not merge natural pairs (e.g., the asymmetric codes of Region III'). On the other hand, when nothing stands in the way of a natural pair, they tend to merge, resulting in the symmetric codes of regions II and IV. A similar situation exists for $d > \frac{1}{4}$, but with a twist. There, again, the optimal construction will not merge natural pairs in Region III', and it will in regions II' and IV'. Nevertheless, regardless of whether they are merged or not, symbols in a natural pair end up with equal code lengths in the three sub-regions of Region III. This is due to the fact that the optimal code is a Golomb code of even parameter, and that every integer has a natural pair when $d > \frac{1}{4}$.

Terminology. We will often loosely refer to the "code length assigned to probability p" rather than the more cumbersome "code length assigned to the symbol whose probability is p." Also, we will characterize prefix codes in terms of their codeword length distributions, rather than the actual binary patterns they assign to the symbols. Thus, the actual code described will not always match the underlying tree induced by the iterative construction. For example, while the three sub-regions comprising Region III for a given ℓ (regions II', III', and IV') admit the same optimal prefix code, the constructions leading to the optimal length distribution and their underlying trees are quite different. A discussion of

the number of different coding trees that can be optimal for a given distribution, including some infinite alphabet cases, can be found in [26].

Additional definitions. The following definitions will aid in the proof of Theorem 2. For all integers j, we define

$$p_j \;=\; \begin{cases} C(\theta,d)\theta^{\lceil j/2\rceil - d}, & j \text{ odd}, \\ C(\theta,d)\theta^{(j/2)+d}, & j \text{ even}. \end{cases} \tag{33}$$

Notice that when $j \geq 0$, we have $p_j = P_{(\theta,d)}(\mu(j))$, i.e., p_j is the j-th probability in the ranking (32). Let $m \geq 0$, $L > 0$, and i be integers. We define a *single tail* $f_{L,i}(m)$ as follows:

$$f_{L,i}(m) = \sum_{j=0}^{\infty} p_{2m+i+jL}. \tag{34}$$

For all integers j, let

$$\eta_j = \; p_{2j-1} + p_{2j}. \tag{35}$$

Notice that, for $j \geq 1$, we have $\eta_j = \Pr(|x| = j)$. For even values of L, we define a *symmetric double tail* $F_{L,i}(m)$ as

$$F_{L,i}(m) = f_{L,2i-1}(m) + f_{L,2i}(m) = \sum_{j=0}^{\infty} \eta_{m+i+j\frac{L}{2}}. \tag{36}$$

The claims of the following lemma follow immediately from the definitions (2), (33), (34), and (36), and from straightforward geometric sum calculations.

Lemma 2 *Let $L > 0$, $m \geq 0$, and i be integers.*
(i) For any integer $k \geq i$, we have

$$f_{L,i}(m) \geq f_{L,k}(m)\,,$$

and, for even L,

$$F_{L,i}(m) \geq F_{L,k}(m)\,.$$

(ii) For integers k and $h \geq -m$, we have

$$f_{L,i+2k}(m+h) = \theta^{k+h} f_{L,i}(m),$$

and, for even L,

$$F_{L,i+k}(m+h) = \theta^{k+h} F_{L,i}(m). \tag{37}$$

(iii) We have

$$f_{L,i}(m) = C(\theta,d)\theta^{m+1}\hat{f}_{L,i} \tag{38}$$

and

$$F_{L,i}(m) = C(\theta,d)\theta^{m+1}\hat{F}_{L,i}, \tag{39}$$

where

$$\hat{f}_{L,i} = \begin{cases} \dfrac{\theta^{j-1+d}}{1-\theta^{2\ell-1}}(1+\theta^{\ell-2d}) & L=2\ell{-}1,\ i=2j, \\ \dfrac{\theta^{j-d}}{1-\theta^{2\ell-1}}(1+\theta^{\ell-1+2d}) & L=2\ell{-}1,\ i=2j{+}1, \\ \dfrac{\theta^{j-1+d}}{1-\theta^{\ell}} & L=2\ell,\ i=2j, \\ \dfrac{\theta^{j-d}}{1-\theta^{\ell}} & L=2\ell,\ i=2j{+}1. \end{cases} \tag{40}$$

and

$$\hat{F}_{L,i} = \frac{\theta^{i-1}(\theta^{-d}+\theta^{d})}{1-\theta^{\ell}}, \quad L=2\ell. \tag{41}$$

Let $\overline{\delta} = \frac{1}{2} - \delta$. We define the auxiliary functions $\overline{r}_i(\ell,\theta,d)$, $i=0,1$, by substituting $\overline{\delta}$ for δ in $r_i(\ell,\theta,d)$ as defined in (25)-(26). Since $\overline{\delta} \geq \frac{1}{4} \geq \delta$, the following inequalities hold:

$$\overline{r}_0(\ell,\theta,d) \geq r_0(\ell,\theta,d), \tag{42}$$

$$\overline{r}_1(\ell,\theta,d) \leq r_1(\ell,\theta,d), \tag{43}$$

with equalities holding only for $d = \frac{1}{4}$.

Proof of Theorem 2.

Region I. Let $L = 2\ell - 1$. We recall that Region I is characterized by the conditions $\theta_0(\ell,d) < \theta \leq \theta_1(\ell,d)$ or, equivalently $r_0(\ell,\theta,d) > 0$ and $r_1(\ell,\theta,d) \leq 0$. We refer to the latter two conditions as $\mathrm{C}_{1\mathrm{a}}$ and $\mathrm{C}_{1\mathrm{b}}$ respectively. By the inequalities (42) and (43), $\mathrm{C}_{1\mathrm{a}}$ and $\mathrm{C}_{1\mathrm{b}}$ imply the weaker conditions $\overline{r}_0(\ell,\theta,d) > 0$ and $\overline{r}_1(\ell,\theta,d) \leq 0$, which will be referred to, respectively, as $\overline{\mathrm{C}}_{1\mathrm{a}}$ and $\overline{\mathrm{C}}_{1\mathrm{b}}$.

We define an mth order reduced source $\mathcal{A}_{L,m}$, $m \geq 0$, as the multiset of probabilities

$$\mathcal{A}_{L,m} = \{p_0,\, p_1,\, \cdots,\, p_{2m-1},\, f_{L,0}(m),\, f_{L,1}(m),\, \cdots\, f_{L,L-1}(m)\}.$$

(When $m = 0$, the source includes tails only.)

We build an optimal prefix code for $\mathcal{A}_{L,m}$, $m > 0$. For real numbers a, b, c, d, we use the notation $\{a,b\} \leq \{c,d\}$ to denote $\max\{a,b\} \leq \min\{c,d\}$. This notation is extended in the natural way to relations of the form $a \leq \{c,d\}$ and $a \geq \{c,d\}$. We claim that the probabilities in $\mathcal{A}_{L,m}$ are ordered as follows:

$$\begin{aligned} \{p_{2m-1}, f_{L,L-1}(m)\} \;&\leq\; \{p_{2m-2}, f_{L,L-2}(m)\} \leq \{p_{2m-3}, f_{L,L-3}(m)\} \leq \cdots \\ \cdots \;&\leq\; \{p_{2m-L}, f_{L,0}(m)\} \leq p_{2m-L-1} \leq \cdots \leq p_0\,. \end{aligned} \tag{44}$$

To prove the claim, it suffices to prove the two leftmost inequalities, since the remaining inequalities are scaled versions of the first two. For $L = 1$ or $m = 1$, some of the symbols involved in the two leftmost inequalities are not part of $\mathcal{A}_{L,m}$, but the inequalities still

apply (with some negative-indexed p_j and $f_{L,i}$). To prove the leftmost inequality in (44), after applying (38), it suffices to show $\theta^{-1-d} < \hat{f}_{L,L-2}$ and $\hat{f}_{L,L-1} \le \theta^{-2+d}$. Using the expression for $\hat{f}_{L,L-2}$ from (40), the former inequality is equivalent to the condition

$$\theta^{2\ell-1}(1+\theta^{-1+2d}) + \theta^{\ell-1} - 1 > 0,$$

which in turn is equivalent to $\overline{\mathrm{C}}_{1\mathrm{a}}$ if $d \le \frac{1}{4}$, or to $\mathrm{C}_{1\mathrm{a}}$ otherwise. Similarly, $\hat{f}_{L,L-1} \le \theta^{-2+d}$ is equivalent to either $\mathrm{C}_{1\mathrm{b}}$ or $\overline{\mathrm{C}}_{1\mathrm{b}}$.

For the second leftmost inequality in the chain (44), it suffices to prove $\theta^{-2+d} < \hat{f}_{L,L-3}$ and $\hat{f}_{L,L-2} \le \theta^{-2-d}$. As before, the first inequality is equivalent to, or dominated by $\mathrm{C}_{1\mathrm{a}}$, while the second one is in the same situation with respect to $\mathrm{C}_{1\mathrm{b}}$.

It follows that the first merge of the Huffman algorithm on $\mathcal{A}_{L,m}$ produces the probability

$$p_{2m-1} + f_{L,L-1}(m) = p_{2m-1} + \sum_{j=0}^{\infty} p_{2m+L-1+jL} = \sum_{j=0}^{\infty} p_{2m-1+jL} = f_{L,1}(m-1).$$

Notice that $f_{L,1}(m-1) = f_{L,-1}(m)$, so after scaling by $\theta^{-\ell-1}$, the second leftmost inequality in (44) implies that the newly created probability satisfies $f_{L,1}(m-1) \ge \{p_{2m-L}, f_{L,0}(m)\}$. Hence, the next merge in the Huffman algorithm produces the probability

$$p_{2m-2} + f_{L,L-2}(m) = p_{2m-2} + \sum_{j=0}^{\infty} p_{2m+L-2+jL} = \sum_{j=0}^{\infty} p_{2m-2+jL} = f_{L,0}(m-1). \tag{45}$$

If $L = 1$, (45) still applies, since the second step uses the probability $f_{L,-1}(m) = f_{L,1}(m-1)$ produced in the first step.

Also, by (34), we have $f_{L,i}(m) = f_{L,i+2}(m-1)$, $0 \le i \le L-3$. Thus, after two steps (one round) of the Huffman algorithm, $\mathcal{A}_{L,m}$ is transformed into $\mathcal{A}_{L,m-1}$. The process continues for a total of m rounds, building up the tails $f_{L,i}$, until $\mathcal{A}_{L,0}$ is reached. This reduced source is given, in ascending probability order, by

$$\mathcal{A}_{L,0} = \{\, f_{L,L-1}(0),\, f_{L,L-2}(0),\, \cdots,\, f_{L,0}(0) \,\}.$$

We say that a finite source with probabilities $\pi_0 \le \pi_1 \le \cdots \le \pi_{N-1}$ is *quasi-uniform* if either $N \le 2$ or $\pi_0 + \pi_1 \ge \pi_{N-1}$. As noted in [14], an optimal prefix code for a quasi-uniform source of N probabilities admits at most two distinct codeword lengths, and it consists of $2^{\lceil \log N \rceil} - N$ codewords of length $\lfloor \log N \rfloor$, and $2N - 2^{\lceil \log N \rceil}$ codewords of length $\lceil \log N \rceil$, the shorter codewords being assigned to the larger probabilities.

We claim that $\mathcal{A}_{L,0}$ is quasi-uniform. For $L = 1$, there is nothing to prove. Otherwise, we need to show $f_{L,L-1}(0) + f_{L,L-2}(0) - f_{L,0}(0) \ge 0$. By Lemma 2(iii), after straightforward manipulations, the latter inequality is equivalent to

$$\theta^{2\ell-1}(\theta^{-1} + \theta^{-2d}) + \theta^{\ell-1} - 1 + \theta^{\ell-2d}(\theta^{-1} - 1) \ge 0. \tag{46}$$

By definition (25), since $\theta^{-1} > 1$ and $d \geq \delta$, the left-hand side of (46) is larger than $r_0(\ell, \theta, d)$, which is positive by $\mathrm{C}_{1\mathrm{a}}$. Therefore, $\mathcal{A}_{L,0}$ is quasi-uniform, and its optimal prefix code is constructed as described above, with $N = L$.

Tracing the way $\mathcal{A}_{L,0}$ "unfolds" into $\mathcal{A}_{L,m}$, it is apparent that the code length assigned to p_j, for m such that $2m - 1 \geq j$, is $\lfloor j/L \rfloor + 1 + \Lambda_{\mathcal{A}}(f_{L,j'})$, where $\Lambda_{\mathcal{A}}(f)$ is the code length assigned by the optimal prefix code for $\mathcal{A}_{L,0}$ to f, and $j' = j \mod L$. This is precisely the code length assigned by the Lth order Golomb code to j, as claimed by Theorem 2 for Region I.

Region II. Let $L = 2\ell$. In Region II we have $d \leq \frac{1}{4}$, $r_1(\ell, \theta, d) > 0$ and $r_2(\ell, \theta, d) \leq 0$. The latter two inequalities are referred to, respectively, as conditions $\mathrm{C}_{2\mathrm{a}}$ and $\mathrm{C}_{2\mathrm{b}}$.

We use a reduced source $\mathcal{S}_{L,m}$, defined by

$$\begin{aligned}\mathcal{S}_{L,m} = \ \{\, & p_0,\, p_1,\, \cdots,\, p_{2m},\, \eta_{m+1},\, \eta_{m+2},\, \cdots,\, \eta_{m+\ell-1}, \\ & F_{L,1}(m+\ell-1),\, F_{L,2}(m+\ell-1),\, \cdots,\, F_{L,\ell}(m+\ell-1)\,\}.\end{aligned}$$

(For $\ell = 1$, $\mathcal{S}_{L,m}$ contains no η_i's.)

We claim that the probabilities in $\mathcal{S}_{L,m}$ are ordered as follows (listed in ascending order, with inequality signs omitted):

$$\begin{aligned}& p_{2m},\, p_{2m-1},\, F_{L,\ell}(m+\ell-1),\, \eta_{m+\ell-1},\, p_{2m-2},\, p_{2m-3},\, F_{L,\ell-1}(m+\ell-1),\, \cdots \\ & \qquad \cdots,\, \eta_{m+1},\, p_{2(m-\ell+1)},\, p_{2(m-\ell+1)-1},\, F_{L,1}(m+\ell-1),\, p_{2(m-\ell)},\, \cdots,\, p_1,\, p_0\ .\end{aligned}$$

When $m < \ell$, the sequence of original probabilities p_i stops at p_0 "in the middle" of the chain (i.e., just before one of the $F_{L,j}$'s), but the order relations between all remaining symbols are still claimed to hold.

To prove the claim, it suffices to show that (a) $p_{2m-1} < F_{L,\ell}(m+\ell-1)$, (b) $F_{L,\ell}(m+\ell-1) \leq \eta_{m+\ell-1}$, and (c) $\eta_{m+\ell-1} \leq p_{2m-2}$. The rest of the chain follows by virtue of scaling. We use the expression for $F_{L,\ell}$ from (39) and (41). Inequality (a) is equivalent to $\mathrm{C}_{2\mathrm{a}}$. By definition (35), and after eliminating common factors, inequality (b) is equivalent to

$$2\theta^{\ell} - 1 \leq 0. \tag{47}$$

Clearly, (47) is implied by

$$\theta^{\ell}(1 + \theta^{-2\delta}) - 1 \leq 0,$$

the latter inequality being $\mathrm{C}_{2\mathrm{b}}$. Inequality (c), in turn, is also equivalent to $\mathrm{C}_{2\mathrm{b}}$, as $d = \delta$.

The first step of the Huffman algorithm on $\mathcal{S}_{L,m}$ merges p_{2m} and p_{2m-1}, creating η_m. By inequality (b) above, suitably scaled by $\theta^{\ell-1}$, we have $\eta_m \geq F_{L,\ell}(m) = F_{L,1}(m+\ell-1)$. Hence, the second Huffman step joins $F_{L,\ell}(m+\ell-1)$ with $\eta_{m+\ell-1}$. Recalling the definition of $F_{L,\ell}$ in (36), we obtain

$$F_{L,\ell}(m+\ell-1) + \eta_{m+\ell-1} = \sum_{j=0}^{\infty} \eta_{m+2\ell-1+j\ell} + \eta_{m+\ell-1} = \sum_{j=0}^{\infty} \eta_{m+\ell-1+j\ell} = F_{L,1}(m+\ell-2).$$

When $\ell = 1$, the probability η_m created in the first step is used in the second one. Notice that, by (36), we can also write $F_{L,j}(m+\ell-1) = F_{L,j+1}(m+\ell-2)$ for $1 \le j \le \ell-1$. Hence, after two steps (one round) of the Huffman algorithm, $\mathcal{S}_{L,m}$ is transformed into $\mathcal{S}_{L,m-1}$. After m rounds, we obtain $\mathcal{S}_{L,0}$, given in ascending probability order by

$$\mathcal{S}_{L,0} \;=\; \{\, p_0,\, F_{L,\ell}(\ell-1),\, \eta_{\ell-1},\, F_{L,\ell-1}(\ell-1),\, \cdots,\, F_{L,2}(\ell-1),\, \eta_1,\, F_{L,1}(\ell-1) \,\}. \tag{48}$$

When $\ell = 1$, (48) translates to $\mathcal{S}_{2,0} = \{\, p_0,\, F_{2,1}(0) \,\}$.
We claim that $\mathcal{S}_{L,0}$ is quasi-uniform, i.e., $p_0 + F_{L,\ell}(\ell-1) - F_{L,1}(\ell-1) \ge 0$ whenever $\ell > 1$. By (37), (39) and (41), the required condition is equivalent to

$$\theta^d + \frac{(\theta^{\ell-1}-1)\theta^\ell(\theta^{-d}+\theta^d)}{1-\theta^\ell} \;\ge\; 0.$$

Multiplying by $(1-\theta^\ell)$, and rearranging terms, the above inequality is equivalent to

$$\theta^d(1-2\theta^\ell) + \theta^{\ell-d} r_3(\ell{-}1,\theta,d) \;\ge\; 0.$$

By (47), and since $r_3(\ell-1,\theta,d) > 0$, as in Region II we have $\theta > \theta_1(\ell,d) > \theta_0(\ell,d) \ge \theta_3(\ell-1,\theta,d)$, it follows that $\mathcal{S}_{L,0}$ is quasi-uniform.
We now construct an optimal prefix code for $\mathcal{S}_{L,0}$, and show how it translates into an optimal code for $\mathcal{S}_{L,m}$ and, thus, for the integers under the distribution (2). Let r be the integer satisfying $2^{r-1} \le \ell < 2^r$, and $s = 2^r - \ell$. Assume first that $s < \ell$. Since $\mathcal{S}_{L,0}$ has 2ℓ probabilities, an optimal prefix code for it assigns code length r to the $2s$ largest probabilities, namely,

$$\eta_s,\, F_{L,s}(\ell-1),\, \cdots,\, \eta_2,\, F_{L,2}(\ell-1),\, \eta_1,\, F_{L,1}(\ell-1)\,. \tag{49}$$

The code assigns length $r+1$ to the $2\ell - 2s$ remaining probabilities, namely,

$$p_0,\, F_{L,\ell}(\ell-1),\, \eta_{\ell-1},\, \cdots,\, \eta_{s+1},\, F_{L,s+1}(\ell-1)\,. \tag{50}$$

Notice that in the iterative construction of the Huffman code for $\mathcal{S}_{L,m}$, $m \ge 1$, pairs p_{2j-1}, p_{2j}, $j \ge 1$, which correspond to integers of opposite signs, merge to form η_j. Thus, it suffices to characterize the code length assigned by the construction to the η_j, $j > 0$, and p_0. Similarly to Region I, tracing this time the way $\mathcal{S}_{L,0}$ "unfolds" into $\mathcal{S}_{L,m}$, we observe that $\eta_{i+j\ell}$, $0 \le i < \ell$, $1 \le j \le m$, is a leaf in a subtree rooted at $F_{L,i+1}(\ell{-}1)$, j levels down from the root of the subtree. Thus, the code length assigned to $\eta_{i+j\ell}$ when m is sufficiently large is $j + \Lambda_{\mathcal{S}}(F_{L,i+1}(\ell-1))$, $j \ge 1$. Here, $\Lambda_{\mathcal{S}}(\cdot)$ is the code length assignment of the Huffman code for $\mathcal{S}_{L,0}$, i.e., $\Lambda_{\mathcal{S}}(\pi) = r$ for probabilities π in the list (49), and $\Lambda_{\mathcal{S}}(\pi) = r+1$ for probabilities π in the list (50). Code lengths for p_0 and η_i, $0 < i < \ell$, are assigned directly by $\Lambda_{\mathcal{S}}$. The foregoing discussion is summarized in the following lemma.

Lemma 3 *If $s < \ell$, the code lengths assigned by our construction for Region II are as follows:*

$$\begin{aligned}
\underline{\textit{Probability}} &: \underline{\textit{Code Length}} \\
p_0 &: r+1, \\
\eta_{j\ell} &: r+j, \;\; j \geq 1, \\
\eta_{k+j\ell} &: r+j, \;\; 1 \leq k \leq s-1, \;\; j \geq 0, \\
\eta_s &: r, \\
\eta_{s+j\ell} &: r+1+j, \;\; j \geq 1, \\
\eta_{h+j\ell} &: r+1+j, \;\; s+1 \leq h \leq \ell-1, \;\; j \geq 0.
\end{aligned}$$

For ease of comparison, the following lemma explicitly lists the code length assignment of an ℓth order Golomb code (in a purposely redundant manner), with r and s defined as in Theorem 2.

Lemma 4 *The code lengths assigned by an ℓth order Golomb code on the nonnegative integers are as follows:*

$$\begin{aligned}
\underline{\textit{Symbol}} &: \underline{\textit{Code Length}} \\
0 &: r, \\
j\ell &: r+j, \;\; j \geq 1, \\
k+j\ell &: r+j, \;\; 1 \leq k \leq s-1, \;\; j \geq 0, \\
s &: r+1, \\
s+j\ell &: r+1+j, \;\; j \geq 1, \\
h+j\ell &: r+1+j, \;\; s{+}1 \leq h \leq \ell-1, \;\; j \geq 0.
\end{aligned}$$

Comparing the length assignment in Lemma 3 with that of Lemma 4, we observe that our construction assigns to p_0 and η_i, $i > 0$, the same code length that an ℓth order Golomb code assigns to $i = 0$ and $i > 0$, respectively, except that the code lengths for $i = 0$ and $i = s$ have been exchanged. This is due to the fact that η_s and p_0 have "swapped places" in the lists (49) and (50) with respect to their "natural" places in a Golomb code. When $s = \ell$, i.e., ℓ is a power of two, all the probabilities in $\mathcal{S}_{L,0}$ are assigned the same code length r, and the swapping has no effect, even though p_0 is still the lowest probability in $\mathcal{S}_{L,0}$. To complete the proof of Theorem 2 for Region II, the code for the original probabilities p_{2j-1}, p_{2j}, $j \geq 1$, is obtained by appending a sign bit to the code for the corresponding η_j.

Region IV. Let $L = 2\ell$. Region IV is characterized by $d \leq \frac{1}{4}$, and the conditions $r_3(\ell, \theta, d) > 0$ and $r_0(\ell+1, \theta, d) \leq 0$, referred to, respectively, as conditions $\mathrm{C}_{4\mathrm{a}}$ and $\mathrm{C}_{4\mathrm{b}}$. The construction in Region IV follows along similar lines as that of Region II, and it will only be outlined here.

We use a reduced source $\mathcal{T}_{L,m}$, defined by

$$\begin{aligned}
\mathcal{T}_{L,m} = \; \{ & \, p_0, \, p_1, \, \cdots, \, p_{2m}, \, \eta_{m+1}, \, \eta_{m+2}, \, \cdots, \, \eta_{m+\ell}, \\
& F_{L,1}(m+\ell), \, F_{L,2}(m+\ell), \, \cdots, \, F_{L,\ell}(m+\ell) \, \}.
\end{aligned}$$

It follows from C_{4a} and C_{4b} that the probabilities in $\mathcal{T}_{L,m}$ are sorted in ascending order as follows:

$$p_{2m},\, p_{2m-1},\, \eta_{m+\ell},\, F_{L,\ell}(m+\ell),\, p_{2m-2},\, p_{2m-3},\, \eta_{m+\ell-1},\, F_{L,\ell-1}(m+\ell),\, \cdots$$
$$\cdots,\, p_{2(m-\ell+1)},\, p_{2(m-\ell+1)-1},\, \eta_{m+1},\, F_{L,1}(m+\ell),\, p_{2(m-\ell)},\, \cdots,\, p_2,\, p_1,\, p_0\,.$$

Similarly to Region II, a round consisting of two steps of the Huffman algorithm on $\mathcal{T}_{L,m}$ leads to $\mathcal{T}_{L,m-1}$, with the pair p_{2m}, p_{2m-1} merging to form η_m, and $\eta_{m+\ell}$ merging with $F_{L,\ell}(m+\ell)$ to form $F_{L,1}(m+\ell-1)$. After m rounds, we obtain the reduced source $\mathcal{T}_{L,0}$, which consists of the following $2\ell+1$ probabilities, in ascending order:

$$p_0,\, \eta_\ell,\, F_{L,\ell}(\ell),\, \eta_{\ell-1},\, F_{L,\ell-1}(\ell),\, \cdots,\, \eta_1,\, F_{L,1}(\ell)\,.$$

It follows from C_{4a} and C_{4b} that $\mathcal{T}_{L,0}$ is quasi-uniform. Now, if r is the integer satisfying $2^{r-1} \le \ell < 2^r$, then r also satisfies $2^r < 2\ell+1 < 2^{r+1}$. Thus, an optimal prefix code for $\mathcal{T}_{L,0}$ contains $2^{r+1} - 2\ell - 1 = 2s-1$ codewords of length r, and $2\ell - 2s + 2$ codewords of length $r+1$. The list of probabilities corresponding to codewords of length r is given by

$$F_{L,s}(\ell),\, \eta_{s-1},\, F_{L,s-1}(\ell),\, \cdots,\, \eta_2,\, F_{L,2}(\ell),\, \eta_1,\, F_{L,1}(\ell),$$

while the list of probabilities corresponding to codewords of length $r+1$ is given by

$$p_0,\, \eta_\ell,\, F_{L,\ell}(\ell),\, \cdots,\, \eta_{s+1},\, F_{L,s+1}(\ell),\, \eta_s\,.$$

(If $\ell = 1$, the first list consists just of $F_{L,1}(1)$, while the second list consists of p_0 and η_1.) From this code length distribution for $\mathcal{T}_{L,0}$ we derive a code length distribution for p_0 and η_j, $j \ge 0$, described in the following lemma.

Lemma 5 *The code lengths assigned by our construction for Region IV are as follows:*

Probability	:	*Code Length*
p_0	:	$r+1$,
$\eta_{j\ell}$	:	$r+j,\ \ j \ge 1$,
$\eta_{k+j\ell}$	:	$r+j,\ \ i \le k \le s-1,\ \ j \ge 0$,
η_s	:	$r+1$,
$\eta_{s+j\ell}$	:	$r+j,\ \ j \ge 1$,
$\eta_{h+j\ell}$	:	$r+1+j,\ \ s+1 \le h \le \ell-1,\ \ j \ge 0$.

We observe that our construction assigns to p_0 and η_i, $i > 0$, the same code length that an ℓth order Golomb code assigns to $i = 0$ and $i > 0$, respectively, except that p_0 is assigned a code one bit longer, and all $\eta_{s+j\ell}$, $j \ge 1$, are assigned a code one bit shorter. Since length transitions in an ℓth order Golomb code occur precisely at integers congruent to s modulo ℓ (in that the Golomb code length for $s + j\ell$ is one more than the code length for

$s-1+j\ell$), this shortening is equivalent to assigning to η_i, $i > s$, the code length that a Golomb code would assign to $i-1$. The codewords for p_0 (that needs to grow by one bit relative to $G_\ell(0)$ of length r) and η_s (that retains the length $r+1$ of $G_\ell(s)$ but just lost that codeword to η_{s+1}) are obtained by appending a bit to $G_\ell(0)$. As in the case of Region II, the proof is completed by observing that the η_j, $j > 0$, split into original probabilities, amounting to the appendage of a sign bit to the code for η_j.

Regions II' and IV'. These regions satisfy the same conditions as regions II and IV (resp.), but are defined for $d > \frac{1}{4}$, i.e., we have $\delta = \frac{1}{2} - d$. Consider the distribution $P_{(\theta,\delta)}(\cdot)$, and let $\overline{p}_j = P_{(\theta,\delta)}(\mu(j))$ for $j \geq 0$, extending in analogy with p_j for $j < 0$. We can write

$$P_{(\theta,d)}(x) = C(\theta,d)\theta^{|x+d|} = \begin{cases} C(\theta,d)\theta^{-\frac{1}{2}}\theta^{x+1-\delta} = \gamma P_{(\theta,\delta)}(-x-1) & x \geq 0, \\ C(\theta,d)\theta^{-\frac{1}{2}}\theta^{-x+\delta} = \gamma P_{(\theta,\delta)}(-x) & x < 0, \end{cases}$$

for a constant γ. Therefore, we have $p_j = \gamma \overline{p}_{j+1}$ for all integers j, which means that once the probabilities are ordered, and ignoring scaling factors, the distribution $P_{(\theta,d)}(\cdot)$, $d > \frac{1}{4}$, "looks" exactly like the distribution $P_{(\theta,\delta)}(\cdot)$, with p_0 removed. Noting that in the constructions for reduced sources in regions II and IV, p_0 was involved only in the final stage, when an optimal code for the "core" sources $\mathcal{S}_{L,0}$ or $\mathcal{T}_{L,0}$ was determined, we conclude that the same constructions can be used for regions II' and IV', except for that final stage. We now formalize this idea.

Let $L = 2\ell$. For all integers j, let

$$\nu_j = p_{2j-2} + p_{2j-1},$$

and, for all integers i define the *asymmetric double tail*

$$K_{L,i}(n) = \sum_{j=0}^{\infty} \nu_{n+i+j\ell}.$$

It follows from the discussion above that we can write

$$\nu_j = \gamma \overline{\eta}_j, \quad j \in \mathbf{Z},$$

and

$$K_{L,i}(n) = \gamma \overline{F}_{L,i}(n), \quad i \in \mathbf{Z},$$

where $\overline{\eta}_j$ and $\overline{F}_{L,i}$ are analogous to η_j and $F_{L,i}$, respectively, but defined for the distribution $P_{(\theta,\delta)}$. Assume (θ, d) falls in Region II', and define the reduced source

$$\begin{aligned} \overline{\mathcal{S}}_{L,m} = \ & \{\, p_0, p_1, \cdots, p_{2m-1}, \nu_{m+1}, \nu_{m+2}, \cdots, \nu_{m+\ell-1}, \\ & K_{L,1}(m+\ell-1), K_{L,2}(m+\ell-1), \cdots, K_{L,\ell}(m+\ell-1) \,\}. \end{aligned}$$

This source is equivalent to a scaled version of $\mathcal{S}_{L,m} - \{p_0\}$, but for the distribution $P_{(\theta,\delta)}$. Therefore, the iteration leading from $\mathcal{S}_{L,m}$ to $\mathcal{S}_{L,0}$ applies, and after $2m$ steps of the Huffman algorithm, we have

$$\overline{\mathcal{S}}_{L,0} = \{\, K_{L,\ell}(\ell-1),\ \nu_{\ell-1},\ K_{L,\ell-1}(\ell-1),\ \nu_{\ell-2},\ \cdots,\ K_{L,2}(\ell-1),\ \nu_1,\ K_{L,1}(\ell-1) \,\},$$

where the probabilities are listed in ascending order. The construction now departs from that of Region II. We observe that, without an analog of p_0 in the way, we can carry out $\ell - 1$ additional merges, which, by the definition of $K_{L,i}$, yield

$$K_{L,i}(\ell - 1) + \nu_{i-1} = K_{L,i-1}(0), \;\; 2 \leq i \leq \ell \, .$$

In addition, we have $K_{L,1}(\ell - 1) = K_{L,\ell}(0)$. Thus, we obtain a further reduced source

$$\overline{\mathcal{S}}_{L,-1} = \{ \, K_{L,\ell}(0), \, K_{L,\ell-1}(0), \, \cdots \, , \, K_{L,1}(0) \, \}.$$

It can be readily verified that $\overline{\mathcal{S}}_{L,-1}$ is quasi-uniform. Thus, the construction yields an ℓth order Golomb code for $\nu_1, \nu_2, \nu_3, \cdots$, where ν_j, $j \geq 1$, gets assigned a code length corresponding to that of $G_\ell(j-1)$. This translates into an Lth order Golomb code for $p_0, p_1, p_2, \cdots$, as claimed by Theorem 2 for Region II' (as part of Region III).

Similar considerations apply to Region IV', where we define a reduced source $\overline{\mathcal{T}}_{L,m}$ analogous to $\mathcal{T}_{L,m} - \{p_0\}$. This leads to a core source

$$\overline{\mathcal{T}}_{L,0} = \{ \, \nu_\ell, \, K_{L,\ell}(\ell), \, \nu_{\ell-1}, \, K_{L,\ell-1}(\ell), \, \cdots \, , \, \nu_2, \, K_{L,2}(\ell), \, \nu_1, \, K_{L,1}(\ell) \, \},$$

which can be further reduced to obtain

$$\overline{\mathcal{T}}_{L,-1} = \{ \, K_{L,\ell}(0), \, K_{L,\ell-1}(0), \, \cdots \, , \, K_{L,1}(0) \, \}.$$

This source, again, leads to an Lth order Golomb code for the original sorted probabilities.

Region III'. Let $L = 2\ell$. Region III' is characterized by the conditions $r_2(\ell, \theta, d) > 0$ and $r_3(\ell, \theta, d) \leq 0$, referred to, respectively, as conditions $\mathrm{C}_{3\mathrm{a}}$ and $\mathrm{C}_{3\mathrm{b}}$.

Here, we define a reduced source $\mathcal{U}_{L,m}$ given by

$$\mathcal{U}_{L,m} = \{ \, p_0, \, p_1, \, \cdots \, , \, p_{2m-1}, \, f_{L,0}(m), \, f_{L,1}(m), \, \cdots \, , \, f_{L,L-1}(m) \, \}.$$

This reduced source appears to be formally identical to $\mathcal{A}_{L,m}$ used in Region I. However, inspecting (38) and (40), we note that the expressions for $f_{L,i}(m)$ when L is even are quite different from those applying when L is odd. Nevertheless, after appropriate reinterpretation of $f_{L,i}$, the order relations claimed in (44) hold for Region III', being implied this time by $\mathrm{C}_{3\mathrm{a}}$ and $\mathrm{C}_{3\mathrm{b}}$ rather than $\mathrm{C}_{1\mathrm{a}}$ and $\mathrm{C}_{1\mathrm{b}}$. Similarly, the evolution from $\mathcal{U}_{L,m}$ to $\mathcal{U}_{L,0}$ by way of the Huffman procedure is formally identical to that from $\mathcal{A}_{L,m}$ to $\mathcal{A}_{L,0}$. Finally, the quasi-uniformity of $\mathcal{U}_{L,0}$ follows from $\mathrm{C}_{3\mathrm{a}}$, and thus, the construction in Region III' yields a code whose length assignment for p_i, $i \geq 0$, is identical to that of an Lth order Golomb code for $i \geq 0$.

Optimality. The optimality of the codes prescribed by Theorem 2 follows from the same argument presented in [14], applied separately to each region. The main formal requirement is the convergence of the average code length, which is established in Lemma 6 below. This completes the proof of Theorem 2. □

We refer to the prefix codes defined in Theorem 2 for regions I, II, III, and IV as codes of *Type I, II, III, and IV*, respectively. The expected code lengths for these codes when

applied to the TSG distribution (2) are obtained from their definitions in Theorem 2 by applying straightforward geometric sums, and derived sums of the general form $\sum_i ix^i$. The resulting code lengths are given in the following lemma. Notice that the expected code lengths apply to all allowable parameter values (θ, d), and not just to the region for which a code is optimal.

Lemma 6 *Let ℓ be an arbitrary positive integer, and let $\overline{\lambda}_{X,\ell}(\theta, d)$ denote the average code length for a code of type X (X=I,II,III,IV), for the given value of ℓ, when applied to a TSG source with parameters (θ, d). Let r and s be defined as in Theorem 2, and let $s' = s \bmod 2^{r-1}$. Then, we have*

$$\overline{\lambda}_{I,\ell}(\theta, d) = 1 + \lfloor \log(2\ell - 1) \rfloor + \frac{\theta^{s'+1}}{1 - \theta^{2\ell-1}}(1 - P_{(\theta,d)}(0) + \theta^{\ell}) \,.$$

$$\overline{\lambda}_{II,\ell}(\theta, d) = 1 + \lceil \log \ell \rceil + (1 - P_{(\theta,d)}(0))\theta^{s'} \left(1 + \frac{\theta^{\ell-1}}{1 - \theta^{\ell}}\right) .$$

$$\overline{\lambda}_{III,\ell}(\theta, d) = 1 + \lfloor \log(2\ell) \rfloor + \frac{\theta^{s}}{1 - \theta^{\ell}} \,.$$

$$\overline{\lambda}_{IV,\ell}(\theta, d) = 2 + \lfloor \log \ell \rfloor + (1 - P_{(\theta,d)}(0))\theta^{s-1} \left(1 + \frac{\theta^{\ell+1}}{1 - \theta^{\ell}}\right) .$$

4 Low complexity adaptive codes

In this section, we consider the case where the values of the parameters θ and d are unknown a-priori, in the framework of symbol-by-symbol coding. Even though in general adaptive strategies are easier to implement with arithmetic codes, the structured family of codes introduced in Section 3 provides a reasonable alternative for low complexity adaptive coding of TSG models: Based on x^t, select a code in the family sequentially, and use this code to encode x_{t+1}. Unlike in Section 2, here the set of available coding strategies for each symbol is discrete, and the 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, namely the Huffman code for the unknown parameter values.

Let $\mathcal{C} = \{C^{(1)}, C^{(2)}, \cdots, C^{(j)}, \cdots\}$ denote a countable family of codes, and let $C_t \in \mathcal{C}$ denote the code chosen by an on-line algorithm to encode x_{t+1}. For a fixed code $C^* \in \mathcal{C}$, let $\overline{\Delta\lambda}(j)$ denote the expected code length difference between $C^{(j)}$ and C^*. Then, it can be readily verified that the expectation of the code length difference $\Delta\Lambda(x^n)$ over the entire sequence x^n is given by

$$E[\Delta\Lambda(X^n)] = \sum_{j=1}^{\infty} \overline{\Delta\lambda}(j) \sum_{t=1}^{n} \Pr\{C_t = C^{(j)}\} \,. \tag{51}$$

Equation (51) motivates the following on-line strategy for adaptive coding of a TSG distribution: Given an estimate of θ and d based on the sufficient statistics S_t and N_t (as defined in equations (11) and (13), Section 2), use the corresponding optimal prefix code prescribed by Theorem 2 to encode x_{t+1}. 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 $\overline{\Delta\lambda}(j)$ 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 choosing the code that would have performed best on x^t. The latter approach was used in [27] to encode OSG distributions. Yet, 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, the sub-family of codes used in practical schemes such as [15] and [5] 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 (the simplicity of the power-of-2 case was already noted in [13]). The code length is $r + 1 + \lfloor z/2^r \rfloor$ bits. Furthermore, following [15] and [5], we will consider only the asymmetric codes of types I and III. The codeword assigned to an integer x is denoted $\Gamma_r(x) = G_{2^r}(M(x))$. Given the nature of the mapping (31), which privileges 0 over -1, 1 over -2, etc., we consider an additional code Γ'_0, which corresponds to $r = 0$ but with the mapping (31) modified so that negative and nonnegative values are interleaved in the sequence $-1, 0, -2, 1, ...$ This code accounts for values of d larger than $\frac{1}{2}$, for which -1 is more frequent than 0, and small θ. For $r > 0$ (larger values of θ) this re-mapping (which actually realizes the transformation $x \to -(x+1)$ mentioned in the beginning of Section 3) is irrelevant, as the values in the pairs $(0, -1)$, $(1, -2)$, ..., are given the same code length. Thus, this final section presents a further practical compromise, according to which we consider adaptive coding for the above reduced family of codes. This reduced family represents a specific optimality-complexity trade-off, and similar derivations are possible with other sub-families. For example, the symmetric codes of Type II are included in [28], leading to a more complex analysis for region determination. The specific trade-off in this section is motivated by the success of the lossless image compression algorithm LOCO-I [5], in which only the asymmetric sub-family $\mathcal{C} = \{\Gamma_r\} \bigcup \Gamma'_0$ is employed.[5] Based on a sequential estimate of θ and ρ (rather than d), LOCO-I chooses a code from the above sub-family. Lemma 7 below presents optimal decision regions to choose among these codes for given values of θ and ρ under the modified TSG distribution (6).

Lemma 7 *Let $S \stackrel{\Delta}{=} \theta/(1-\theta)$ and $\phi \stackrel{\Delta}{=} (\sqrt{5}+1)/2$. Given S and ρ, the following decision rules minimize the expected codeword length over the sub-family of codes $\mathcal{C}$.*

a. If $S \leq \phi$, compare S, ρ, and $1-\rho$. If S is largest, choose code Γ_1. Otherwise, if ρ is

[5] 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 [10] 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.

largest, choose Γ_0. *Otherwise, choose* Γ'_0.

b. If $S > \phi$, *choose code* Γ_{r+1}, $r \geq 1$ *provided that*

$$\frac{1}{\phi^{(2^{-r+1})} - 1} < S \leq \frac{1}{\phi^{(2^{-r})} - 1}\,. \tag{52}$$

Proof. Let $\overline{\lambda}_r(S,\rho)$ denote the average code length for code Γ_r, $r \geq 0$, and let $\overline{\lambda'}_0(S,\rho)$ denote the average code length for code Γ'_0. 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), $E_{(\theta,\rho)}[M(x)] = 2S + 1 - \rho$. Thus,

$$\overline{\lambda}_0(S,\rho) = \overline{\lambda'}_0(S, 1-\rho) = 2 + 2S - \rho\,. \tag{53}$$

The other codes in the sub-family under consideration ($r > 0$) are of Type III, so we apply Lemma 6 with $\ell = 2^{r-1}$ to obtain

$$\overline{\lambda}_r(S,\rho) = \overline{\lambda}_{\mathrm{III},\ell}(\theta, d) = r + 1 + \frac{\theta^\ell}{1-\theta^\ell}\,.$$

Thus, the code selection for $r > 0$ is done according to the sign of

$$\overline{\lambda}_{r+1}(S,\rho) - \overline{\lambda}_r(S,\rho) = 1 - \frac{\theta^\ell}{1-\theta^{2\ell}}\,. \tag{54}$$

By (54), the maximum value of θ for which $\Gamma_r\,, r > 0$, can be optimal, is such that $\theta^{2^{r-1}} = 1 - \theta^{2^r}$, namely,

$$\theta = (\phi - 1)^{(2^{-r+1})}\,. \tag{55}$$

Thus, the maximum value of S for which $\Gamma_r\,, r > 0$, can be optimal, is

$$S = \frac{t}{\phi^{(2^{-r+1})} - 1}\,. \tag{56}$$

The decision rule of Lemma 7 follows from equations (53) and (56). □

Lemma 7 extends results in [15] and [27]. The golden ratio is mentioned in connection with Golomb codes in [26], and also in [27].

Theorem 3 below states that, in a probabilistic setting, an on-line strategy based on ML estimation and the decision regions of Lemma 7 performs essentially as well as the best code in the sub-family $\mathcal{C}$. The code selection for x_{t+1} is based on the sufficient statistics S_t and N_t.

Theorem 3 *Encode* x_{t+1} *using the code prescribed by Lemma 7 after substituting* S_t/t *for* S *and* N_t/t *for* $1-\rho$, $0 \leq t < n$, *and let* $\Lambda(x^n)$ *denote the code length resulting from applying this adaptation strategy to the sequence* x^n. *Let* Λ^* *denote the minimum expected codeword length over codes in* $\mathcal{C}$ *for the (unknown) parameters* θ *and* ρ. *Then,*

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

Discussion.

Relation to prior work. A result analogous to Theorem 3 is proved in [27] 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 $O(1/\sqrt{n})$. In fact, this alternative approach was also analyzed (for general loss functions) in the individual sequence setting in [29], where it was shown that by introducing randomization it is possible to match the performance on *every sequence* up to $O(1/\sqrt{n})$. It should be pointed out that, in our case, these two plug-in strategies indeed differ. For example, for the sequence $x^6 = 022222$, $S_t/t = 5/3 > \phi$, so the approach based on ML estimation encodes x_7 with the code Γ_2. On the other hand, direct inspection reveals that the best code for x^6 is Γ_1.

Low complexity approximation. The decision region boundaries (56) admit a low complexity approximation, for which it is useful to define the functions $S(r)$ and $\gamma(r)$, $r > 0$, by

$$S(r) \stackrel{\Delta}{=} \frac{1}{\phi^{(2^{-r+1})} - 1} \stackrel{\Delta}{=} \frac{2^{r-1}}{\ln \phi} - \frac{1}{2} + \gamma(r)\,. \tag{57}$$

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 $1/\ln\phi \approx 2.078$, (57) 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 (52), a good approximation to the decision rule of Theorem 3 for encoding x_{t+1} is:

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

a. If $S'_t \le 2t$, compare S_t, N_t, and $t - N_t$. If S_t is larger, choose code Γ_1. Otherwise, if $t - N_t$ is larger, choose Γ_0. Otherwise, choose Γ'_0.

b. If $S'_t > 2t$, choose code Γ_{r+1}, $r \ge 1$ provided that

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

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

Proof of Theorem 3.

It suffices to prove that the right-hand side of (51) is upper-bounded by a constant as $n \to \infty$ for the family $\mathcal{C}$ of codes under consideration. Let $\overline{r}^*$ denote the *largest* value of r for which Γ_r is optimum or, in case no such r exists (hence Γ'_0 is optimum), let $\overline{r}^* = 0$. Similarly, let $\underline{r}^*$ denote the *smallest* value of r for which Γ_r is optimum or, in case Γ'_0 is optimum, let $\underline{r}^* = 0$. Notice that, in general, $\overline{r}^* = \underline{r}^*$, except when S is on one of the code selection boundaries defined in Lemma 7. We divide the outer sum in the right-hand side of (51) in two parts, one corresponding to codes Γ_r such that $r > \overline{r}^*$, which yields a sum

Δ_1, and one for the other codes in $\mathcal{C}$ which are not optimum (codes Γ_r such that $r < \underline{r}^*$, and, if not optimum, Γ'_0), which yields a sum Δ_2. Thus, (51) takes the form

$$E_{(\theta,\rho)}[\Lambda(x^n)] = n\Lambda^* + \Delta_1 + \Delta_2 \,. \tag{58}$$

Clearly, if $r > \overline{r}^*$ then the code length difference between Γ_r and $\Gamma_{\overline{r}^*}$ can be at most $r - \overline{r}^*$ bits per encoding, due to a longer binary part using Γ_r ($r > \overline{r}^*$ cannot increase the unary part). Thus,

$$\Delta_1 \leq \sum_{r=\overline{r}^*+1}^{\infty} (r - \overline{r}^*) \sum_{t=1}^{n} \Pr\{\mathcal{C}_t = \Gamma_r\} = \sum_{r=\overline{r}^*}^{\infty} \sum_{t=1}^{n} \Pr\{r(t) > r\} \tag{59}$$

where $\Gamma_{r(t)} \stackrel{\Delta}{=} \mathcal{C}_t$. Now, with $S(r)$ defined as in Equation (57), $r > 0$, and $S(0) \stackrel{\Delta}{=} \max\{\rho, 1-\rho\}$, the proposed on-line selection rule is such that

$$\Pr\{r(t) > r\} = \Pr\{S_t > tS(r)\} \,. \tag{60}$$

Define $\theta(r) \stackrel{\Delta}{=} S(r)/(1 + S(r))$, which is an increasing function of r. By Lemma 7 and the definition of $\overline{r}^*$, 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

$$\Pr\{S_t > tS(r)\} \leq 2^{-tD(\theta(r)||\theta)} \tag{61}$$

where $D(\theta(r)||\theta)$ denotes the informational divergence between OSG sources with parameters $\theta(r)$ and θ, respectively, which is positive for $r \geq \overline{r}^*$. It follows from (59), (60), and (61), that

$$\Delta_1 \leq \sum_{r=\overline{r}^*}^{\infty} \frac{1}{2^{D(\theta(r)||\theta)} - 1} \stackrel{\Delta}{=} \sum_{r=\overline{r}^*}^{\infty} a_r \,. \tag{62}$$

To prove the convergence of the series in the right-hand side of (62) we upper-bound the ratios a_{r+1}/a_r. Since $D(\theta(r)||\theta)$ is an increasing function of r for $r \geq \overline{r}^*$, we have

$$\frac{a_{r+1}}{a_r} = \frac{2^{D(\theta(r)||\theta)} - 1}{2^{D(\theta(r+1)||\theta)} - 1} \leq 2^{D(\theta(r)||\theta) - D(\theta(r+1)||\theta)} \,. \tag{63}$$

It can be readily verified that

$$D(\theta_1||\theta_2) = \frac{D_B(\theta_1||\theta_2)}{1 - \theta_1}$$

where the informational divergence $D_B(\cdot||\cdot)$ *for Bernoulli processes* is defined in Equation (17), Section 2. Thus, by (63),

$$\frac{a_{r+1}}{a_r} \leq 2^{-\frac{D_B(\theta(r+1)||\theta) - D_B(\theta(r)||\theta)}{1-\theta(r+1)}} \,. \tag{64}$$

Now,

$$\begin{aligned} D_B(\theta(r+1)||\theta) - D_B(\theta(r)||\theta) &= [\theta(r+1)-\theta(r)]\left[-\log S - \frac{h(\theta(r+1)) - h(\theta(r))}{\theta(r+1)-\theta(r)}\right] \\ &\geq [\theta(r+1)-\theta(r)][-\log S - h'(\theta(r))] \\ &= [\theta(r+1)-\theta(r)]\log\frac{S(r)}{S} \end{aligned} \tag{65}$$

where the inequality follows from $\theta(r+1) > \theta(r)$ and the concavity of the entropy function. By (55), we have $\theta(r) = [\theta(r+1)]^2$ for all $r \geq 1$, which together with (64) and (65) implies

$$\frac{a_{r+1}}{a_r} \leq 2^{-\theta(r+1)\log\frac{S(r)}{S}}\,. \tag{66}$$

Define $S^* \stackrel{\Delta}{=} S(\overline{r}^*)$, which is the smallest value of $S(r)$ larger than S and depends only on the actual parameter value. Since $\theta(r+1) \geq \theta(1) > \frac{1}{2}$, (66) yields for all $r \geq \max\{\overline{r}^*, 1\}$

$$\frac{a_{r+1}}{a_r} \leq \sqrt{\frac{S}{S^*}} < 1\,.$$

Thus,

$$\Delta_1 < \infty\,. \tag{67}$$

As for the sum Δ_2, we consider two cases: $\underline{r}^* > 0$ and $\underline{r}^* = 0$. In the first case, a code that contributes to the sum Δ_2 is selected to encode x_{t+1} whenever $S_t < tS(\underline{r}^* - 1)$. In addition, for codes Γ_r, $0 \leq r < \underline{r}^*$, or Γ'_0, the code length increase with respect to $\Gamma_{\underline{r}^*}$ for an integer x can be at most $M(x)$ bits per encoding, due to a longer unary part (the binary part decreases at least by one). Thus, the expected code length increase in (51) is uniformly upper-bounded for all codes that contribute to Δ_2, implying

$$\Delta_2 \leq E_{(\theta,\rho)}[M(x)] \sum_{t=1}^{n} \Pr\{S_t < tS(\underline{r}^* - 1)\}\,. \tag{68}$$

Substituting for $E_{(\theta,\rho)}[M(x)]$ in (68) as in Equation (53), and using the Chernoff bounding technique as we have $\theta > \theta(\underline{r}^* - 1)$, we obtain

$$\Delta_2 \leq \left[\frac{1+\theta}{1-\theta} - \rho\right] \sum_{t=1}^{\infty} 2^{-tD(\theta(\underline{r}^*-1)||\theta)} = \left[\frac{1+\theta}{1-\theta} - \rho\right] \frac{1}{2^{D(\theta(\underline{r}^*-1)||\theta)} - 1}\,. \tag{69}$$

Finally, in the case $\underline{r}^* = 0$, the magnitude of the average code length discrepancy between Γ_0 and Γ'_0 is $\rho' \stackrel{\Delta}{=} 2\max\{\rho, 1-\rho\} - 1$. In addition, the decision between the two codes is governed by N_t, which are statistics for a Bernoulli process with parameter ρ. Consequently,

$$\Delta_2 \leq \rho' \sum_{t=1}^{\infty} 2^{-tD_B(\frac{1}{2}||\rho)} = \frac{\rho'}{2^{D_B(\frac{1}{2}||\rho)} - 1}\,. \tag{70}$$

Theorem 3 follows from equations (58), (67), (69), and (70). □

5 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, "LOCO-I: A low complexity, context-based, lossless image compression algorithm," in *Proc. of the 1996 Data Compression Conference*, (Snowbird, Utah, USA), pp. 140–149, Mar. 1996.

[6] X. Wu, "An algorithmic study on lossless image compression," in *Proc. of the 1996 Data Compression Conference*, (Snowbird, Utah, USA), pp. 150–159, Mar. 1996.

[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] ISO/IEC JTC1/SC29 WG1 (JPEG/JBIG), "Information technology - Lossless and near-lossless compression of continuous-tone still images, Draft International Standard DIS14495-1 (JPEG-LS)," 1998.

[11] T. Linder, V. Tarokh, and K. Zeger, "Existence of optimal prefix codes for infinite source alphabets," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 2026–2028, Nov. 1997.

[12] A. Kato, T. S. Han, and H. Nagaoka, "Huffman coding with an infinite alphabet," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 977–984, May 1996.

[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 and J.-J. Lee, "Some practical universal noiseless coding techniques - Part II," Tech. Rep. JPL-83-17, Jet Propulsion Laboratory, Pasadena, CA, Mar. 1983.

[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] J. Abrahams, "Huffman-type codes for infinite source distributions," *J. Franklin Inst.*, vol. 331B, no. 3, pp. 265–271, 1994.

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

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

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

[21] 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.

[22] 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.

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

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

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

[26] S. W. Golomb, "Sources which maximize the choice of a Huffman coding tree," *Information and Control*, vol. 45, pp. 263–272, June 1980.

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

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

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