Quantization of Multiple Sources Using Nonnegative Integer Bit Allocation

Benjamin Farber and Kenneth Zeger, *Fellow, IEEE*

***Abstract*—Asymptotically optimal real-valued bit allocation among a set of quantizers for a finite collection of sources was derived in 1963 by Huang and Schultheiss, and an algorithm for obtaining an optimal nonnegative integer-valued bit allocation was given by Fox in 1966. We prove that, for a given bit budget, the set of optimal nonnegative integer-valued bit allocations is equal to the set of nonnegative integer-valued bit allocation vectors which minimize the Euclidean distance to the optimal real-valued bit-allocation vector of Huang and Schultheiss. We also give an algorithm for finding optimal nonnegative integer-valued bit allocations. The algorithm has lower computational complexity than Fox's algorithm, as the bit budget grows. Finally, we compare the performance of the Huang–Schultheiss solution to that of an optimal integer-valued bit allocation. Specifically, we derive upper and lower bounds on the deviation of the mean-squared error (MSE) using optimal integer-valued bit allocation from the MSE using optimal real-valued bit allocation. It is shown that, for asymptotically large transmission rates, optimal integer-valued bit allocations do not necessarily achieve the same performance as that predicted by Huang–Schultheiss for optimal real-valued bit allocations.**

***Index Terms*—Data compression, high-resolution quantization, source coding.**

I. Introduction

The classical bit allocation problem for lossy source coding is to determine the individual rates of a finite collection of scalar quantizers so as to minimize the sum of their distortions, subject to a constraint on the sum of the quantizer rates. Bit allocation arises in applications such as speech, image, and video coding. It has been shown [1], [21] that finding optimal integer bit allocations is NP-hard (as the number of sources grows), via reduction to the multiple-choice knapsack problem.

Huang and Schultheiss [20] analytically solved the bit allocation problem when the mean-squared error (MSE) of each quantizer decreases exponentially as its rate grows. The results in [20] were generalized in [26] by finding optimal real-valued bit allocations when the MSE of each quantizer is a convex function of its rate. Other generalizations were given in [17] and [24]. Bit allocation was studied in [3], in the context of trading off the total bit budget and the quantization error, a generalization of the Lagrangian approach.

The formulaic solution given in [20] allows arbitrary real-valued bit allocations. However, applications generally impose integer-value constraints on the rates used. In practice, bit allocations may be obtained by using some combinatorial optimization method such as integer linear programming or dynamic programming [10], [15], [16], [19], [30], [31], [34] or by optimizing with respect to the convex hull of the quantizers' rate-versus-distortion curves [6], [7], [25], [29], [32]. These techniques generally ignore the Huang–Schultheiss solution. Alternatively, a widely used technique is to explicitly use an optimal real-valued bit allocation as a starting point and then home in on an integer-valued bit allocation that is close by. As noted in the textbook by Gersho and Gray [14, pp. 230-231]:

> "In practice, … if an integer valued allocation is needed, then each non-integer allocation b_i is adjusted to the nearest integer. These modifications can lead to a violation of the allocation quota, B, so that some incremental adjustment is needed to achieve an allocation satisfying the quota. The final integer valued selection can be made heuristically. Alternatively, a local optimization of a few candidate allocations that are close to the initial solution obtained from [the Huang–Schultheiss solution] can be performed by simply computing the overall distortion for each candidate and selecting the minimum. … Any simple heuristic procedure, however, can be used to perform this modification."

In 1966, Fox [12] gave an algorithm for finding nonnegative integer-valued bit allocations. His algorithm is greedy in that at each step it allocates one bit to the quantizer whose distortion will be reduced the most by receiving an extra bit. Fox proved this intuitive approach is optimal for any convex decreasing quantizer distortion function. There are many other algorithmic techniques in the literature for obtaining integer-valued bit allocations. Some examples of these include [1], [4], [5], [13], [22], [23], [27], [35].

In this paper, we first prove that, for a given bit budget, the set of optimal nonnegative integer-valued bit allocations is equal to the set of nonnegative integer-valued bit allocation vectors which minimize the Euclidean distance to the optimal real-valued bit-allocation vector of Huang and Schultheiss. The proof of this result yields an alternate algorithm to that given by Fox for finding optimal nonnegative integer-valued bit allocations. This algorithm uses asymptotically (as the bit budget grows) less computational complexity than Fox's algorithm.

Despite the wealth of knowledge about bit allocation algorithms, there has been no published theoretical analysis com-

Manuscript received May 29, 2005; revised August 1, 2006. This work was supported by the National Science Foundation, Ericsson, and the UCSD Center for Wireless Communications. The material in this paper was presented in part at the Data Compression Conference, Snowbird, UT, March 2005.

B. Farber is with Fair Isaac Corp., San Diego, CA 92130 USA (e-mail: farber@code.ucsd.edu).

K. Zeger is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: zeger@ucsd.edu).

Communicated by S. A. Savari, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2006.883630

0018-9448/$20.00 © 2006 IEEE

paring the performance of optimal bit allocations with integer constraints to the performance obtained using the real-valued allocations due to Huang and Schultheiss.

We provide some such theoretical analysis. Specifically, we derive upper and lower bounds on the deviation of the MSE using optimal integer-valued bit allocation from the MSE using optimal real-valued bit allocation. Informally speaking, we show that no matter what bit budget is chosen, optimal integer-valued bit allocation might be as much as 6% worse than optimal real-valued bit allocation, but never more than 26% worse.

Our main results are summarized in the following (for $k \geq 2$).

i) For any k scalar sources and any bit budget, the set of optimal nonnegative integer-valued bit allocations is the same as the set of nonnegative integer-valued bit allocation vectors (with the same bit budget) which are closest to the optimal real-valued bit-allocation vector of Huang and Schultheiss (Theorem III.13).

ii) An algorithm is given for finding the set of optimal nonnegative integer-valued bit allocations from the Huang–Schultheiss optimal real-valued bit allocation (Algorithm III.14).

iii) For any k scalar sources, suppose the optimal real-valued bit allocation is *never* integer-valued for any bit budget. Then, the ratio of the MSE due to optimal nonnegative integer-valued bit allocation and the MSE due to optimal real-valued bit allocation is bounded away from 1 over all bit budgets (Theorem IV.3).

iv) There exist k scalar sources, such that for all bit budgets, the MSE due to optimal nonnegative integer-valued bit allocation is at least 6% greater than the MSE due to optimal real-valued bit allocation (Theorem IV.5).

v) For any k scalar sources and for all bit budgets, the MSE due to optimal integer-valued bit allocation is at most 26% greater than the MSE due to optimal real-valued bit allocation (Theorem V.2).

Our results are for memoryless scalar quantizers and possibly correlated sources. Cases i) and ii) are first established for integer-valued bit allocations and then extended to such allocations with nonnegative components. In case ii), the problem of finding an optimal nonnegative integer-valued bit allocation is reduced to first computing a particular real-valued bit allocation for the same bit budget, and then performing a (low-complexity) nearest neighbor search in a certain lattice using the real-valued bit allocation vector as the input to the search procedure. In each of the cases iii), iv), and v), we derive explicit bounds on the MSE penalty paid for using integer-valued bit allocation rather than real-valued bit allocation. A preliminary version of our results without the nonnegativity constraint was presented in [11].

This paper is organized as follows. Section II gives definitions, notation, and some lemmas. Section III shows the equivalence of closest nonnegative integer-valued bit allocation and optimal nonnegative integer-valued bit allocation. Section IV characterizes, for a given set of sources, the set of bit budgets for which no penalty occurs when using integer-valued bit allocation instead of real-valued bit allocation. Also, a lower bound is given on the ratio of the MSEs achieved by using optimal integer-valued bit allocation and optimal real-valued bit allocation. Section V presents an upper bound on the ratio of the MSEs achieved by using optimal integer-valued bit allocation and optimal real-valued bit allocation. The Appendix contains proofs of lemmas.

II. Preliminaries

Let $X_1, \dots, X_k$ be real-valued, possibly correlated, random variables (i.e., scalar sources) with variances $\sigma_1^2, \dots, \sigma_k^2$. Throughout this paper, we assume $k \geq 2$ and

$$0 < \sigma_1^2, \dots, \sigma_k^2 < \infty.$$

The sources $X_1, \dots, X_k$ are memoryless scalar quantized with resolutions $b_1, \dots, b_k$, respectively, measured in bits. The goal in bit allocation is to determine the k quantizer resolutions, subject to a constraint on their sum, so as to minimize the sum of the resulting MSEs.

Let $\mathbb{R}$ denote the reals and $\mathbb{Z}$ denote the integers. We will use the following notation:

$$\begin{aligned} b &= (b_1, \dots, b_k) \\ |u| &= \sum_{i=1}^{k} u_i \quad \forall u \in \mathbb{R}^k \\ g &= \left(\prod_{i=1}^{k} \sigma_i^2 \right)^{1/k} \\ \mathcal{A}_R(B) &= \{u \in \mathbb{R}^k : |u| = B\} \\ \mathcal{A}_I(B) &= \{u \in \mathbb{Z}^k : |u| = B\} \\ \mathcal{A}_I^+(B) &= \{u \in \mathbb{Z}^k : u_i \geq 0 \, \forall i, \ |u| = B\}. \end{aligned}$$

The vector b will be called a *bit allocation* and the integer $B \geq 1$ a *bit budget*. We say that b is a nonnegative bit allocation if $b_i \geq 0$ for all i. $\mathcal{A}_R(B)$, $\mathcal{A}_I(B)$, and $\mathcal{A}_I^+(B)$ are, respectively, the sets of all real-valued, integer-valued, and nonnegative integer-valued bit allocations b with bit budgets B. Bit allocations in $\mathcal{A}_I(B)$ and $\mathcal{A}_I^+(B)$ are said to be integer bit allocations. We use the notation $B \bmod k$ to represent the unique integer x satisfying $k \mid (B-x)$ and $0 \leq x \leq k-1$. If the components of two vectors are the same but ordered differently, then each vector is said to be a *permutation* of the other vector.

We will assume the MSE of the ith quantizer is equal to

$$d_i = h_i \sigma_i^2 4^{-b_i} \tag{1}$$

where h_i is a quantity dependent on the distribution of X_i, but independent of b_i. It is known that (1) is satisfied for asymptotically optimal scalar quantization [14], in which case

$$h_i = (1/12) \left(\int |f_{X_i/\sigma_i}|^{1/3} \right)^3$$

where f_{X_i} denotes the probability density function of X_i. Also, uniform quantizers satisfy (1), but with a different constant h_i. Many useful quantizers have distortions of the form in (1), as the distortion d_i in (1) often represents a reasonable approximation even for non-asymptotic bit rates.

The total (MSE) resulting from the bit allocation b is

$$d = \sum_{i=1}^{k} d_i.$$

We will also assume that

$$h_i = h$$

for all i. It is straightforward to generalize our results to the case where d is a weighted combination of the d_i's and where not all the h_i's are equal. Such multiplicative constants can essentially be absorbed by σ_i^2.

For any k scalar sources and for each bit budget B, let

$$\begin{aligned} a_{or}(B) &= \operatorname*{argmin}_{b \in \mathcal{A}_R(B)} \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} \\ d_{or} &= \sum_{i=1}^{k} h\sigma_i^2 4^{-a_{or}(B)_i}. \end{aligned}$$

We call $a_{or}(B)$ the *optimal real-valued bit allocation* and d_{or} the *MSE achieved by* $a_{or}(B)$. In 1963, Huang and Schultheiss [20] derived the optimal high-resolution real-valued bit allocation for the multiple-source quantization problem. Their result, stated in the following lemma, shows that $a_{or}(B)$ is unique.

Lemma II.1: For any k scalar sources and for each bit budget B,

$$\begin{aligned} a_{or}(B) &= \frac{B}{k}(\underbrace{1,\ldots,1}_{k}) + \frac{1}{2}\left(\log_2 \frac{\sigma_1^2}{g}, \ldots, \log_2 \frac{\sigma_k^2}{g}\right) \\ d_{or} &= khg4^{-B/k}. \end{aligned}$$

Lemma II.1 implies that the components of the bit allocation $a_{or}(B)$ are positive for a sufficiently large bit budget B; however, $a_{or}(B)$ need not be an integer bit allocation for any particular bit budget. The next lemma follows immediately from Lemma II.1. Let

$$(a_{or}(B) - b)_i$$

denote the ith component of the vector obtained from subtracting b component-wise from $a_{or}(B)$.

Lemma II.2: For any k scalar sources, for each bit budget B, and for any bit allocation $b \in \mathcal{A}_R(B)$, the MSE resulting from b is

$$d = hg4^{-B/k} \cdot \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i}.$$

For any k scalar sources and for each bit budget B, let

$$\begin{aligned} d_{oi} &= \min_{b \in \mathcal{A}_I(B)} \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} \\ \mathcal{A}_{oi}(B) &= \left\{ b \in \mathcal{A}_I(B) : \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} = d_{oi} \right\} \\ d_{oi}^{+} &= \min_{b \in \mathcal{A}_I^{+}(B)} \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} \\ \mathcal{A}_{oi}^{+}(B) &= \left\{ b \in \mathcal{A}_I^{+}(B) : \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} = d_{oi}^{+} \right\}. \end{aligned}$$

By Lemma II.1, these equations are equivalent to

$$\begin{aligned} d_{oi} &= \min_{b \in \mathcal{A}_I(B)} hg4^{-B/k} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} \\ \mathcal{A}_{oi}(B) &= \left\{ b \in \mathcal{A}_I(B) : hg4^{-B/k} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} = d_{oi} \right\} \end{aligned} \tag{2}$$

$$\begin{aligned} d_{oi}^{+} &= \min_{b \in \mathcal{A}_I^{+}(B)} hg4^{-B/k} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} \\ \mathcal{A}_{oi}^{+}(B) &= \left\{ b \in \mathcal{A}_I^{+}(B) : hg4^{-B/k} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} = d_{oi}^{+} \right\}. \end{aligned}$$

We call $\mathcal{A}_{oi}(B)$ the set of *optimal integer bit allocations* and d_{oi} the *MSE achieved by any bit allocation in* $\mathcal{A}_{oi}(B)$. The set $\mathcal{A}_{oi}^{+}(B)$ and the scalar d_{oi}^{+} are the analogous quantities for nonnegative bit allocations. In order to analyze $\mathcal{A}_{oi}^{+}(B)$ and d_{oi}^{+}, we will first obtain results about $\mathcal{A}_{oi}(B)$ and d_{oi}.

A. Lattice Tools

We next introduce some notation and terminology related to lattices that will be useful throughout the paper. We exploit certain facts from lattice theory to establish bit allocation results, specifically Theorems IV.5 and V.2. Most of the following definitions and notation are adapted from [8].

For any $w \in \mathbb{R}^m$, denote a set $\Gamma \subset \mathbb{R}^m$ translated by the vector w by

$$\Gamma + w = \{u + w : u \in \Gamma\}.$$

For any $k \geq 1$, define[1] the following lattice:

$$\Lambda_k = \{u \in \mathbb{Z}^{k+1} : |u| = 0\}.$$

The lattice Λ_{k-1} is useful for analyzing bit allocations for k scalar sources since it consists of points with k integer coordinates which sum to zero. For $0 \leq j \leq k$, define the $(k+1)$-dimensional vector

$$c(k,j) = \frac{1}{k+1}(\underbrace{-j,\ldots,-j}_{k+1-j}, \underbrace{k+1-j,\ldots,k+1-j}_{j}). \tag{3}$$

Note that

$$|c(k,j)| = 0$$

for all j and k.

[1] Usually denoted A_k in the literature. We use alternate notation to avoid confusion with sets of bit allocations.

Let $\|w\|$ denote the Euclidean norm of w. For any $k \geq 1$ and $w \in \mathbb{R}^{k+1}$, define

$$\Phi_k(w) = \left\{ u \in \Lambda_k : \|w-u\| = \min_{v \in \Lambda_k} \|w-v\| \right\}$$

i.e., the closest lattice points in Λ_k to w. Typically, $\Phi_k(w)$ contains a single point, however it can contain more than one point when w equidistant from muliple lattice points.

Lemma II.3: For any $w, y \in \mathbb{R}^{k+1}$

$$\left\{ u \in \Lambda_k + y : \|w-u\| = \min_{v \in \Lambda_k + y} \|w-v\| \right\} = \Phi_k(w-y) + y.$$

Denote the *Voronoi cell* associated with any point y in the lattice Λ_k by

$$V(y) = \{u \in \mathbb{R}^{k+1} : \|u-y\| \leq \|u-w\|, \ \forall w \in \Lambda_k\}.$$

This definition implies $V(y)$ shares boundary points with neighboring cells. Let

$$\mathcal{H}^k = \left\{ u \in \mathbb{R}^{k+1} : |u| = 0 \right\}.$$

The lattice Λ_k is a subset of $\mathbb{R}^{k+1}$ and also a subset of the k-dimensional hyperplane $\mathcal{H}^k$. Define the quantity

$$V_k(y) = V(y) \cap \mathcal{H}^k.$$

III. Closest Integer Bit Allocation

In this section, we first demonstrate the equivalence of closest integer bit allocation and optimal integer bit allocation. Then, we extend this equivalence to the case where the bit allocations must have nonnegative integer components. Finally, we obtain an algorithm for finding optimal nonnegative integer bit allocations.

For any k scalar sources and for each bit budget B, let

$$\begin{aligned}
\mathcal{A}_{ci}(B) &= \Bigg\{ b \in \mathcal{A}_I(B) : \|b - a_{or}(B)\| \\
&\qquad = \min_{\hat{b} \in \mathcal{A}_I(B)} \|\hat{b} - a_{or}(B)\| \Bigg\} \\
\mathcal{D}_{ci} &= \left\{ \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} : b \in \mathcal{A}_{ci}(B) \right\} \\
\Delta &= \mathcal{A}_{ci}(B) - a_{or}(B) \\
\mathcal{A}_{ci}^{+}(B) &= \Bigg\{ b \in \mathcal{A}_I^{+}(B) : \|b - a_{or}(B)\| \\
&\qquad = \min_{\hat{b} \in \mathcal{A}_I^{+}(B)} \|\hat{b} - a_{or}(B)\| \Bigg\} \\
\mathcal{D}_{ci}^{+} &= \left\{ \sum_{i=1}^{k} h\sigma_i^2 4^{-b_i} : b \in \mathcal{A}_{ci}^{+}(B) \right\}.
\end{aligned}$$

For a given bit budget B, $\mathcal{A}_{ci}(B)$ is the set of closest integer bit allocations, with respect to Euclidean distance, to the optimal real-valued bit allocation. Note that each $b \in \mathcal{A}_{ci}(B)$ is, in general, different from a bit allocation obtained by finding the closest integer to each component of $a_{or}(B)$, since such a component-wise closest bit allocation might result in using either more or less than B bits. The set Δ is a translate of $\mathcal{A}_{ci}(B)$ and is a function of $\sigma_1^2, \ldots, \sigma_k^2$ and B, although we will notationally omit these dependencies. $\mathcal{A}_{ci}^{+}(B)$ and $\mathcal{D}_{ci}^{+}$ are the analogous quantities to $\mathcal{A}_{ci}(B)$ and $\mathcal{D}_{ci}$, respectively, for nonnegative bit allocations.

The following lemma will be used to prove Lemmas III.2 and IV.4, and Theorem V.2. Define the quantities

$$\begin{aligned}
\mu &= \frac{1}{2}\left(\log_2 \frac{\sigma_1^2}{g}, \ldots, \log_2 \frac{\sigma_k^2}{g} \right) - c(k-1, B \bmod k) \\
M_B &= \bigcup_{\sigma_1^2, \ldots, \sigma_k^2} \Delta
\end{aligned}$$

and note that $\mu \in \mathcal{H}^{k-1}$. The union in the definition of M_B is over all k-tuples of sources that satisfy the assumptions made in Section II.

Lemma III.1: For any k scalar sources with variances $\sigma_1^2, \ldots, \sigma_k^2$ and for each bit budget B

$$\Delta = \Phi_{k-1}(\mu) - \mu.$$

Furthermore, $M_B = V_{k-1}(0)$ for all B.

The next lemma states that the smallest distance (in the Euclidean sense) that a closest integer bit allocation can be to the optimal real-valued bit allocation vector must occur when the bit budget is at most the number of sources.

Lemma III.2: For any k scalar sources

$$\inf_{\substack{w \in \Delta \\ B \geq 1}} \|w\| = \min_{\substack{w \in \Delta \\ 1 \leq B \leq k}} \|w\|.$$

A. An Algorithm for Finding $\mathcal{A}_{ci}(B)$

The following theorem is adapted from [9, pp. 230-231] and immediately yields an algorithm for finding closest integer bit allocation vectors (the components of the resulting bit allocation vectors need not all be nonnegative). For all $u \in \mathbb{R}$, define

$$\begin{aligned}
r(u) &= \lfloor u + (1/2) \rfloor \\
\rho(u) &= u - r(u).
\end{aligned}$$

The quantity $r(u)$ is a closest integer to u.

Theorem III.3: Let B be a bit budget,

$$\hat{b} = (r(a_{or}(B)_1), \ldots, r(a_{or}(B)_k)),$$

and

$$t = |\hat{b}| - B \in \mathbb{Z}.$$

Let $\mathcal{I}_k$ denote the set of all permutations $(i_1, \ldots, i_k)$ of $\{1, \ldots, k\}$ such that

$$-\frac{1}{2} \le \rho(a_{or}(B)_{i_1}) \le \cdots \le \rho(a_{or}(B)_{i_k}) < \frac{1}{2}$$

and let

$$\mathcal{R}^+ = \left\{ b \in \mathcal{A}_I(B) : \exists (i_1, \ldots, i_k) \in \mathcal{I}_k \text{ such that} \right.$$
$$\left. b_j = \begin{cases} \hat{b}_j - 1, & \text{if } j \in \{i_1, \ldots, i_t\} \\ \hat{b}_j, & \text{if } j \in \{i_{t+1}, \ldots, i_k\} \end{cases} \right\}$$
$$\mathcal{R}^- = \left\{ b \in \mathcal{A}_I(B) : \exists (i_1, \ldots, i_k) \in \mathcal{I}_k \text{ such that} \right.$$
$$\left. b_j = \begin{cases} \hat{b}_j + 1, & \text{if } j \in \{i_{k-t+1}, \ldots, i_k\} \\ \hat{b}_j, & \text{if } j \in \{i_1, \ldots, i_{k-t}\} \end{cases} \right\}.$$

Then

$$\mathcal{A}_{ci}(B) = \begin{cases} \{\hat{b}\}, & \text{if } t = 0 \\ \mathcal{R}^+, & \text{if } t > 0 \\ \mathcal{R}^-, & \text{if } t < 0\,. \end{cases}$$

Proof: For any $w \in \mathbb{Z}^k$

$$\left\| a_{or}(B) - \hat{b} \right\| \le \| a_{or}(B) - w \|\,.$$

Suppose $t = 0$. Then

$$\hat{b} \in \mathcal{A}_I(B) \subset \mathbb{Z}^k.$$

Thus, $\hat{b}$ is a point in $\mathcal{A}_I(B)$ of minimum distance to $a_{or}(B)$. This means that $\hat{b} \in \mathcal{A}_{ci}(B)$. Since $r(u)$ is a closest integer to u and since r breaks ties by rounding upward, any other integer bit allocation b with minimum distance from $a_{or}(B)$ must satisfy $|b| < |a_{or}(B)|$. Thus, $b \notin \mathcal{A}_I(B)$ and hence, $\mathcal{A}_{ci}(B) = \{\hat{b}\}$.

Suppose $t \ne 0$ and let

$$\mathcal{R} = \begin{cases} \mathcal{R}^+, & \text{if } t > 0 \\ \mathcal{R}^-, & \text{if } t < 0\,. \end{cases}$$

It can be seen that every element of $\mathcal{R}$ is a bit allocation $b \in \mathcal{A}_I(B)$ which minimizes the difference between $\|a_{or}(B) - b\|$ and $\|a_{or}(B) - \hat{b}\|$. Since $\|a_{or}(B) - \hat{b}\|$ does not depend on such b, we have

$$\begin{aligned} \mathcal{R} \subset \{ b \in \mathcal{A}_I(B) : \|a_{or}(B) - b\| \le \|a_{or}(B) - b'\| \\ \forall b' \in \mathcal{A}_I(B)\} \\ = \mathcal{A}_{ci}(B). \end{aligned}$$

To finish the proof, we will show that $\mathcal{A}_{ci}(B) \subset \mathcal{R}$. Let $b \in \mathcal{A}_{ci}(B)$. For any i and j, the following identity holds:

$$\begin{aligned} &[(b_i - 1) - a_{or}(B)_i]^2 + [(b_j + 1) - a_{or}(B)_j]^2 \\ &\quad - [b_i - a_{or}(B)_i]^2 - [b_j - a_{or}(B)_j]^2 \\ &= 2[1 + a_{or}(B)_i - b_i + b_j - a_{or}(B)_j]. \end{aligned} \tag{4}$$

Suppose there exists an i such that

$$b_i - a_{or}(B)_i \ge 1.$$

Then there must exist a j such that

$$b_j - a_{or}(B)_j < 0$$

since

$$\sum_l b_l = \sum_l a_{or}(B)_l = B.$$

But then the right-hand side of (4) would be negative which would imply $b \notin \mathcal{A}_{ci}(B)$, since subtracting 1 from b_i and adding 1 to b_j would result in an integer bit allocation closer than b to $a_{or}(B)$. A similar contradiction results in the case where

$$b_i - a_{or}(B)_i \le -1.$$

Thus, for every i, we must have

$$b_i \in \{\lfloor a_{or}(B)_i \rfloor, \lceil a_{or}(B)_i \rceil\}.$$

Since

$$\hat{b}_i \in \{\lfloor a_{or}(B)_i \rfloor, \lceil a_{or}(B)_i \rceil\}$$

we conclude that

$$|\hat{b}_i - b_i| \le 1$$

for all i.

Now, suppose $t > 0$. Then there exists at least one i such that

$$b_i = \hat{b}_i - 1 = \lfloor a_{or}(B)_i \rfloor < a_{or}(B)_i.$$

For each j, it cannot be the case that

$$b_j = \hat{b}_j + 1 = \lceil a_{or}(B)_j \rceil > a_{or}(B)_j,$$

for otherwise the Euclidean distance between b and $a_{or}(B)$ could be reduced by adding 1 to b_i and subtracting 1 from b_j, which violates the fact that $b \in \mathcal{A}_{ci}(B)$. Thus, for all i, we have

$$\hat{b}_i - b_i \in \{0, 1\}.$$

To minimize the distance between b and $a_{or}(B)$, the t components of b for which $\hat{b}_i - b_i = 1$ must be those components with the smallest values of $\rho(a_{or}(B)_i)$. Thus $b \in \mathcal{R}^+$.

A similar argument shows that if $t < 0$, then for all i, we have $\hat{b}_i - b_i \in \{0, -1\}$; this then implies that the t components of b for which $\hat{b}_i - b_i = -1$ must be those components with the largest values of $\rho(a_{or}(B)_i)$, i.e., $b \in \mathcal{R}^-$. In summary, $b \in \mathcal{R}$. □

Note that in practice $\mathcal{A}_{ci}(B)$ will usually consist of a single bit allocation, although in principle it can contain more than one bit allocation.

We note that Guo and Meng [18] gave a similar algorithm to that implied by Theorem III.3. Instead of rounding each component of the Huang–Schultheiss solution $a_{or}(B)$ to the nearest integer, they round each component down to the nearest integer from below. Then, they added 1 bit at a time to the rounded components, based on which components were rounded down the most. The technique implied from our Theorem III.3 uses the same idea, but also adds bits to components which were rounded up too far. The authors of [18] did not claim that their resulting bit allocation gave a closest integer bit allocation. They did, however, assert that their resulting bit allocation was optimal; but, in fact, their proof was not valid. They attempted to show that adding bits, one at a time, in the manner they described was optimal among all ways to add bits to the rounded bit allocation. However, their proof did not eliminate the possibility of adding more than two bits to multiple components of the rounded bit allocation. Nor did they rule out the possibility of subtracting extra bits from some components in order to add even more bits to other components. We believe their algorithm is indeed correct, despite the lack of proof.

Wintz and Kurtenbach [33, p. 656] also gave a similar algorithm for obtaining integer-valued bit allocations. Their technique was to round off the components of the Huang–Schultheiss solution to the nearest integer, and then add or subtract bits to certain components until the bit budget was satisfied. However, their choice of which components to adjust up or down was based on the magnitudes of the components, rather than how much they were initially truncated. The authors of [33] note that their technique is suboptimal.

The algorithm in [18] assumes the Huang–Schultheiss solution has nonnegative components, as does the algorithm implied by our Theorem III.3. However, in Section III-C, we generalize the result of Theorem III.3 to give an algorithm for finding optimal nonnegative integer bit allocations without any such assumptions about the Huang–Schultheiss solution.

B. Equivalence of Closest Integer Bit Allocations and Optimal Integer Bit Allocations

In this subsection, we allow bit allocations to have negative components. In Section III-C we will add the nonnegativity constraint. The next two technical lemmas are used to prove Lemma III.6.

Lemma III.4: For any k scalar sources and for each bit budget B, let $\beta \in \Delta$ be such that $\beta_j \in (-1/2, 1/2]$ for some j.

If $\beta_i < -1/2$, then

$$\begin{aligned}\beta_i &= -\rho(a_{or}(B)_i) - 1\\ \beta_j &= -\rho(a_{or}(B)_j)\\ \rho(a_{or}(B)_i) &\le \rho(a_{or}(B)_j).\end{aligned}$$

If $\beta_i > 1/2$, then

$$\begin{aligned}\beta_i &= -\rho(a_{or}(B)_i) + 1\\ \beta_j &= -\rho(a_{or}(B)_j)\\ \rho(a_{or}(B)_i) &\ge \rho(a_{or}(B)_j).\end{aligned}$$

Lemma III.5: For any k scalar sources and for each bit budget B, let

$$t = r(a_{or}(B)_1) + \cdots + r(a_{or}(B)_k) - B.$$

Then for any $\beta \in \Delta$ and for all i

$$\beta_i \in \begin{cases} (-1/2, 1/2], & \text{if } t = 0\\ (-1, 1/2], & \text{if } t > 0\\ (-1/2, 1), & \text{if } t < 0.\end{cases}$$

For each i and j, define a k-dimensional vector $\omega(i,j)$ whose components are

$$\omega(i,j)_l = \begin{cases} 1, & \text{if } l = i\\ -1, & \text{if } l = j\\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

Lemma III.6: For any k scalar sources, for each bit budget B, and for any $b \in \mathcal{A}_{ci}(B)$, let $\beta = b - a_{or}(B)$. Then for all i, j

$$\beta_j - \beta_i \le 1.$$

If $\beta_j - \beta_i = 1$, then

$$b + \omega(i,j) \in \mathcal{A}_{ci}(B). \tag{6}$$

The following theorem establishes that for each bit budget, the closest integer bit allocations and the optimal integer bit allocations are the same collections.

Theorem III.7: For any k scalar sources and for each bit budget B

$$\begin{aligned}\mathcal{A}_{ci}(B) &= \mathcal{A}_{oi}(B)\\ \mathcal{D}_{ci} &= \{d_{oi}\}.\end{aligned}$$

Proof: First, we show that $\mathcal{A}_{ci}(B) \subset \mathcal{A}_{oi}(B)$. Let $b \in \mathcal{A}_I(B)$ and $\tilde{b} \in \mathcal{A}_{ci}(B)$, and let d and $\tilde{d}$ denote the resulting MSEs, respectively. It suffices to show that $d \ge \tilde{d}$.

Define

$$\begin{aligned}\eta^+ &= \{l : b_l - \tilde{b}_l > 0\}\\ \eta^- &= \{l : b_l - \tilde{b}_l < 0\}\end{aligned}$$

and consider any sequence of integer bit allocation vectors

$$\tilde{b} = b^{(0)}, \dots, b^{(n)} = b \tag{7}$$

such that for each $m = 0, \dots, n-1$ there exists an $i \in \eta^+$ and a $j \in \eta^-$ such that

$$b^{(m+1)} - b^{(m)} = \omega(i,j). \tag{8}$$

By (8) we have that

$$b^{(m)} \in \mathcal{A}_I(B)$$

for each m since exactly one element of $b^{(m)}$ is incremented and exactly one element of $b^{(m)}$ is decremented going from $b^{(m)}$ to $b^{(m+1)}$. Such a sequence is guaranteed to exist since $|\tilde{b}| = |b|$. For each m, let $d^{(m)}$ be the MSE achieved by $b^{(m)}$. To establish $d \ge \tilde{d}$, we will show that $d^{(m)}$ is monotonic nondecreasing in m.

The construction of the sequence $b^{(0)}, \dots, b^{(n)}$ implies that for each $m = 0, \dots, n-1$

$$\begin{aligned}(b^{(m)} - b^{(0)})_i &\ge 0\\ (b^{(m)} - b^{(0)})_j &\le 0\end{aligned}$$

where $i \in \eta^+$ and $j \in \eta^-$ are defined by (8), and are functions of m. Thus,

$$(b^{(m)} - b^{(0)})_j \le (b^{(m)} - b^{(0)})_i.$$

Let

$$\beta^{(m)} = b^{(m)} - a_{or}(B).$$

Then

$$\beta_j^{(0)} - \beta_i^{(0)} \le 1 \quad \text{[from Lemma III.6]}$$

and therefore for each $m = 0, \ldots, n-1$, we get

$$\beta_j^{(0)} - \beta_i^{(0)} - (b^{(m)} - b^{(0)})_i \le 1 - (b^{(m)} - b^{(0)})_j$$

or equivalently (by the definition of $\beta^{(0)}$)

$$\begin{aligned} &-(b^{(0)} - a_{or}(B))_i - (b^{(m)} - b^{(0)})_i - 1 \\ &\qquad \le -(b^{(0)} - a_{or}(B))_j - (b^{(m)} - b^{(0)})_j. \qquad (9)\end{aligned}$$

Canceling terms in (9) and raising 4 to the remaining quantity on each side of the inequality gives

$$4^{-\beta_i^{(m)}-1} \le 4^{-\beta_j^{(m)}} \qquad (10)$$

or equivalently

$$\begin{aligned} 4^{-\beta_i^{(m)}} + 4^{-\beta_j^{(m)}} &\le 4^{-(\beta_i^{(m)}+1)} + 4^{-(\beta_j^{(m)}-1)} \\ &= 4^{-\beta_i^{(m+1)}} + 4^{-\beta_j^{(m+1)}} \quad \text{[from (8)]}\end{aligned}$$

which implies

$$\begin{aligned} d^{(m)} &= hg4^{-B/k} \cdot \sum_{l=1}^{k} 4^{-\beta_l^{(m)}} \quad \text{[from Lemma II.2]} \\ &\le hg4^{-B/k} \cdot \sum_{l=1}^{k} 4^{-\beta_l^{(m+1)}} \\ &= d^{(m+1)} \qquad \text{[from Lemma II.2]}. \quad (11)\end{aligned}$$

Thus, $d^{(m)}$ is monotonic and therefore we have shown

$$\mathcal{A}_{ci}(B) \subset \mathcal{A}_{oi}(B).$$

The fact that $\mathcal{D}_{ci} = \{d_{oi}\}$ then immediately follows.

Next, we show that $\mathcal{A}_{oi}(B) \subset \mathcal{A}_{ci}(B)$. Let $b \in \mathcal{A}_{oi}(B)$. Since $\mathcal{A}_{oi}(B) \subset \mathcal{A}_I(B)$, a decomposition as in (7) still holds. Our goal is to show $b \in \mathcal{A}_{ci}(B)$, which will be accomplished by showing $b^{(n)} \in \mathcal{A}_{ci}(B)$. By the optimality of b, we must have $d \le \tilde{d}$, which by the monotonicity of $d^{(m)}$ implies

$$d^{(n)} = d^{(0)}.$$

Hence, equality holds in (11) and therefore also in (10), which implies for each $m = 0, \ldots, n-1$ that

$$\beta_j^{(m)} - \beta_i^{(m)} = 1. \qquad (12)$$

Now, we use induction to show $b^{(n)} \in \mathcal{A}_{ci}(B)$. The $m = 0$ case holds since

$$b^{(0)} = \tilde{b} \in \mathcal{A}_{ci}(B).$$

Now suppose for all $m \le l$ (where $l \ge 1$) that

$$b^{(m)} \in \mathcal{A}_{ci}(B).$$

Then we can apply Lemma III.6 to (12) in the case $m = l$, and use (8) to obtain

$$b^{(l+1)} \in \mathcal{A}_{ci}(B). \qquad \square$$

Note that an immediate consequence of Theorem III.7 is that the components of every element in $\mathcal{A}_{oi}(B)$ tend to infinity as the bit budget grows without bound.

C. Equivalence of Closest Nonnegative Integer Bit Allocations and Optimal Nonnegative Integer Bit Allocations

The problem of finding nonnegative bit allocations was addressed by Segall [26], but his solution did not assure integer-valued quantizer resolutions. Fox [12] gave a greedy algorithm for finding nonnegative integer bit allocations by allocating one bit at a time to a set of quantizers. His algorithm is optimal for any convex decreasing distortion function, and in particular, it is optimal for the distortion function we assume in (1).

In this subsection, we prove (in Theorem III.13) that optimal nonnegative integer bit allocations are equivalent to closest nonnegative integer bit allocations. Our proof leads to an alternate algorithm for finding optimal nonnegative integer bit allocations. The algorithm is faster than Fox's algorithm (as the bit budget grows).

First we introduce some useful notation and then establish five lemmas that will be used to prove Theorem III.13.

For any bit budget B and any nonempty set $S \subset \{1, 2, \ldots, k\}$, define a vector $a_{or}(B, S) \in \mathbb{R}^k$ whose components are

$$a_{or}(B,S)_i = \begin{cases} \frac{B}{|S|} + \frac{1}{2}\log_2 \frac{\sigma_i^2}{g(S)}, & \text{if } i \in S \\ 0, & \text{otherwise} \end{cases}$$

where

$$g(S) = \left(\prod_{i \in S} \sigma_i^2 \right)^{1/|S|}.$$

Lemma II.1 shows that the $|S|$-dimensional vector obtained by extracting the coordinates of $a_{or}(B, S)$, corresponding to the elements of S, is the optimal real-valued bit allocation for the quantizers corresponding to the elements of S. For any bit budget B, any bit allocation $b \in \mathcal{A}_R(B)$, and any nonempty set $S \subset \{1, \ldots, k\}$, let

$$\begin{aligned} \theta_1(b) &= \|b - a_{or}(B,S)\| \\ \theta_2(b) &= hg4^{-B/k} \sum_{i=1}^{k} 4^{(a_{or}(B,S)-b)_i}\end{aligned}$$

and for any set $T \subset \mathbb{Z}^k$ and any function $f : T \rightarrow \mathbb{R}$, let

$$Q(T, f) = \left\{ b \in T : f(b) = \min_{\hat{b} \in T} f(\hat{b}) \right\}.$$

Define the quantities

$$\begin{aligned}
\mathbb{Z}_S^k &= \{u \in \mathbb{Z}^k : u_j = 0 \ \forall j \notin S\} \\
\mathcal{A}_I(B, S) &= \{u \in \mathbb{Z}_S^k : |u| = B\} \\
\mathcal{A}_I^+(B, S) &= \{u \in \mathcal{A}_I(B, S) : u_i \geq 0 \ \forall i\} \\
\mathcal{A}_{ci}(B, S) &= Q(\mathcal{A}_I(B, S), \theta_1) \\
\mathcal{A}_{ci}^+(B, S) &= Q(\mathcal{A}_I^+(B, S), \theta_1) \\
\mathcal{A}_{oi}^+(B, S) &= Q(\mathcal{A}_I^+(B, S), \theta_2).
\end{aligned}$$

For a given bit budget B, $\mathcal{A}_{ci}(B, S)$ is the set of closest integer bit allocations to $a_{or}(B, S)$, and $\mathcal{A}_{ci}^+(B, S)$ is the set of closest nonnegative integer bit allocations to $a_{or}(B, S)$.

Lemma III.8: If $Q(W, f) \subset V \subset W$, then $Q(W, f) = Q(V, f)$.

The following lemma shows that to find a closest integer bit allocation to $a_{or}(B, S)$, one can assume without loss of generality that zeros are located in bit allocation vector components corresponding to integers not in S.

Lemma III.9: For each bit budget B and for any nonempty set $S \subset \{1, 2, \ldots, k\}$

$$\mathcal{A}_{ci}(B, S) = Q(\mathcal{A}_I(B), \theta_1).$$

Lemma III.10: For any k scalar sources, for each bit budget B, and for any nonempty set $S \subset \{1, 2, \ldots, k\}$, if $a_{or}(B, S)$ is nonnegative, then every bit allocation in $\mathcal{A}_{ci}(B, S)$ is nonnegative.

Lemma III.11: Consider k scalar sources with bit budget B and a nonempty set $S \subset \{1, 2, \ldots, k\}$. If $\mathcal{A}_{ci}^+(B) \subset \mathcal{A}_I(B, S)$, then $\mathcal{A}_{ci}^+(B) = \mathcal{A}_{ci}^+(B, S)$. If $\mathcal{A}_{oi}^+(B) \subset \mathcal{A}_I(B, S)$, then $\mathcal{A}_{oi}^+(B) = \mathcal{A}_{oi}^+(B, S)$.

Lemma III.12: Consider k scalar sources with bit budget B and a nonempty set $S \subset \{1, 2, \ldots, k\}$. Suppose $\mathcal{A}_{ci}^+(B), \mathcal{A}_{oi}^+(B) \subset \mathcal{A}_I(B, S)$ and there exists an $i \in S$ such that $a_{or}(B, S)_i < 0$. Then $b_i = 0$ for all $b \in \mathcal{A}_{ci}^+(B) \cup \mathcal{A}_{oi}^+(B)$.

The following theorem shows that optimal nonnegative integer bit allocation is equivalent to closest nonnegative integer bit allocation. In other words, minimizing the distortion among all nonnegative integer bit allocations is equivalent to finding which nonnegative integer bit allocation vectors are closest in Euclidean distance to the Huang-Schultheiss real-valued bit-allocation vector. This, in turn, can be accomplished with a nearest neighbor search in a lattice. Following Theorem III.13, we give an efficient algorithm for finding optimal nonnegative integer bit allocation vectors.

Theorem III.13: For any k scalar sources and for each bit budget B

$$\begin{aligned}
\mathcal{A}_{ci}^+(B) &= \mathcal{A}_{oi}^+(B) \\
\mathcal{D}_{ci}^+ &= \{d_{oi}^+\}.
\end{aligned}$$

Proof: Let $S^{(0)} = \{1, \ldots, k\}$ and consider the sequence of bit allocations

$$a_{or}(B, S^{(0)}), \ldots, a_{or}(B, S^{(n)})$$

where

$$S^{(m+1)} = \{i \in S^{(m)} : a_{or}(B, S^{(m)})_i \geq 0\}$$

and n is the smallest nonnegative integer such that

$$a_{or}(B, S^{(n)})_i \geq 0$$

for all i. Such an integer n exists since the following hold.

- $|S^{(m)}| \geq 1$, for all m.
- If $|S^{(m)}| = 1$, then $a_{or}(B, S^{(m)})_i \geq 0$, for all i.
- $|S^{(m)}|$ is monotone decreasing in m.

We will show that both $\mathcal{A}_{ci}^+(B)$ and $\mathcal{A}_{oi}^+(B)$ are equal to $\mathcal{A}_{ci}(B, S^{(n)})$. The fact that $\mathcal{D}_{ci}^+ = \{d_{oi}^+\}$ then follows from the definition of $\mathcal{D}_{ci}^+$.

Note that for any $m \geq 0$, if

$$\mathcal{A}_{oi}^+(B), \mathcal{A}_{ci}^+(B) \subset \mathcal{A}_I(B, S^{(m)})$$

then (by Lemma III.12) any optimal or closest nonnegative integer bit allocation b must satisfy

$$b_i = 0$$

for $i \notin S^{(m+1)}$, and therefore,

$$\mathcal{A}_{oi}^+(B), \mathcal{A}_{ci}^+(B) \subset \mathcal{A}_I(B, S^{(m+1)}).$$

Thus, since

$$\mathcal{A}_{oi}^+(B), \mathcal{A}_{ci}^+(B) \subset \mathcal{A}_I(B) = \mathcal{A}_I(B, S^{(0)})$$

we obtain by induction that

$$\mathcal{A}_{oi}^+(B), \mathcal{A}_{ci}^+(B) \subset \mathcal{A}_I(B, S^{(n)}). \tag{13}$$

Now using (13) and Lemma III.11, we have

$$\mathcal{A}_{ci}^+(B) = \mathcal{A}_{ci}^+(B, S^{(n)}) \tag{14}$$

$$\mathcal{A}_{oi}^+(B) = \mathcal{A}_{oi}^+(B, S^{(n)}). \tag{15}$$

Since $a_{or}(B, S^{(n)})$ is nonnegative by definition, Lemma III.10 implies that each $b \in \mathcal{A}_{ci}(B, S^{(n)})$ is nonnegative, i.e.,

$$\mathcal{A}_{ci}(B, S^{(n)}) \subset \mathcal{A}_I^+(B, S^{(n)}). \tag{16}$$

From (16) and the fact that

$$\mathcal{A}_I^+(B, S^{(n)}) \subset \mathcal{A}_I(B, S^{(n)})$$

we can apply Lemma III.8 with

$$\begin{aligned}
W &= \mathcal{A}_I(B, S^{(n)}) \\
V &= \mathcal{A}_I^+(B, S^{(n)}) \\
f &= \theta_1
\end{aligned}$$

to obtain

$$\mathcal{A}_{ci}(B, S^{(n)}) = \mathcal{A}_{ci}^{+}(B, S^{(n)}).$$

Thus, we have

$$\mathcal{A}_{ci}^{+}(B) = \mathcal{A}_{ci}(B, S^{(n)}) \qquad \text{[from (14)]}.$$

Now consider a set of sources $\hat{X}_1, \ldots, \hat{X}_k$ with variances

$$\hat{\sigma}_i^2 = \begin{cases} \frac{\sigma_i^2}{g(S^{(n)})} 4^{\frac{B(k-|S^{(n)}|)}{k|S^{(n)}|}}, & \text{if } i \in S^{(n)} \\ 4^{-B/k}, & \text{if } i \notin S^{(n)} . \end{cases}$$

Lemma II.1 shows that $a_{or}(B, S^{(n)})$ is the optimal real-valued bit allocation for $\hat{X}_1, \ldots, \hat{X}_k$ (mimicking the argument from the proof of Lemma III.10). Therefore, by Lemma III.9, $\mathcal{A}_{ci}(B, S^{(n)})$ is the set of closest integer bit allocations (without requiring any zero components) for $\hat{X}_1, \ldots, \hat{X}_k$. Hence, by Theorem III.7, $\mathcal{A}_{ci}(B, S^{(n)})$ is also the set of optimal integer bit allocations for $\hat{X}_1, \ldots, \hat{X}_k$. Thus,

$$\begin{aligned} \mathcal{A}_{ci}(B, S^{(n)}) &= Q(\mathcal{A}_I(B), \theta_2) && \text{[from (2)]} \quad (17) \\ &\subset \mathcal{A}_I^{+}(B, S^{(n)}) && \text{[from (16), (17)]} \\ &\subset \mathcal{A}_I(B). \end{aligned}$$

Now applying Lemma III.8 with

$$\begin{aligned} W &= \mathcal{A}_I(B) \\ V &= \mathcal{A}_I^{+}(B, S^{(n)}) \\ f &= \theta_2 \end{aligned}$$

gives

$$Q(\mathcal{A}_I(B), \theta_2) = \mathcal{A}_{oi}^{+}(B, S^{(n)}).$$

Therefore, we have

$$\mathcal{A}_{oi}^{+}(B) = \mathcal{A}_{ci}(B, S^{(n)}) \qquad \text{[from (15), (17)]}. \qquad \square$$

The proof of Theorem III.13 yields an alternative procedure to that given by Fox [12] for finding optimal nonnegative integer bit allocations. The main idea is to remove any negative components in the Huang-Schultheiss real-valued solution and then re-compute the Huang-Schultheiss solution for the surviving quantizers, iteratively repeating this procedure until no negative components remain. Then, the set of closest integer-valued vectors (with the same bit budget) to the resulting nonnegative real-valued vector is computed as the output of the algorithm.

Algorithm III.14: ***(Procedure to Find $\mathcal{A}_{oi}^{+}(B)$ and $\mathcal{A}_{ci}^{+}(B)$)***

For any k scalar sources and for each bit budget B, the following procedure generates a set of bit allocations which is both the set $\mathcal{A}_{oi}^{+}(B)$ and the set $\mathcal{A}_{ci}^{+}(B)$.

- Step 1: Set $S = \{1, 2, \ldots, k\}$.
- Step 2: Compute $a_{or}(B, S)$ and let
$$J = \{i \in S : a_{or}(B, S)_i \geq 0\}.$$
- Step 3: If $J = S$ go to Step 4.
Otherwise, set $S = J$ and go to Step 2.
- Step 4: Set $a_{or}(B)$ equal to $a_{or}(B, S)$ in Theorem III.3 and then compute $\mathcal{A}_{ci}(B)$.
Set $\mathcal{A}_{oi}^{+}(B) = \mathcal{A}_{ci}^{+}(B) = \mathcal{A}_{ci}(B)$.

Remark: We briefly remark on the computational complexity of the algorithm above as a function of the bit budget B, for a fixed k. When there exists a unique closest nonnegative integer bit allocation, the computational complexity of the algorithm reduces to the complexity of determining $\mathcal{A}_{ci}(B)$. The complexity of this lattice search is known to be constant in B (e.g., see [9, p. 231]). In contrast, Fox's algorithm has complexity linear in B. Thus, for large B, Algorithm III.14 is faster than Fox's algorithm.

Also, by examining the proof of Theorem III.13, one can readily verify a possible modification to Algorithm III.14. Namely, in Step 2 of the algorithm, instead of zeroing out all negative components of the Huang–Schultheiss bit allocation, one could zero out one negative component per iteration in the algorithm. The optimal nonnegative integer bit allocation is still achieved.

IV. Distortion Penalty for Integer Bit Allocations

For any k scalar sources and for each bit budget B, let

$$p^{oi+} = \frac{d_{oi}^{+}}{d_{or}}.$$

We call p^{oi+} the *distortion penalty* resulting from optimal nonnegative integer bit allocation. The distortion penalty measures the increase in distortion when one uses (the practical) optimal nonnegative integer bit allocation instead of (the fictitious) optimal real-valued bit allocation given by the Huang–Schultheiss formula. For any $b \in \mathcal{A}_{oi}^{+}(B)$, we have

$$p^{oi+} = \frac{1}{k} \sum_{i=1}^{k} 4^{(a_{or}(B) - b)_i} \qquad \text{[from (2), Lemma II.1]}. \quad (18)$$

Also, clearly $p^{oi+} \geq 1$.

It is straightforward to see that for any k scalar sources with variances $\sigma_1^2, \ldots, \sigma_k^2$ and a bit budget B, the following three statements are equivalent.

i) $p^{oi+} = 1$.

ii) The optimal real-valued bit allocation is a nonnegative integer bit allocation.

iii) $\frac{1}{2} \log_2 \frac{\sigma_i^2}{g} + \frac{B \bmod k}{k} \in \mathbb{Z} \quad \forall i$.

Lemma IV.1: For any $w \in \mathcal{A}_R(0)$

$$\begin{aligned} &4^{-\|w\|\sqrt{(k-1)/k}} + (k-1)4^{\|w\|\sqrt{1/(k(k-1))}} \\ &\quad \leq \sum_{i=1}^{k} 4^{-w_i} \\ &\quad \leq 4^{\|w\|\sqrt{(k-1)/k}} + (k-1)4^{-\|w\|\sqrt{1/(k(k-1))}}. \end{aligned}$$

For any $b \in \mathcal{A}_R(B)$, if $w = b - a_{or}(B)$, then Lemma IV.1 gives bounds on the sum in Lemma II.2. Moreover, both the upper and lower bounds in Lemma IV.1 are functions only of k

and $\|w\|$, both bounds are monotone increasing with $\|w\|$, and as $\|w\| \to 0$ the bounds become tight.

Lemma IV.2: For any k scalar sources, for each bit budget B, and for any bit allocation $b \in \mathcal{A}_I(B)$, the MSE d resulting from b satisfies

$$\begin{aligned} &hg4^{-B/k} \cdot \Big(4^{-\|b-a_{or}(B)\|\sqrt{(k-1)/k}} \\ &\quad +(k-1)4^{\|b-a_{or}(B)\|\sqrt{1/(k(k-1))}}\Big) \\ &\le d \\ &\le hg4^{-B/k} \cdot \Big(4^{\|b-a_{or}(B)\|\sqrt{(k-1)/k}} \\ &\quad +(k-1)4^{-\|b-a_{or}(B)\|\sqrt{1/(k(k-1))}}\Big). \end{aligned}$$

For any k scalar sources, define

$$\delta = \min_{B\ge 1} \min_{b\in\mathcal{A}_{oi}(B)} \|b - a_{or}(B)\|.$$

The quantity δ is the minimum distance, for k fixed sources, between an optimal integer bit allocation (with possibly negative values) and the optimal real-valued bit allocation vector, over all bit budgets. Lemma III.2 and Theorem III.7 show that we can write

$$\delta = \min_{1\le B\le k} \min_{b\in\mathcal{A}_{ci}(B)} \|b - a_{or}(B)\|.$$

Hence, δ is simple to compute since $\mathcal{A}_{ci}(B)$ typically consists of a single bit allocation, easily found using Theorem III.3. The minimal value of the quantity δ is 0, which occurs when $a_{or}(B)$ is integer-valued for some bit budget. The maximum value of δ is

$$\sqrt{\frac{(k+1)(k-1)}{12k}}$$

which is the covering radius of the dual of the lattice Λ_{k-1} (see [8, p. 115]).

One can show that for k scalar sources, if $\delta = 0$, then there exists a nonnegative integer $n \le k-1$ such that for each sufficiently large bit budget B

$$p^{oi+} = 1 \text{ if and only if } B \bmod k = n.$$

Theorem IV.3 examines the case when $\delta > 0$. In this case

$$a_{or}(B) \notin \mathcal{A}_{oi}(B)$$

for all B. Theorem IV.3 shows that, in fact, if

$$a_{or}(B) \notin \mathcal{A}_{oi}^{+}(B)$$

for all B, then the distortion penalty resulting from optimal nonnegative integer bit allocation is bounded away from 1 for all bit budgets. This may appear surprising since one might expect the distortion penalty due to optimal nonnegative integer bit allocation to tend to 1 as the bit budget grows.

Theorem IV.3: Consider k scalar sources. If $\delta > 0$, then for every bit budget B

$$p^{oi+} \ge \frac{1}{k}\left(4^{-\delta\sqrt{(k-1)/k}} + (k-1)4^{\delta\sqrt{1/(k(k-1))}}\right) > 1.$$

Proof: For each bit budget B, for any $b \in \mathcal{A}_{oi}^{+}(B)$, and for any $b' \in \mathcal{A}_{oi}(B)$, we have

$$b \in \mathcal{A}_{ci}^{+}(B)$$

and

$$b' \in \mathcal{A}_{ci}(B)$$

by Theorems III.13 and III.7, respectively. Since bit allocations in $\mathcal{A}_{ci}^{+}(B)$ minimize the distance to $a_{or}(B)$ over a smaller set of integer bit allocations than bit allocations in $\mathcal{A}_{ci}(B)$

$$\|b - a_{or}(B)\| \ge \|b' - a_{or}(B)\|. \quad (19)$$

The definition of δ and (19) imply

$$\|b - a_{or}(B)\| \ge \delta > 0. \quad (20)$$

Define a function $f : [0,\infty) \to (0,\infty)$ by

$$f(u) = 4^{-u\sqrt{(k-1)/k}} + (k-1)4^{u\sqrt{1/(k(k-1))}}.$$

For each bit budget B and for every $b \in \mathcal{A}_{oi}^{+}(B)$

$$\begin{aligned} p^{oi+} &= \frac{d_{oi}^{+}}{d_{or}} \\ &\ge \frac{hg4^{-B/k}}{d_{or}} \cdot f(\|b - a_{or}(B)\|) \quad \text{[from Lemma IV.2]} \\ &= \frac{1}{k} f(\|b - a_{or}(B)\|) \quad \text{[from Lemma II.1]} \\ &\ge \frac{1}{k} f(\delta) \quad \text{[from (20) and the monotonicity of } f\text{]} \\ &> 1 \quad \text{[from the arithmetic-geometric mean inequality].} \end{aligned}$$

□

A. Lower Bound on Worst Case Distortion Penalty for Integer Bit Allocations

For any particular set of k sources, the distortion obtained by using optimal nonnegative integer-valued bit allocation may be larger than the distortion predicted by optimal real-valued bit allocation. Theorem IV.5 below illustrates how much worse nonnegative integer-valued bit allocation can be compared to real-valued bit allocation.

Let

$$\gamma_k = \frac{1}{2k+2}\big(-k, -k+2, \ldots, k-2, k\big).$$

Lemma IV.4: If the variances $\sigma_1^2, \ldots, \sigma_k^2$ of k scalar sources satisfy

$$\frac{1}{2}\left(\log_2 \frac{\sigma_1^2}{g}, \ldots, \log_2 \frac{\sigma_k^2}{g}\right) = \gamma_{k-1}$$

then for each bit budget B and for any $b \in \mathcal{A}_{oi}^{+}(B)$, the vector $b - a_{or}(B)$ is a permutation of γ_{k-1}.

Theorem IV.5: For each k, there exist k scalar sources, such that for any bit budget, the distortion penalty resulting from optimal nonnegative integer bit allocation satisfies

$$p^{oi+} = \frac{3 \cdot 2^{(k-1)/k}}{k(4 - 4^{(k-1)/k})} > 1.$$

The distortion penalty in Theorem IV.5 is monotone increasing with k and is bounded as

$$1.06 \approx \frac{3\sqrt{2}}{4} \le p^{oi+} \le \frac{3}{4 \ln 2} \approx 1.08$$

where the lower bound is attained at $k = 2$ and the upper bound is approached as $k \to \infty$. Thus, the theorem guarantees that for some sources, the MSE due to optimal nonnegative integer-valued bit allocation is at least 6% greater (and as much as 8% greater for large k) than the MSE due to optimal real-valued bit allocation. We do not claim this is the largest or smallest possible distortion penalty—indeed p^{oi+} can range from 1, when $a_{or}(B)$ happens to be nonnegative integer-valued, to ∞ as shown by Theorem V.1. Rather, Theorem IV.5 demonstrates that p^{oi+} can be bounded away from 1. Unfortunately, one cannot qualify the distortion penalty in Theorem IV.5 as typical, or atypical, without first defining what constitues a typical set of sources. We leave this task to the reader for any particular application.

Proof of Theorem IV.5: Let $a > 0$ be arbitrary. For each $i \le k$, consider a scalar source whose variance is given by

$$\sigma_i^2 = a 4^{(\gamma_{k-1})_i}.$$

Then

$$g = \left(\prod_{i=1}^{k} \sigma_i^2 \right)^{1/k} = a$$

$$\frac{1}{2} \left(\log_2 \frac{\sigma_1^2}{g}, \ldots, \log_2 \frac{\sigma_k^2}{g} \right) = \gamma_{k-1}$$

and Lemma IV.4 implies that for each bit budget B and for any $b \in \mathcal{A}_{oi}^{+}(B)$, the vector $b - a_{or}(B)$ is a permutation of γ_{k-1}. Hence, for each B

$$\begin{aligned} p^{oi+} &= \frac{1}{k} \sum_{i=1}^{k} 4^{-(\gamma_{k-1})_i} \qquad \text{[from (18)]} \\ &= \frac{1}{k} \sum_{i=0}^{k-1} 4^{-[(-(k-1)+2i)/2k]} \\ &= \frac{2^{(k-1)/k}}{k} \sum_{i=0}^{k-1} 4^{-i/k} \qquad (21) \\ &= \frac{2^{(k-1)/k}}{k} \cdot \frac{1 - \left(4^{-1/k}\right)^k}{1 - 4^{-1/k}} \\ &= \frac{3 \cdot 2^{(k-1)/k}}{k(4 - 4^{(k-1)/k})}. \end{aligned}$$

Applying the arithmetic-geometric mean inequality to (21) gives $p^{oi+} > 1$. □

We note that for the sources used in the proof of Theorem IV.5, the lower bound in Theorem IV.5 is greater than that given in Theorem IV.3, for all k.

V. Upper Bound on Distortion Penalty for Integer Bit Allocations

The Huang–Schultheiss formula gives a bit allocation which can include the fictitious concept of "negative" bits. In practice, such negative bits tend to disappear as the bit budget B grows. However, for any bit budget B, there exist collections of pathological sources that always lead to negative bits in the Huang–Schultheiss allocation. As a result, when restricted to using nonnegative integer bit allocations, these pathological sources prevent one from obtaining a finite uniform upper bound on the distortion penalty p^{oi+}. This fact is demonstrated in Theorem V.1 below.

In contrast, by mathematically allowing negative integer bit allocations (to more closely approximate Huang–Schultheiss allocations containing negative bits), a useful upper bound on a new distortion penalty p^{oi} can be obtained. For any k scalar sources and for each bit budget B, let

$$p^{oi} = \frac{d_{oi}}{d_{or}}.$$

We call p^{oi} the *distortion penalty* resulting from optimal integer bit allocation.

One implication of Lemma III.5 and Theorem III.7 is that no component of an optimal integer bit allocation can differ form the corresponding component of the Huang–Schultheiss allocation by more than one, i.e., $|a_{or}(B)_i - b_i| < 1$, for all i. Thus, we have $p^{oi} \le p^{oi+}$, with equality whenever the Huang–Schultheiss bit allocation is nonnegative. An upper bound on p^{oi} is only practical, however, for sources whose Huang–Schultheiss bit allocation is nonnegative. Such an upper bound is given in Theorem V.2.

Theorem V.1: For each $k \ge 2$ and for any bit budget

$$\sup_{\sigma_1^2, \ldots, \sigma_k^2} p^{oi+} = \infty$$

where the supremum is taken over all k-tuples of sources with positive, finite variances.

Proof: For a bit budget B, let $N > B/k$ and suppose the variances $\sigma_1^2, \ldots, \sigma_k^2$ of k scalar sources satisfy

$$\sigma_i^2 = \begin{cases} 2^{2kN}, & \text{if } i = 1 \\ 1, & \text{if } 1 < i \le k. \end{cases}$$

Then

$$g = \left(\prod_{i=1}^{k} \sigma_i^2 \right)^{1/k} = 2^{2N}$$

$$\frac{1}{2} \left(\log_2 \frac{\sigma_1^2}{g}, \ldots, \log_2 \frac{\sigma_k^2}{g} \right) = \left(N(k-1), -N, \ldots, -N \right)$$

and Lemma II.1 implies

$$a_{or}(B)_i = \begin{cases} (B/K) + N(k-1), & \text{if } i = 1 \\ (B/k) - N, & \text{if } 1 < i \le k\,. \end{cases}$$

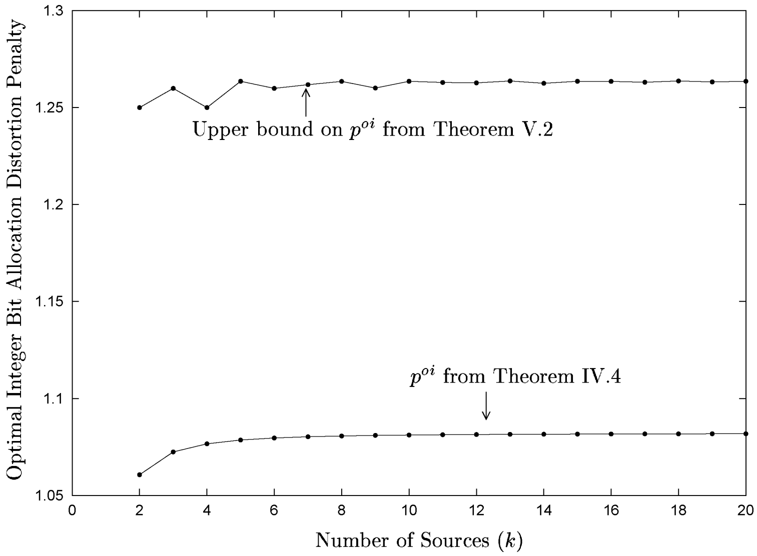

Fig. 1. Plot of the achievable distortion penalty from Theorem IV.5 and the upper bound on the distortion penalty from Theorem V.2.

Algorithm III.14 shows that

$$\mathcal{A}_{oi}^{+}(B) = \{(B, 0, \ldots, 0)\}.$$

Hence, by (18) we have

$$\begin{aligned} p^{oi+} &= \frac{1}{k}\left(4^{(k-1)[N-(B/k)]} + (k-1)4^{(B/k)-N}\right) \\ &\longrightarrow \infty \text{ as } N \to \infty. \qquad \square \end{aligned}$$

In the following theorem we give an upper bound on the distortion penalty resulting from optimal integer bit allocation. The bound does not depend on the source distribution or the bit budget.

Theorem V.2: For each $k \geq 2$, for any k scalar sources, and for any bit budget, the distortion penalty resulting from optimal integer bit allocation is upper-bounded as

$$p^{oi} \leq 4^{\tau}\left(1 - \frac{3\tau}{4}\right)$$

where

$$\tau = \frac{1}{k}\left\lceil \frac{4k}{3} - \frac{1}{1-4^{-1/k}} \right\rceil.$$

The upper bound on p^{oi} in Theorem V.2 is tight since, for arbitrary $a > 0$, if

$$\sigma_i^2 = a4^{-c(k-1,k\tau)_i}, \qquad 1 \leq i \leq k$$

and the bit budget B is a multiple of k, then by Theorem III.3, Theorem III.7, and (18) we have

$$p^{oi} = 4^{\tau}\left(1 - \frac{3\tau}{4}\right).$$

For all $k \geq 2$, the upper bound on p^{oi} in Theorem V.2 satisfies

$$1.25 \leq 4^{\tau}\left(1 - \frac{3\tau}{4}\right) < \frac{3}{e2^{1/3}\ln 2} \approx 1.26 \qquad (22)$$

where the lower bound in (22) is attained at $k = 2$ and $k = 4$ and the upper bound in (22) is approached as $k \to \infty$. Thus, Theorem V.2 guarantees that for any k scalar sources and for all bit budgets, the MSE due to optimal integer-valued bit allocation is at most 26% greater than the MSE due to optimal real-valued bit allocation.

Fig. 1 compares the upper bound in Theorem V.2 with the distortion penalty from Theorem IV.5.

Proof: We show that

$$\sup_{B} \sup_{\sigma_1^2,\ldots,\sigma_k^2} \frac{d_{oi}}{d_{or}} = 4^{\tau}\left(1 - \frac{3\tau}{4}\right)$$

where, for a fixed k, the suprema are taken over all possible k-tuples of sources and over all bit budgets.

Define a mapping $f : \mathbb{R}^k \to \mathbb{R}$ by

$$f(u) = \sum_{i=1}^{k} 4^{-u_i}.$$

Then we have

$$\begin{aligned} &\sup_{B} \sup_{\sigma_1^2,\ldots,\sigma_k^2} \frac{d_{oi}}{d_{or}} \\ &= \frac{1}{k} \sup_{B} \sup_{\sigma_1^2,\ldots,\sigma_k^2} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} \;\; \forall b \in \mathcal{A}_{oi}(B) \;\; \text{[from (18)]} \\ &= \frac{1}{k} \sup_{B} \sup_{\sigma_1^2,\ldots,\sigma_k^2} \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} \;\; \forall b \in \mathcal{A}_{ci}(B) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{[from Theorem III.7]} \\ &= \frac{1}{k} \sup_{B} \sup_{\sigma_1^2,\ldots,\sigma_k^2} \sum_{i=1}^{k} 4^{-u_i} \;\; \forall u \in \Delta \; \text{[from the definition of } \Delta\text{]} \\ &= \frac{1}{k} \sup_{B} \sup_{u \in M_B} f(u) \qquad\qquad \text{[from the definition of } M_B\text{]} \\ &= \frac{1}{k} \sup_{u \in V_{k-1}(0)} f(u) \qquad\qquad \text{[from Lemma III.1]} \\ &= \frac{1}{k} \max_{0 \leq j \leq k-1} \sum_{i=1}^{k} 4^{-c(k-1,j)_i} \qquad\qquad (23) \\ &= \max_{0 \leq j \leq k-1} 4^{j/k}\left(1 - \frac{3}{4k}j\right) \qquad\qquad \text{[from (3)]} \end{aligned}$$

where (23) follows from the fact that the convex function f, restricted to the closed and bounded polytope $V_{k-1}(0)$, achieves a global maximum (e.g., see [28, Theorem 6.12 on p. 154]) on the polytope's set of vertices, which consists of all coordinate permutations [8, pp. 461–462] of

$$c(k-1,0),\ldots,c(k-1,k-1).$$

For $j = 0,\ldots,k-1$, define

$$g(j) = 4^{j/k}\left(1 - \frac{3}{4k}j\right).$$

Since $g(j) > 0$ if and only if $j < 4k/3$, the function g must attain its maximum when $j < 4k/3$. In the range $0 \le j < 4k/3$, the ratio

$$\frac{g(j+1)}{g(j)} = 4^{1/k}\left(1 - \frac{1}{\frac{4k}{3} - j}\right)$$

is greater than 1 if and only if

$$j < \frac{4k}{3} - \frac{1}{1 - 4^{-1/k}}$$

so g attains is maximum when

$$j = \left\lceil \frac{4k}{3} - \frac{1}{1 - 4^{-1/k}} \right\rceil. \qquad \square$$

Appendix

Proof of Lemma II.3:

$$\begin{aligned}
&\left\{u \in \Lambda_k + y : \|w-u\| = \min_{v \in \Lambda_k + y} \|w - v\|\right\} \\
&\quad = y + \left\{u \in \Lambda_k : \|w - (u+y)\| = \min_{v \in \Lambda_k} \|w - (v+y)\|\right\} \\
&\quad = \Phi_k(w-y) + y. \qquad \square
\end{aligned}$$

Proof of Lemma III.1: First, note that for each $k \ge 1$ and for any $u \in \mathcal{H}^k$, the symmetry of Λ_k implies that

$$u \in V_k(0) \text{ if and only if } -u \in V_k(0). \tag{A1}$$

Also, note that since Λ_{k-1} consists of all vectors with k integer coordinates which sum to 0, and since

$$\frac{B}{k}(\underbrace{1,\ldots,1}_{k}) + c(k-1, B \bmod k) \in \mathcal{A}_I(B)$$

it follows that

$$\mathcal{A}_I(B) = \Lambda_{k-1} + \frac{B}{k}(\underbrace{1,\ldots,1}_{k}) + c(k-1, B \bmod k). \tag{A2}$$

Now, Lemma II.3 and (A2) imply that

$$\begin{aligned}
\mathcal{A}_{ci}(B) &= \frac{B}{k}(\underbrace{1,\ldots,1}_{k}) + c(k-1, B \bmod k) \\
&+\Phi_{k-1}\left(a_{or}(B) - \left[\frac{B}{k}(\underbrace{1,\ldots,1}_{k}) + c(k-1, B \bmod k)\right]\right).
\end{aligned}$$

Thus, Lemma II.1 gives

$$\Delta = \Phi_{k-1}(\mu) - \mu.$$

Since $\mu \in V_{k-1}\big(w\big)$ for all $w \in \Phi_{k-1}(\mu)$, we have that for each $w \in \Phi_{k-1}(\mu)$

$$\|\mu - w\| \le \|\mu - y\| \quad \forall y \in \Lambda_{k-1} \tag{A3}$$

$$= \|(\mu - w) - (y - w)\| \quad \forall y \in \Lambda_{k-1}. \tag{A4}$$

Since $w \in \Lambda_{k-1}$, we have $y - w \in \Lambda_{k-1}$ for all $y \in \Lambda_{k-1}$. Thus, by (A1) and the definition of $V_{k-1}(0)$, (A3)–(A4) imply $w - \mu \in V_{k-1}(0)$. Hence, $\Delta \subset V_{k-1}(0)$, and therefore,

$$M_B \subset V_{k-1}(0).$$

Now, for any $v \in V_{k-1}(0)$ and for arbitrary $a > 0$, setting

$$\sigma_i^2 = a4^{c(k-1,B \bmod k)_i - v_i}$$

for $1 \le i \le k$ results in

$$\begin{aligned}
g &= \left(\prod_{i=1}^{k} \sigma_i^2\right)^{1/k} = a \\
\frac{1}{2}&\left(\log_2 \frac{\sigma_1^2}{g},\ldots,\log_2 \frac{\sigma_k^2}{g}\right) = c(k-1, B \bmod k) - v
\end{aligned}$$

and therefore

$$\Delta = \Phi_{k-1}(-v) + v.$$

Since $v \in V_{k-1}(0)$, by (A1), we also have $-v \in V_{k-1}(0)$. Hence, $0 \in \Phi_{k-1}(-v)$, and thus $v \in \Delta$. So,

$$V_{k-1}(0) \subset M_B$$

and therefore, $M_B = V_{k-1}(0)$. $\square$

Proof of Lemma III.2: From (3) we have

$$\begin{aligned}
&\{c(k-1, B \bmod k) : B \ge 1\} \\
&\qquad = \{c(k-1, B \bmod k) : 1 \le B \le k\}.
\end{aligned}$$

Lemmas II.3 and III.1 imply that for each B, any element of Δ is the difference between the vector

$$(1/2)\big(\log_2(\sigma_1^2/g),\ldots,\log_2(\sigma_k^2/g)\big)$$

and a point (not necessarily unique) closest to it from the set

$$\Lambda_{k-1} + c(k-1, B \bmod k).$$

Hence,

$$\bigcup_{B\geq 1} \Delta \subset \bigcup_{1\leq B\leq k} \Delta$$

so, in fact, these two unions are equal. The result then follows from the fact that for each B, the set Δ is finite. □

Proof of Lemma III.4: Let t and $\mathcal{I}_k$ be defined as in Theorem III.3 and let

$$b = \beta + a_{or}(B) \in \mathcal{A}_{ci}(B).$$

Then for all i

$$\begin{aligned}
&\beta_i = b_i - r(a_{or}(B)_i) - \rho(a_{or}(B)_i) \\
&\qquad\qquad \text{[from the definitions of } \rho \text{ and } r\text{]} \qquad \text{(A5)}\\
&-\frac{1}{2} \leq \rho(a_{or}(B)_i) < \frac{1}{2} \\
&\qquad\qquad \text{[from the definitions of } \rho \text{ and } r\text{]} \qquad \text{(A6)}\\
&b_i - r(a_{or}(B)_i) \in \begin{cases} \{0,-1\}, & \text{if } t>0 \\ \{0,1\}, & \text{if } t<0 \end{cases}\\
&\qquad\qquad \text{[from Theorem III.3].} \qquad \text{(A7)}
\end{aligned}$$

Since $\beta_j \in (-1/2, 1/2]$, we have

$$\begin{aligned}
&-\frac{1}{2} < b_j - r(a_{or}(B)_j) - \rho(a_{or}(B)_j) \leq \frac{1}{2} \quad \text{[from (A5)]} \qquad \text{(A8)}\\
&b_j - r(a_{or}(B)_j) = 0 \qquad \text{[from (A6), (A7), (A8)]} \qquad \text{(A9)}\\
&\beta_j = -\rho(a_{or}(B)_j) \qquad \text{[from (A5), (A9)].}
\end{aligned}$$

Suppose $\beta_i < -1/2$. Then

$$\begin{aligned}
b_i - r(a_{or}(B)_i) &= \beta_i + \rho(a_{or}(B)_i) && \text{[from (A5)]}\\
&< -\frac{1}{2} + \rho(a_{or}(B)_i)\\
&< 0 && \text{[from (A6)]} \qquad \text{(A10)}\\
b_i - r(a_{or}(B)_i) &= -1 && \text{[from (A7), (A10)]} \qquad \text{(A11)}\\
\beta_i &= -\rho(a_{or}(B)_i) - 1 && \text{[from (A5), (A11)].}
\end{aligned}$$

By (A9), (A11), the fact that $b \in \mathcal{A}_{ci}(B)$, and Theorem III.3, there exists $(i_1, \ldots, i_k) \in \mathcal{I}_k$ such that

$$\begin{aligned}
i &\in \{i_1, \ldots, i_t\}\\
j &\in \{i_{t+1}, \ldots, i_k\}\\
\rho(a_{or}(B)_i) &\leq \rho(a_{or}(B)_j).
\end{aligned}$$

Suppose $\beta_i > 1/2$. Then

$$\begin{aligned}
b_i - r(a_{or}(B)_i) &= \beta_i + \rho(a_{or}(B)_i) && \text{[from (A5)]}\\
&> \frac{1}{2} + \rho(a_{or}(B)_i)\\
&> 0 && \text{[from (A6)]} \qquad \text{(A12)}\\
b_i - r(a_{or}(B)_i) &= 1 && \text{[from (A7), (A12)]} \qquad \text{(A13)}\\
\beta_i &= -\rho(a_{or}(B)_i) + 1 && \text{[from (A5), (A13)].}
\end{aligned}$$

By (A9), (A13), the fact that $b \in \mathcal{A}_{ci}(B)$, and Theorem III.3, there exists $(i_1, \ldots, i_k) \in \mathcal{I}_k$ such that

$$\begin{aligned}
i &\in \{i_{k+t+1}, \ldots, i_k\}\\
j &\in \{i_1, \ldots, i_{k+t}\}\\
\rho(a_{or}(B)_i) &\geq \rho(a_{or}(B)_j).
\end{aligned}$$

□

Proof of Lemma III.5: Let $\hat{b}$ and $\mathcal{I}_k$ be defined as in Theorem III.3. If $t = 0$, then the result follows from Theorem III.3 and the definitions of Δ and $r(\cdot)$.

Suppose $t > 0$ and let

$$b = \beta + a_{or}(B) \in \mathcal{A}_{ci}(B).$$

By Theorem III.3, there exists $(i_1, \ldots, i_k) \in \mathcal{I}_k$ such that

$$b_j = \begin{cases} \hat{b}_j - 1, & \text{if } j \in \{i_1, \ldots, i_t\} \\ \hat{b}_j, & \text{if } j \in \{i_{t+1}, \ldots, i_k\}. \end{cases} \tag{A14}$$

Subtracting $a_{or}(B)$ from both sides of (A14) gives

$$\begin{aligned}
\beta_j &= \begin{cases} \hat{b}_j - 1 - a_{or}(B)_j, & \text{if } j \in \{i_1, \ldots, i_t\} \\ \hat{b}_j - a_{or}(B)_j, & \text{if } j \in \{i_{t+1}, \ldots, i_k\} \end{cases}\\
&= \begin{cases} -\rho(a_{or}(B)_j) - 1, & \text{if } j \in \{i_1, \ldots, i_t\} \\ -\rho(a_{or}(B)_j), & \text{if } j \in \{i_{t+1}, \ldots, i_k\}. \end{cases}
\end{aligned}$$

Since

$$-1/2 \leq \rho(a_{or}(B)_j) < 1/2$$

we have

$$-\rho(a_{or}(B)_j) \in (-1/2, 1/2] \subset (-1, 1/2].$$

Thus, it suffices to show that

$$\rho(a_{or}(B)_j) < 0$$

for $j \in \{i_1, \ldots, i_t\}$, since then

$$-\rho(a_{or}(B)_j) - 1 \in (-1, -1/2].$$

Let n denote the number of components of $a_{or}(B)$ such that $\rho(a_{or}(B)_j) < 0$. Since the subscripts i_j are ordered by increasing value of $\rho(a_{or}(B)_j)$, we have $\rho(a_{or}(B)_j) < 0$ for $j \in \{i_1, \ldots, i_n\}$. Hence, it suffices to show that $t \leq n$. We have

$$\begin{aligned}
t &= \left(\sum_{i=1}^{k} r(a_{or}(B)_i)\right) - B\\
&= n + \left(\sum_{i=1}^{k} \lfloor a_{or}(B)_i \rfloor\right) - B\\
&= n - \sum_{i=1}^{k} (a_{or}(B)_i - \lfloor a_{or}(B)_i \rfloor)\\
&\leq n.
\end{aligned}$$

The result then follows by symmetry for $t < 0$. □

Proof of Lemma III.6: Since $\beta \in \Delta$, Lemma III.5 gives $\beta_i, \beta_j \in (-1,1)$. It is easy to verify that

$$\beta_j - \beta_i \le 1$$

in the following three cases:

- $\beta_i, \beta_j \in [0,1)$,
- $\beta_i \in (-1,1), \beta_j \in (-1,0]$,
- $\beta_i \in [-1/2,0], \beta_j \in [0,1/2]$.

The inequality also holds for

$$\beta_i \in (-1,-1/2) \quad \text{and} \quad \beta_j \in [0,1/2]$$

since

$$\beta_j - 1 = -\rho(a_{or}(B)_j) - 1 \le -\rho(a_{or}(B)_i) - 1 = \beta_i \qquad \text{[from Lemma III.4]}$$

and it holds for

$$\beta_i \in (-1/2,0] \quad \text{and} \quad \beta_j \in (1/2,1)$$

since

$$\beta_j - 1 = -\rho(a_{or}(B)_j) \le -\rho(a_{or}(B)_i) = \beta_i \qquad \text{[from Lemma III.4].}$$

Finally, Lemma III.5 implies that it cannot be the case that $\beta_i \in (-1,-1/2]$ and $\beta_j \in (1/2,1)$. Thus, $\beta_j - \beta_i \le 1$ for all i and j.

Let

$$\tilde{b} = b + \omega(i,j)$$

and suppose $\beta_j - \beta_i = 1$. Then

$$\begin{aligned}\tilde{b}_i &= b_i + 1 = \beta_i + 1 + a_{or}(B)_i = \beta_j + a_{or}(B)_i\\ \tilde{b}_j &= b_j - 1 = \beta_j - 1 + a_{or}(B)_j = \beta_i + a_{or}(B)_j.\end{aligned}$$

Hence,

$$\tilde{b}_l - a_{or}(B)_l = \begin{cases}\beta_j, & \text{if } l = i\\ \beta_i, & \text{if } l = j\\ \beta_l, & \text{otherwise .}\end{cases}$$

Therefore,

$$\|\tilde{b} - a_{or}(B)\| = \|\beta\|$$

which, by the definition of Δ, implies $\tilde{b} \in \mathcal{A}_{ci}(B)$. □

Proof of Lemma III.8: Assume

$$Q(W,f) \subset V \subset W.$$

If $b \in Q(W,f)$, then

$$\begin{aligned}f(b) &= \min_{\hat{b}\in W} f(\hat{b}) && \text{[from } b \in Q(W,f)\text{]}\\ &\le \min_{\hat{b}\in V} f(\hat{b}) && \text{[from } V \subset W\text{]}\\ &\le f(b) && \text{[from } b \in V\text{]}\end{aligned}$$

and therefore, $b \in Q(V,f)$. Thus,

$$Q(W,f) \subset Q(V,f).$$

If $b \in Q(V,f)$, then

$$\begin{aligned}f(b) &= \min_{\hat{b}\in V} f(\hat{b}) && \text{[from } b \in Q(V,f)\text{]}\\ &\le \min_{\hat{b}\in Q(W,f)} f(\hat{b}) && \text{[from } Q(W,f) \subset V\text{]}\\ &= \min_{\hat{b}\in W} f(\hat{b}) && \text{[from the definition of } Q(W,f)\text{]}\\ &\le f(b) && \text{[from } b \in V \subset W\text{]}\end{aligned}$$

and therefore, $b \in Q(W,f)$. Thus,

$$Q(V,f) \subset Q(W,f). \qquad \square$$

Proof of Lemma III.9: Suppose

$$b \in Q(\mathcal{A}_I(B), \theta_1).$$

For any i and j, the following identity holds:

$$\begin{aligned}&[(b_i - 1) - a_{or}(B,S)_i]^2 + [(b_j + 1) - a_{or}(B,S)_j]^2\\ &\quad - [b_i - a_{or}(B,S)_i]^2 - [b_j - a_{or}(B,S)_j]^2\\ &= 2[1 + a_{or}(B,S)_i - b_i + b_j - a_{or}(B,S)_j].\end{aligned} \tag{A15}$$

Now, suppose there exists an i such that

$$b_i - a_{or}(B,S)_i \ge 1.$$

Then there must exist a j such that

$$b_j - a_{or}(B,S)_j < 0$$

since

$$\sum_l b_l = \sum_l a_{or}(B,S)_l = B.$$

But then the right-hand side of (A15) would be negative which would imply

$$b \notin Q(\mathcal{A}_I(B), \theta_1)$$

since subtracting 1 from b_i and adding 1 to b_j would result in an integer bit allocation closer than b to $a_{or}(B,S)$. A similar contradiction results in the case where

$$b_i - a_{or}(B,S)_i \le -1.$$

Thus, for every i, we must have

$$b_i \in \{\lfloor a_{or}(B,S)_i \rfloor, \lceil a_{or}(B,S)_i \rceil\}.$$

The definition of $a_{or}(B,S)$ then implies $b_i = 0$ for all $i \notin S$. Thus, $b \in \mathcal{A}_I(B,S)$, and therefore,

$$Q(\mathcal{A}_I(B), \theta_1) \subset \mathcal{A}_I(B,S).$$

Now applying Lemma III.8 with

$$\begin{aligned}W &= \mathcal{A}_I(B)\\ V &= \mathcal{A}_I(B,S)\\ f &= \theta_1\end{aligned}$$

gives

$$Q(\mathcal{A}_I(B), \theta_1) = \mathcal{A}_{ci}(B, S). \qquad \square$$

Proof of Lemma III.10: Consider a set of sources $\hat{X}_1, \ldots, \hat{X}_k$ with variances $\hat{\sigma}_1^1, \ldots, \hat{\sigma}_k^2$ given by

$$\hat{\sigma}_i^2 = \begin{cases} \frac{\sigma_i^2}{g(S)} 4^{\frac{B(k-|S|)}{k|S|}}, & \text{if } i \in S \\ 4^{-B/k}, & \text{if } i \notin S \,. \end{cases}$$

The geometric mean of the variances is

$$\begin{aligned} \left(\prod_{i=1}^{k} \hat{\sigma}_i^2\right)^{1/k} &= \left(\prod_{i\in S} \frac{\sigma_i^2}{g(S)} \cdot 4^{\frac{B(k-|S|)}{k|S|}}\right)^{1/k} \left(\prod_{i\notin S} 4^{-B/k}\right)^{1/k} \\ &= \left(\frac{\prod_{i\in S} \sigma_i^2}{g(S)^{|S|}} \cdot 4^{\frac{B(k-|S|)}{k}}\right)^{1/k} \left(4^{\frac{-B(k-|S|)}{k}}\right)^{1/k} \\ &= \left(\frac{g(S)^{|S|}}{g(S)^{|S|}} \cdot 4^{\frac{B(k-|S|)}{k}}\right)^{1/k} \left(4^{\frac{-B(k-|S|)}{k}}\right)^{1/k} \\ &\qquad \left[\text{from } g(\ S) = \left(\prod_{i\in S} \ \sigma_i^2\right)^{1/|S|}\right] \\ &= 1. \end{aligned}$$

Therefore, substituting the variances and their geometric mean into Lemma II.1 gives

$$\begin{aligned} a_{or}(B)_i &= \frac{B}{k} + \frac{1}{2}\log_2 \frac{\hat{\sigma}_i^2}{1} \\ &= \begin{cases} \frac{B}{|S|} + \frac{1}{2}\log_2 \frac{\sigma_i^2}{g(S)}, & \text{if } i \in S \\ 0, & \text{otherwise} \end{cases} \\ &= a_{or}(B, S)_i. \end{aligned}$$

Hence, $a_{or}(B, S)$ is the optimal real-valued bit allocation for $\hat{X}_1, \ldots, \hat{X}_k$. Thus, by Lemma III.9, $\mathcal{A}_{ci}(B, S)$ is the set of closest integer bit allocations for $\hat{X}_1, \ldots, \hat{X}_k$ (regardless of S). Let

$$t = |(r(a_{or}(B, S)_1), \ldots, r(a_{or}(B, S)_k))| - B$$

and for $b \in \mathcal{A}_{ci}(B, S)$, let

$$\beta = b - a_{or}(B, S) \in \mathcal{A}_{ci}(B, S) - a_{or}(B, S).$$

Then Lemma III.5 implies that for all i

$$\beta_i \in \begin{cases} (-1/2, 1/2], & \text{if } t = 0 \\ (-1, 1/2], & \text{if } t > 0 \\ (-1/2, 1), & \text{if } t < 0. \end{cases} \tag{A16}$$

Combining the fact that $a_{or}(B, S)_i \geq 0$ for all i with (A16) gives $b_i \geq 0$ for all i. $\square$

Proof of Lemma III.11: Let

$$m = \left(\frac{B(k-|S|)}{k|S|} + \frac{1}{2}\log_2 \frac{g}{g(S)}\right).$$

Then, for all $i \in S$

$$a_{or}(B)_i + m = a_{or}(B, S)_i. \tag{A17}$$

Suppose

$$\mathcal{A}_{ci}^{+}(B) \subset \mathcal{A}_I(B, S).$$

Since every vector in $\mathcal{A}_{ci}^{+}(B)$ is nonnegative, we have

$$\begin{aligned} \mathcal{A}_{ci}^{+}(B) &\subset \mathcal{A}_I^{+}(B, S) \\ &\subset \mathcal{A}_I^{+}(B). \end{aligned} \tag{A18}$$

From (A18), we can apply Lemma III.8 with

$$\begin{aligned} W &= \mathcal{A}_I^{+}(B) \\ V &= \mathcal{A}_I^{+}(B, S) \\ f(b) &= \|b - a_{or}(B)\| \end{aligned}$$

to obtain

$$\mathcal{A}_{ci}^{+}(B) = Q\big(\mathcal{A}_I^{+}(B, S), \|b - a_{or}(B)\|\big). \tag{A19}$$

For any $b \in \mathcal{A}_I^{+}(B, S)$

$$\begin{aligned} &\|b - a_{or}(B)\|^2 \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + \sum_{i\in S} |b_i - a_{or}(B)_i|^2 \\ &\qquad [\text{from } b_i = 0 \ \ \forall\, i \notin S] \qquad \text{(A20)} \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + \sum_{i\in S} |m + b_i - (m + a_{or}(B)_i)|^2 \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + |S| \cdot m^2 \\ &\quad + \sum_{i\in S} 2m\,(b_i - a_{or}(B, S)_i) + (b_i - a_{or}(B, S)_i)^2 \\ &\qquad [\text{from (A17)}] \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + |S| \cdot m^2 + \sum_{i\in S} (b_i - a_{or}(B, S)_i)^2 \\ &\qquad \left[\text{from } \sum_{i\in S} \ b_i = \sum_{i\in S} \ a_{or}(\ B,\ S)_i = B\right] \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + |S| \cdot m^2 + \sum_{i=1}^{k} |b_i - a_{or}(B, S)_i|^2 \\ &\qquad [\text{from } b_i = a_{or}(\ B,\ S)_i = 0 \ \ \forall\, i \notin S] \\ &= \sum_{i\notin S} |a_{or}(B)_i|^2 + |S| \cdot m^2 + \|b - a_{or}(B, S)\|^2 \,. \qquad \text{(A21)} \end{aligned}$$

Equations (A20)–(A21) show that the quantities

$$\|b - a_{or}(B)\|$$

and

$$\|b - a_{or}(B, S)\|$$

differ by a constant which is independent of b. Hence, among all bit allocations in $\mathcal{A}_I^{+}(B, S)$, we see that b is of minimal distance from $a_{or}(B, S)$ if and only if b is of minimal distance from $a_{or}(B)$, i.e.,

$$\mathcal{A}_{ci}^{+}(B, S) = Q\big(\mathcal{A}_I^{+}(B, S), \|b - a_{or}(B)\|\big).$$

Thus, by (A19), we have

$$\mathcal{A}^{+}_{ci}(B) = \mathcal{A}^{+}_{ci}(B,S).$$

Now, to show the second part of the lemma, suppose

$$\mathcal{A}^{+}_{oi}(B) \subset \mathcal{A}_{I}(B,S).$$

Since every vector in $\mathcal{A}^{+}_{oi}(B)$ is nonnegative, we have

$$\begin{aligned}\mathcal{A}^{+}_{oi}(B) &\subset \mathcal{A}^{+}_{I}(B,S)\\ &\subset \mathcal{A}^{+}_{I}(B).\end{aligned}\tag{A22}$$

From (A22), we can apply Lemma III.8 with

$$\begin{aligned}W &= \mathcal{A}^{+}_{I}(B)\\ V &= \mathcal{A}^{+}_{I}(B,S)\\ f(b) &= d\end{aligned}$$

to obtain

$$\mathcal{A}^{+}_{oi}(B) = Q\left(\mathcal{A}^{+}_{I}(B,S), d\right).\tag{A23}$$

For any $b \in \mathcal{A}^{+}_{I}(B,S)$

$$\begin{aligned}
d &= hg4^{-B/k} \cdot \sum_{i=1}^{k} 4^{(a_{or}(B)-b)_i} && \text{[from Lemma II.2]}\\
&= hg4^{-B/k} \cdot \sum_{i\notin S} 4^{a_{or}(B)_i} + hg4^{-B/k} \cdot \sum_{i\in S} 4^{(a_{or}(B)-b)_i}\\
& && \text{[from } b_i = 0 \ \forall\, i \notin S\text{]}\\
&= hg4^{-B/k} \cdot \sum_{i\notin S} 4^{a_{or}(B)_i}\\
&\quad + hg4^{-B/k} \cdot \sum_{i\in S} 4^{-m+(m+a_{or}(B)_i)-b_i}\\
&= hg4^{-B/k} \cdot \sum_{i\notin S} 4^{a_{or}(B)_i}\\
&\quad + hg4^{(-B/k)-m} \cdot \sum_{i\in S} 4^{(a_{or}(B,S)-b)_i} && \text{[from (A17)]}\\
&= hg4^{-B/k} \cdot \sum_{i\notin S} 4^{a_{or}(B)_i}\\
&\quad + hg4^{(-B/k)-m} \cdot \left(\sum_{i\in S} 4^{(a_{or}(B,S)-b)_i}\right.\\
&\quad \left. + \sum_{i\notin S} 4^{(a_{or}(B,S)-b)_i} - \sum_{i\notin S} 4^{(a_{or}(B,S)-b)_i}\right)\\
&= hg4^{-B/k} \cdot \sum_{i\notin S} \left(4^{a_{or}(B)_i} - 4^{-m}\right) + 4^{-m} \cdot \theta_2(b)\\
& && \text{[from } a_{or}(B,\ S)_i = b_i = 0 \ \forall\, i \notin S\text{]}
\end{aligned}$$

which is an affine function of $\theta_2(b)$, with coefficients which are independent of b. Therefore, among all bit allocations in $\mathcal{A}^{+}_{I}(B,S)$, we see that b minimizes $\theta_2(b)$ if and only if b minimizes d, i.e.,

$$\mathcal{A}^{+}_{oi}(B,S) = Q\big(\mathcal{A}^{+}_{I}(B,S), d\big).$$

Thus, by (A23), we have

$$\mathcal{A}^{+}_{oi}(B) = \mathcal{A}^{+}_{oi}(B,S). \qquad \square$$

Proof of Lemma III.12: We prove that for all $b \in \mathcal{A}_I(B,S)$, if $b_i > 0$, then

$$b \notin \mathcal{A}^{+}_{ci}(B) \cup \mathcal{A}^{+}_{oi}(B).$$

Let $b \in \mathcal{A}_I(B,S)$ satisfy $b_i > 0$, where $i \in S$ and $a_{or}(B,S)_i < 0$. By Lemma III.11, we know that

$$\mathcal{A}^{+}_{ci}(B) = \mathcal{A}^{+}_{ci}(B,S)$$

and

$$\mathcal{A}^{+}_{oi}(B) = \mathcal{A}^{+}_{oi}(B,S).$$

We will show that $b \notin \mathcal{A}^{+}_{ci}(B,S)$ and $b \notin \mathcal{A}^{+}_{oi}(B,S)$. In particular, we demonstrate that there exists $j \in S$ such that adding 1 bit to b_j and subtracting 1 bit from b_i reduces both $\theta_1(b)$ and $\theta_2(b)$, i.e., the original b chosen could not have been an optimal nor a closest nonnegative integer bit allocation.

Suppose

$$a_{or}(B,S)_i - b_i \geq a_{or}(B,S)_l - b_l - 1, \qquad \forall l \in S.\tag{A24}$$

Then we get

$$\begin{aligned}
(|S|&-1)(a_{or}(B,S)_i - b_i)\\
&= \sum_{\substack{l\in S\\ l\neq i}} a_{or}(B,S)_i - b_i\\
&\geq \sum_{\substack{l\in S\\ l\neq i}} (a_{or}(B,S)_l - b_l - 1) && \text{[from (A24)]}\\
&= (B - a_{or}(B,S)_i) - (B - b_i) - (|S|-1)\\
& && \text{[from } \textstyle\sum_{l\in S} b_l = \sum_{l\in S} a_{or}(B,\ S)_l = B\text{]}\\
&= -(a_{or}(B,S)_i - b_i) - (|S|-1)
\end{aligned}$$

which implies

$$|S|(1 - b_i + a_{or}(B,S)_i) \geq 1,$$

a contradiction, since $b_i \geq 1$ and $a_{or}(B,S)_i < 0$. Therefore, (A24) is false, so there exists $j \in S$ such that $j \neq i$ and

$$a_{or}(B,S)_i - b_i < a_{or}(B,S)_j - b_j - 1.\tag{A25}$$

Multiplying each side of (A25) by -2 and adding

$$(a_{or}(B,S)_i - b_i)^2 + (a_{or}(B,S)_j - b_j)^2$$

to each side gives

$$\begin{aligned}
&(a_{or}(B,S)_i - b_i)^2 - 2(a_{or}(B,S)_i - b_i)\\
&\quad + (a_{or}(B,S)_j - b_j)^2\\
&> (a_{or}(B,S)_i - b_i)^2 - 2(a_{or}(B,S)_j - b_j) + 2\\
&\quad + (a_{or}(B,S)_j - b_j)^2
\end{aligned}$$

or equivalently

$$\begin{aligned}
(a_{or}(B,S)_i &- b_i)^2 + (a_{or}(B,S)_j - b_j)^2\\
&> (a_{or}(B,S)_i - b_i)^2 + 2(a_{or}(B,S)_i - b_i) + 1\\
&\quad + (a_{or}(B,S)_j - b_j)^2 - 2(a_{or}(B,S)_j - b_j) + 1\\
&= (a_{or}(B,S)_i - (b_i - 1))^2 + (a_{or}(B,S)_j - (b_j + 1))^2.
\end{aligned}$$

Thus, subtracting 1 bit from b_i and adding 1 bit to b_j reduces $\theta_1(b)$. Some algebra shows that the inequality in (A25) is equivalent to

$$4^{a_{or}(B,S)_i - b_i} + 4^{a_{or}(B,S)_j - b_j} > 4^{a_{or}(B,S)_i - (b_i - 1)} + 4^{a_{or}(B,S)_j - (b_j + 1)}$$

from which it follows that $\theta_2(b)$ can be reduced by adding 1 bit to b_j and subtracting 1 bit from b_i. □

Proof of Lemma IV.1: The proof is trivial for $||w|| = 0$, so assume $||w|| > 0$. We determine the extrema of

$$\sum_{i=1}^{k} 4^{-w_i} \tag{A26}$$

subject to the constraints

$$\sum_{i=1}^{k} w_i = 0 \tag{A27}$$

$$\sum_{i=1}^{k} w_i^2 = a^2. \tag{A28}$$

Define a Lagrangian J associated with multipliers λ_1 and λ_2 by

$$J = \sum_{i=1}^{k} 4^{-w_i} + \lambda_1 \sum_{i=1}^{k} w_i + \lambda_2 \left(\sum_{i=1}^{k} w_i^2 - a^2 \right).$$

The extrema of J must satisfy (for $1 \le i \le k$)

$$0 = \frac{\partial J}{\partial w_i} = -(\ln 4) 4^{-w_i} + \lambda_1 + 2\lambda_2 w_i. \tag{A29}$$

Suppose $\lambda_2 > 0$. Then $\frac{\partial J}{\partial w_i}$ is monotone increasing in w_i and approaches $\pm\infty$ as $w_i \longrightarrow \pm\infty$. Thus, exactly one w_i satisfies (A29) for each i, and therefore $w_i = w_j$ for all i, j. So, by (A27), it follows that $w_i = 0$ for all i, contradicting $||w|| > 0$.

Thus, we can assume $\lambda_2 < 0$. Since $\frac{\partial J}{\partial w_i}$ is strictly concave, (A29) can have at most two solutions. It cannot be the case that (A29) has only one solution, for otherwise (A27) would again imply that $w_i = 0$ for all i, contradicting $||w|| > 0$. So (A29) has exactly two solutions and by (A27) these two solutions must be of different signs.

Thus, the extrema of J must lie in the set

$$P = P_1 \cup \cdots \cup P_{k-1}$$

where P_j is the set of all $\binom{k}{j}$ component-wise permutations of the vector

$$a \left(\frac{j}{k(k-j)} \right)^{1/2} \cdot \left(\underbrace{-\left(\frac{k-j}{j}\right), \ldots, -\left(\frac{k-j}{j}\right)}_{j} \underbrace{1, \ldots, 1}_{k-j} \right). \tag{A30}$$

The constant factor in (A30) ensures that the elements of P satisfy (A27) and (A28).

Summing both sides of (A29) over i and solving for λ_1 yields

$$\lambda_1 = \frac{\ln 4}{k} \sum_{i=1}^{k} 4^{-w_i}. \tag{A31}$$

From (A29), we obtain

$$(\ln 4) 4^{-w_i} - \lambda_1 = 2\lambda_2 w_i$$

which when squared, summed over i, and simplified using (A28) and (A31) gives

$$\lambda_2 = -\frac{\ln 2}{a} \left[\sum_{i=1}^{k} 16^{-w_i} - \frac{1}{k} \left(\sum_{i=1}^{k} 4^{-w_i} \right)^2 \right]^{1/2}. \tag{A32}$$

Now, for any component w_i of any $w \in P_j$, using (A30), (A31), and (A32) gives

$$\begin{aligned}
&-(\ln 4) 4^{-w_i} + \lambda_1 + 2\lambda_2 w_i \\
&= -(\ln 4) 4^{-w_i} + \frac{\ln 4}{k} \sum_{i=1}^{k} 4^{-w_i} \\
&\quad - \frac{2 w_i \ln 2}{a} \left[\sum_{i=1}^{k} 16^{-w_i} - \frac{1}{k} \left(\sum_{i=1}^{k} 4^{-w_i} \right)^2 \right]^{1/2} \\
&= -(\ln 4) 4^{-w_i} \\
&\quad + \frac{\ln 4}{k} \left(j 4^{a\sqrt{(k-j)/(kj)}} + (k-j) 4^{-a\sqrt{j/(k(k-j))}} \right) \\
&\quad - \frac{w_i \ln 4}{a} \Big[j 16^{a\sqrt{(k-j)/(kj)}} + (k-j) 16^{-a\sqrt{j/(k(k-j))}} \\
&\quad - \frac{1}{k} \left(j 4^{a\sqrt{(k-j)/(kj)}} + (k-j) 4^{-a\sqrt{j/(k(k-j))}} \right)^2 \Big]^{1/2} \\
&= -(\ln 4) 4^{-w_i} \\
&\quad + \frac{\ln 4}{k} \left(j 4^{a\sqrt{(k-j)/(kj)}} + (k-j) 4^{-a\sqrt{j/(k(k-j))}} \right) \\
&\quad - \frac{w_i \ln 4}{a} \sqrt{\frac{j(k-j)}{k}} \left[4^{a\sqrt{(k-j)/(kj)}} - 4^{-a\sqrt{j/(k(k-j))}} \right] \\
&= 0
\end{aligned} \tag{A33}$$

where (A33) follows by considering the cases

$$w_i = -a\sqrt{(k-j)/(kj)}$$

and

$$w_i = a\sqrt{j/(k(k-j))}.$$

Hence, every $w \in P$ satisfies (A29), and therefore, P is the set of solutions to (A29) subject to the constraints in (A27) and (A28).

Substituting an arbitrary element $w \in P$ (i.e., an extremum of J) into (A26) gives

$$\begin{aligned}
\sum_{i=1}^{k} 4^{-w_i} &= j 4^{a\sqrt{(k-j)/(kj)}} + (k-j) 4^{-a\sqrt{j/(k(k-j))}} \\
&= j 4^{||w||\sqrt{(k-j)/(kj)}} + (k-j) 4^{-||w||\sqrt{j/(k(k-j))}} \\
&\qquad [\text{ from (A28)}].
\end{aligned} \tag{A34}$$

To complete the proof it suffices to show that (A34) is decreasing in j. This implies (A26) is upper-bounded by (A34) when $j=1$ and lower-bounded by (A34) when $j=k-1$.

Note that if the right-hand side of (A34) is viewed as a continuous function of j, then its derivative with respect to j is

$$4^{\|w\|\sqrt{(k-j)/(kj)}}\left[1-\|w\|\,(\ln 2)\left(\frac{k}{j(k-j)}\right)^{1/2}\right]$$
$$-4^{-\|w\|\sqrt{j/(k(k-j))}}\left[1+\|w\|\,(\ln 2)\left(\frac{k}{j(k-j)}\right)^{1/2}\right]$$

which is negative if and only if

$$f\left(\|w\|\,(\ln 2)\sqrt{k/(j(k-j))}\right)>0$$

where

$$f(u)=1+u-(1-u)e^{2u}.$$

Since

$$f(0)=f'(0)=0$$

and

$$f''(u)=4ue^{2u}>0$$

for all $u>0$, we have $f(u)>0$ for all $u>0$. □

Proof of Lemma IV.2: The result follows from Lemma II.2 and Lemma IV.1 with $w=b-a_{or}(B)$. □

Proof of Lemma IV.4: First we show the result for each bit budget B and for all $b\in\mathcal{A}_{oi}(B)$. Then we prove that $\mathcal{A}_{oi}(B)=\mathcal{A}_{oi}^{+}(B)$ for each bit budget B.

For any vector u and any permutation π of the positive integers less than or equal to the dimension of u, let $\pi(u)$ denote the component-wise permutation of u according to π. First observe that

$$\Phi_k\big(\pi(\gamma_k)\big)=\{0\}$$

for any k and any permutation π of $\{1,\dots,k+1\}$. To see this, note that for any $w\in\Lambda_k\setminus\{0\}$, since $|(\gamma_k)_i|<1/2$ for all i, we have

$$\begin{aligned}|(\pi(\gamma_k))_i-w_i|&>|(\pi(\gamma_k))_i|, &&\text{if } w_i\neq 0\\ |(\pi(\gamma_k))_i-w_i|&=|(\pi(\gamma_k))_i|, &&\text{if } w_i=0\end{aligned}$$

which implies

$$\|\pi(\gamma_k)-w\|>\|\pi(\gamma_k)\|$$

and therefore,

$$\begin{aligned}\Phi_k\big(\pi(\gamma_k)\big)&=\left\{u\in\Lambda_k:\|\pi(\gamma_k)-u\|=\min_{v\in\Lambda_k}\|\pi(\gamma_k)-v\|\right\}\\&=\{0\}.\end{aligned}$$

Now observe that $\gamma_{k-1}-c(k-1,j)$ is the left-cyclic shift of γ_{k-1} by j positions, for any j, since

$$\begin{aligned}&\gamma_{k-1}-c(k-1,j)\\&\quad=\frac{1}{2k}\big(-(k-1)+2j,-(k-1)+2j+2,\dots,k-1,\\&\qquad-(k-1),\dots,-(k-1)+2j-4,-(k-1)+2j-2\big).\end{aligned}$$

In particular, for each bit budget B, Theorem III.7 and Lemma III.1 imply that for every $b\in\mathcal{A}_{oi}(B)$

$$\begin{aligned}b-a_{or}(B)&\in\Phi_{k-1}\big(\gamma_{k-1}-c(k-1,B \bmod k)\big)\\&\qquad-\big(\gamma_{k-1}-c(k-1,B \bmod k)\big)\\&=\Phi_{k-1}\big(\hat{\gamma}_{k-1}\big)-\hat{\gamma}_{k-1}\\&=\{-\hat{\gamma}_{k-1}\}\end{aligned}$$

where $\hat{\gamma}_{k-1}$ denotes γ_{k-1} left-cyclic shifted by $B \bmod k$ positions. Since the components of γ_{k-1} are the same as those of $-\gamma_{k-1}$, so are the components of $-\hat{\gamma}_{k-1}$. Thus, $-\hat{\gamma}_{k-1}$ is a permutation of γ_{k-1}. Hence, for each bit budget B and for all $b\in\mathcal{A}_{oi}(B)$, the vector $b-a_{or}(B)$ is a permutation of γ_{k-1}.

To show that $\mathcal{A}_{oi}(B)=\mathcal{A}_{oi}^{+}(B)$ for each bit budget B, it suffices to show that $\mathcal{A}_{oi}(B)\subset\mathcal{A}_{I}^{+}(B)$ for each bit budget B. Lemma II.1 and the definition of γ_k imply that for each bit budget B and for $1\le i\le k$

$$a_{or}(B)_i=\frac{-(k-1)+2(i-1)+2B}{2k}.$$

Thus, the definition of $r(\cdot)$ implies that for each bit budget B and for $1\le i\le k$

$$r\left(a_{or}(B)_i\right)=\left\lfloor\frac{1+2(i-1)+2B}{2k}\right\rfloor. \qquad \text{(A35)}$$

For each bit budget B, let

$$\hat{b}=(r(a_{or}(B)_1),\dots,r(a_{or}(B)_k))$$

as in Theorem III.3. Then

$$\begin{aligned}|\hat{b}|&=\sum_{i=1}^{k}\left\lfloor\frac{1+2(i-1)+2B}{2k}\right\rfloor &&\text{(A36)}\\&=\sum_{i=1}^{k}\left\lfloor\frac{i+B}{k}-\frac{1}{2k}\right\rfloor\\&=\sum_{i<k-(B\bmod k)+(1/2)}\left\lfloor\frac{B}{k}\right\rfloor\\&\quad+\sum_{i>k-(B\bmod k)+(1/2)}\left\lfloor\frac{B}{k}\right\rfloor+1\\&=(k-(B \bmod k))\left\lfloor\frac{B}{k}\right\rfloor+(B \bmod k)\left(\left\lfloor\frac{B}{k}\right\rfloor+1\right)\\&=B &&\text{(A37)}\end{aligned}$$

where (A36) follows from (A35). Equation (A37) and Theorem III.3 imply

$$\mathcal{A}_{ci}(B)=\{\hat{b}\}.$$

Therefore, $\mathcal{A}_{oi}(B)=\{\hat{b}\}$ by Theorem III.7. Since (A35) shows that $\hat{b}$ is nonnegative

$$\mathcal{A}_{oi}(B)\subset\mathcal{A}_{I}^{+}(B)$$

for each bit budget B. □

Acknowledgment

The authors wish to thank Bob Gray, Khalid Sayood, and Vivek Goyal for suggesting at the 2005 Data Compression Conference the nonnegative component version of the bit allocation problem. We also are grateful for the extremely detailed and careful reviews provided by both of the anonymous referees.

References

[1] P. Batra and A. Eleftheriadis, "Alternative formulations for bit allocation with dependent quantization," in *Proc. 2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, May 2002, vol. 4, pp. 3505–3508.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984, The Wadsworth Statistics/Probability Series.

[3] A. M. Bruckstein, "On 'soft' bit allocation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 5, pp. 614–617, May 1987.

[4] J. Chen and D. W. Lin, "Optimal bit allocation for coding of video signals over ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1002–1015, Aug. 1997.

[5] G. Cheung and A. Zakhor, "Bit allocation for joint source/channel coding of scalable video," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 340–356, Mar. 2000.

[6] P. A. Chou, "Applications of Information Theory to Pattern Recognition and the Design of Decision Trees and Trellises," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1988.

[7] P. A. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 299–315, Mar. 1989.

[8] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. New York: Springer-Verlag, 1999.

[9] ——, "Fast quantizing and decoding algorithms for lattice quantizers and codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 227–232, Mar. 1982.

[10] J. J. Dubnowski and R. E. Crochiere, "Variable rate coding," in *Proc. 1979 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Washington, DC, Apr. 1979, vol. 4, pp. 445–448.

[11] B. Farber and K. Zeger, "Quantization of multiple sources using integer bit allocation," in *Proc. 2005 Data Compression Conf.*, Snowbird, UT, Mar. 2005, pp. 368–377.

[12] B. Fox, "Discrete optimization via marginal analysis," *Manage. Sci.*, vol. 13, no. 3, pp. 210–216, Nov. 1966.

[13] H. Gazzah and A. K. Khandani, "Optimum non-integer rate allocation using integer programming," *Electron. Lett.*, vol. 33, no. 24, p. 2034, Nov. 1997.

[14] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic, 1991.

[15] J. D. Gibson, "Bounds on performance and dynamic bit allocation for sub-band coders," in *Proc. 1981 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, Mar. 1981, vol. 2, pp. 836–839.

[16] ——, "Notes on bit allocation in the time and frequency domains," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 6, pp. 1609–1610, Dec. 1985.

[17] L. M. Goodman, "Optimum rate allocation for encoding sets of analog messages," *Proce. IEEE*, vol. 53, no. 11, pp. 1776–1777, Nov. 1965.

[18] L. Guo and Y. Meng, "Round-up of integer bit allocation," *Electron. Lett.*, vol. 38, no. 10, pp. 466–467, May 2002.

[19] M. Honda and F. Itakura, "Bit allocation in time and frequency domains for predictive coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 3, pp. 465–473, Jun. 1984.

[20] J. J. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. COM-11, no. 9, pp. 289–296, Sep. 1963.

[21] A. E. Mohr, "Bit allocation in sub-linear time and the multiple-choice knapsack problem," in *Proc. 2002 Data Compression Conf.*, Snowbird, UT, Mar. 2002, pp. 352–361.

[22] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis-based buffered compression and fast approximations," *IEEE Trans. Image Process.*, vol. 3, no. 1, pp. 26–40, Jan. 1994.

[23] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[24] T. A. Ramstad, "Considerations on quantization and dynamic bit-allocation in subband coders," in *Proce. 1986 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, Apr. 1986, vol. 2, pp. 841–844.

[25] E. A. Riskin, "Optimal bit allocation via the generalized BFOS algorithm," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 400–402, Mar. 1991.

[26] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 162–169, Mar. 1976.

[27] Y. Sermadevi and S. S. Hemami, "Efficient bit allocation for dependent video coding," in *Proc. 2004 Data Compression Conf.*, Snowbird, UT, Mar. 2004, pp. 232–241.

[28] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[29] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.

[30] A. V. Trushkin, "Bit number distribution upon quantization of a multivariate random variable," Transl.: from Russian *Probl. Inf. Transm.*, vol. 16, no. 1, pp. 76–79, 1980.

[31] ——, "Optimal bit allocation algorithm for quantizing a random vector," Transl.: from Russian *Probl. Inf. Transm.*, vol. 17, no. 3, pp. 156–161, 1981.

[32] P. H. Westerink, J. Biemond, and D. E. Boekee, "An optimal bit allocation algorithm for sub-band coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, New York, Apr. 1988, pp. 757–760.

[33] P. A. Wintz and A. J. Kurtenbach, "Waveform error control in PCM telemetry," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 5, pp. 650–661, Sep. 1968.

[34] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1278–1288, Oct. 1986.

[35] X. Wu, "Globally optimal bit allocation," in *Proc. 1993 Data Compression Conf.*, Snowbird, UT, Mar. 1993, pp. 22–31.