

Fixed-to-Variable Length Distribution Matching

Rana Ali Amjad and Georg Böcherer

Institute for Communications Engineering

Technische Universität München, Germany

Email: raa2463@gmail.com, georg.boecherer@tum.de

Abstract—Fixed-to-variable length (f2v) matchers are used to reversibly transform an input sequence of independent and uniformly distributed bits into an output sequence of bits that are (approximately) independent and distributed according to a target distribution. The degree of approximation is measured by the informational divergence between the output distribution and the target distribution. An algorithm is developed that efficiently finds optimal f2v codes. It is shown that by encoding the input bits blockwise, the informational divergence per bit approaches zero as the block length approaches infinity. A relation to data compression by Tunstall coding is established.

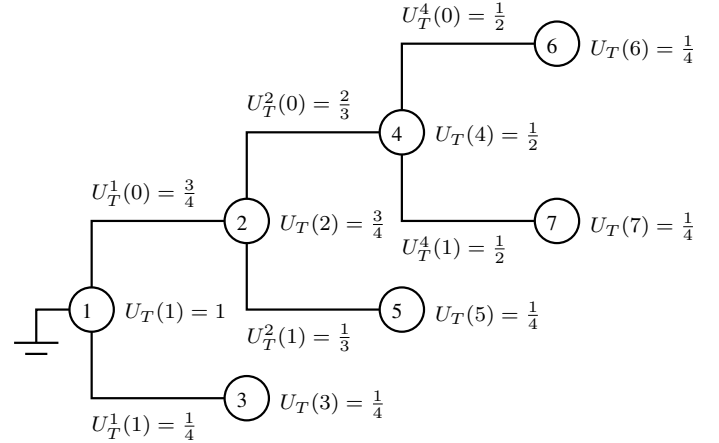
I. INTRODUCTION

Distribution matching considers the problem of mapping uniformly distributed bits to symbols that are approximately distributed according to a target distribution. In difference to the simulation of random processes [1] or the exact generation of distributions [2], distribution matching requires that the original bit sequence can be *recovered* from the generated symbol sequence. We measure the degree of approximation by the normalized informational divergence (I-divergence), which is an appropriate measure when we want to achieve channel capacity of noisy and noiseless channels [3, Sec. 3.4.3 & Chap. 6] by using a matcher. A related work is [4], [3, Chap. 3], where it is shown that *variable-to-fixed length* (v2f) matching is optimally done by geometric Huffman coding and the relation to *fixed-to-variable length* (f2v) source encoders is discussed. In the present work, we consider binary distribution matching by prefix-free f2v codes.

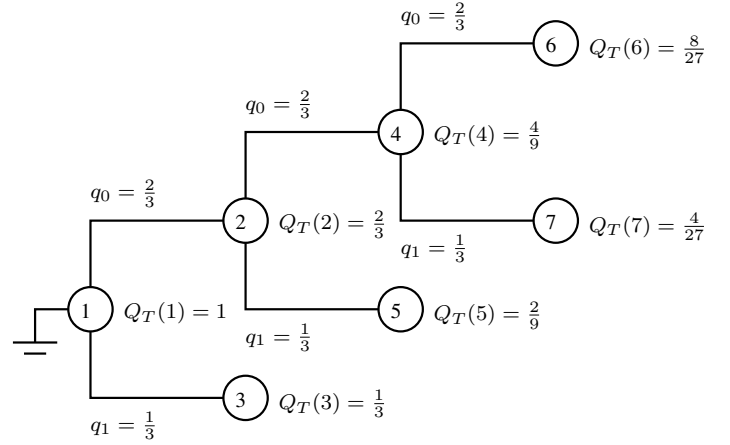
A. Rooted Trees With Probabilities

We use the framework of rooted trees with probabilities [5], [6]. Let \mathcal{T} be the set of all binary trees with 2^m leaves and consider some tree $T \in \mathcal{T}$. Index all nodes by the numbers $\mathcal{N} := \{1, 2, 3, \dots\}$ where 1 is the root. Note that there are at least $|\mathcal{N}| \geq 2^{m+1} - 1$ nodes in the tree, with equality if the tree is complete. A tree is complete if any right-infinite binary sequence starts with a path from the root to a leaf. Let $\mathcal{L} \subset \mathcal{N}$ be the set of leaf nodes and let $\mathcal{B} := \mathcal{N} \setminus \mathcal{L}$ be the set of branching nodes. Probabilities can be assigned to the tree by defining a distribution over the 2^m paths through the tree. For each $i \in \mathcal{N}$, denote by $P_T(i)$ the probability that a path is chosen that passes through node i . Since each path ends at a different leaf node, P_T defines a leaf distribution, i.e.,

This work was supported by the German Ministry of Education and Research in the framework of an Alexander von Humboldt Professorship.



(a) tree probabilities defined by a leaf distribution



(b) tree probabilities defined by branching distributions

Fig. 1. A rooted tree with probabilities. The set of branching nodes is $\mathcal{B} = \{1, 2, 4\}$ and the set of leaf nodes is $\mathcal{L} = \{3, 5, 6, 7\}$. In (a), a uniform leaf distribution U_T is chosen, i.e., for each $i \in \mathcal{L}$, $U_T(i) = \frac{1}{4}$. The leaf distribution determines the node probabilities and branching distributions. In (b), identical branching distributions are chosen, i.e., for each $i \in \mathcal{B}$, $Q_T^i = Q$, where $Q(0) =: q_0 = \frac{2}{3}$ and $Q(1) =: q_1 = \frac{1}{3}$. The branching distributions determine the resulting node probabilities, which we denote by $Q_T(i)$. Since the tree is complete, $\sum_{i \in \mathcal{L}} Q_T(i) = 1$ and Q_T defines a leaf distribution.

$\sum_{i \in \mathcal{L}} P_T(i) = 1$. For each branching node $i \in \mathcal{B}$, denote by P_T^i the branching distribution, i.e., the probabilities of branch 0 and branch 1 after passing through node i . The probabilities on the tree are completely defined either by defining the branching distributions $\{P_T^i, i \in \mathcal{B}\}$ or by defining the leaf distribution $\{P_T(i), i \in \mathcal{L}\}$. See Fig. 1 for an example.

B. v2f Source Encoding and f2v Distribution Matching

Consider a binary distribution Q with $q_1 = Q(0)$, $q_1 = Q(1)$, $0 < q_0 < 1$, and a binary tree T with 2^m leaves. Let $Q_T(i)$, $i \in \mathcal{N}$, be the node probabilities that result from having all branching distributions equal to Q , i.e. $Q_T^i = Q$ for each $i \in \mathcal{B}$. See Fig. 1(b) for an example. Let U_T be a uniform leaf distribution, i.e., $U_T(i) = 2^{-m}$ for each $i \in \mathcal{L}$, see Fig. 1(a) for an example. We use the tree as a v2f source code for a *discrete memoryless source* (DMS) Q . To guarantee lossless compression, the tree for a v2f source encoder has to be complete. Consequently, Q_T defines a leaf distribution, i.e., $\sum_{i \in \mathcal{L}} Q_T(i) = 1$. We denote the set of complete binary trees with 2^m leaves by \mathcal{C} . Each code word consists of $\log_2 2^m = m$ bits and the resulting entropy rate at the encoder output is

$$\begin{aligned} \frac{1}{m} \mathbb{H}(Q_T) &= \sum_{i \in \mathcal{L}} Q_T(i) [-\log_2 Q_T(i)] \\ &= \frac{1}{m} \sum_{i \in \mathcal{L}} Q_T(i) [-\log_2 Q_T(i) + \log_2 U_T(i) - \log_2 U_T(i)] \\ &= 1 - \frac{1}{m} \mathbb{D}(Q_T \| U_T) \end{aligned} \quad (1)$$

where $\mathbb{H}(Q_T)$ is the entropy of the leaf distribution defined by Q_T and where $\mathbb{D}(Q_T \| U_T)$ is defined accordingly. From (1), we conclude that the objective is to solve

$$\min_{T \in \mathcal{C}} \mathbb{D}(Q_T \| U_T). \quad (2)$$

The solution is known to be attained by Tunstall coding [7]. The tree in Fig. 1 is a Tunstall code for Q : $q_0 = \frac{2}{3}$, $q_1 = \frac{1}{3}$ and $m = 2$ and the corresponding v2f source encoder is

$$000 \mapsto 00, 001 \mapsto 01, 01 \mapsto 10, 1 \mapsto 11. \quad (3)$$

The dual problem is f2v distribution matching. Q is now a binary target distribution and we generate the codewords defined by the paths through a (not necessarily complete) binary tree uniformly according to U_T . For example, the f2v distribution matcher defined by the tree in Fig. 1 is

$$00 \mapsto 000, 01 \mapsto 001, 10 \mapsto 01, 11 \mapsto 1. \quad (4)$$

Denote by $\ell_i, i \in \mathcal{L}$ the path lengths and let L be a random variable that is uniformly distributed over the path lengths according to U_T . We want the I-divergence per output bit of U_T and Q_T to be small, i.e., we want to solve

$$\min_{T \in \mathcal{T}} \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)}. \quad (5)$$

In contrast to (2), the minimization is now over the set of all (not necessarily complete) binary trees with 2^m leaves. Note that although for a non-complete tree we have $\sum_{i \in \mathcal{L}} Q_T(i) < 1$, the problem (5) is well-defined, since there is always a complete tree with leaves $\mathcal{L}' \supseteq \mathcal{L}$ and $\sum_{i \in \mathcal{L}'} Q_T(i) = 1$. The sum in (5) is over the support of U_T , which is \mathcal{L} . Solving (5) is the problem that we consider in this work.

C. Outline

In Sec. II and Sec. III, we restrict attention to complete trees. We show that Tunstall coding applied to Q minimizes $\mathbb{D}(U_T \| Q_T)$ and that iteratively applying Tunstall coding to weighted versions of Q minimizes $\mathbb{D}(U_T \| Q_T) / \mathbb{E}_{U_T}(L)$. In Sec. IV we derive conditions for the optimality of complete trees and show that the I-divergence per bit can be made arbitrarily small by letting the blocklength m approach infinity. Finally, in Sec. V, we illustrate by an example that source decoders are sub-optimal distribution matchers and vice-versa, distribution dematchers are sub-optimal source encoders.

II. MINIMIZING I-DIVERGENCE

Let \mathbf{R} be the set of real numbers. For a finite set \mathcal{S} , we say that $W: \mathcal{S} \rightarrow \mathbf{R}$ is a *weighted distribution* if for each $i \in \mathcal{S}$, $W(i) > 0$. We allow for $\sum_{i \in \mathcal{S}} W(i) \neq 1$. The I-divergence of a distribution P and a weighted distribution W is

$$\mathbb{D}(P \| W) = \sum_{i \in \text{supp } P} P(i) \log_2 \frac{P(i)}{W(i)} \quad (6)$$

where supp denotes the support of P . The reason why we need this generalization of the notion of distributions and I-divergence will become clear in the next section.

Proposition 1. *Let Q be a weighted binary target distribution, and let*

$$T^* = \operatorname{argmin}_{T \in \mathcal{C}} \mathbb{D}(U_T \| Q_T) \quad (7)$$

be an optimal complete tree. Then we find that

- i. *An optimal complete tree T^* can be constructed by applying Tunstall coding to Q .*
- ii. *If $0 \leq q_0 \leq 1$ and $0 \leq q_1 \leq 1$, then T^* also minimizes $\mathbb{D}(U_T \| Q_T)$ among all possibly non-complete binary trees \mathcal{T} , i.e., the optimal tree is complete.*

Proof: Part i. We write

$$\begin{aligned} \mathbb{D}(U_T \| Q_T) &= \sum_{i \in \mathcal{L}} 2^{-m} \log_2 \frac{2^{-m}}{Q_T(i)} \\ &= -m - 2^{-m} \sum_{i \in \mathcal{L}} \log Q_T(i) \end{aligned} \quad (8)$$

and hence

$$\operatorname{argmin}_{T \in \mathcal{C}} \mathbb{D}(U_T \| Q_T) = \operatorname{argmax}_{T \in \mathcal{C}} \sum_{i \in \mathcal{L}} \log_2 Q_T(i) \quad (9)$$

Consider now an arbitrary complete tree $T \in \mathcal{C}$. Since the tree is complete, there exist (at least) two leaves that are siblings, say j and $j+1$. Denote by k the corresponding branching node. The contribution of these two leaves to the objective function on the right-hand side of (9) can be written as

$$\begin{aligned} \log_2 Q_T(j) + \log_2 Q_T(j+1) &= \log[Q_T(k)q_0] + \log[Q_T(k)q_1] \\ &= \log Q_T(k) + \log Q_T(k) + \log q_0 + \log q_1. \end{aligned} \quad (10)$$

Now consider the tree T' that results from removing the nodes j and $j+1$. The new set of leaf nodes is $\mathcal{L}' = \mathcal{L} \setminus \{j, j+1\}$ and the new set of branching nodes is $\mathcal{B}' = \mathcal{B} \setminus k$. Also Q_T defines a weighted leaf distribution on \mathcal{L}' . The same procedure can be applied repeatedly by defining $T = T'$, until T' consists only of the root node. We use this idea to re-write the objective function of the right-hand side of (9) as follows.

$$\begin{aligned} \sum_{i \in \mathcal{L}} \log_2 Q_T(i) &= \sum_{i \in \mathcal{L}'} \log_2 Q_T(i) + \log_2 Q_T(k) + \log_2 q_0 + \log_2 q_1 \\ &= \sum_{k \in \mathcal{B}} \log_2 Q_T(k) + (2^m - 1)[\log_2 q_0 + \log_2 q_1]. \end{aligned} \quad (11)$$

Since $(2^m - 1)[\log_2 q_0 + \log_2 q_1]$ is a constant independent of the tree T , we have

$$\operatorname{argmax}_{T \in \mathcal{C}} \sum_{i \in \mathcal{L}} \log_2 Q_T(i) = \operatorname{argmax}_{T \in \mathcal{C}} \sum_{k \in \mathcal{B}} \log_2 Q_T(k). \quad (12)$$

The right-hand side of (12) is clearly maximized by the complete tree with the branching nodes with the greatest weighted probabilities. According to [8, p. 47], this is exactly the tree that is constructed when Tunstall coding is applied to the weighted distribution Q .

Part ii. We now consider $q_0 \leq 1$ and $q_1 \leq 1$. Assume we have constructed a non-complete binary tree. Because of non-completeness, we can remove a branch from the tree. Without loss of generality, assume that this branch is labeled by a zero. Denote by \mathcal{S} the leaves on the subtree of the branch. Denote the tree after removing the branch by T' . Now,

$$Q_{T'}(i) = \frac{Q_T(i)}{q_0} \geq Q_T(i), \text{ for each } i \in \mathcal{S} \quad (13)$$

where the inequality follows because by assumption $q_0 \leq 1$. Thus, for the new tree T' , the objective function (11) is bounded as

$$\begin{aligned} \sum_{i \in \mathcal{L}} \log_2 Q_{T'}(i) &= \sum_{i \in \mathcal{L} \setminus \mathcal{S}} \log_2 Q_T(i) + \sum_{i \in \mathcal{S}} \log_2 \frac{Q_T(i)}{q_0} \\ &\geq \sum_{i \in \mathcal{L}} \log_2 Q_T(i). \end{aligned} \quad (14)$$

In summary, under the assumption $q_0 \leq 1$ and $q_1 \leq 1$, the objective function (11) that we want to maximize does not decrease when removing branches, which shows that there is an optimal complete tree. This proves the statement ii. of the proposition. ■

III. MINIMIZING I-DIVERGENCE PER BIT

The following two propositions relate the problem of minimizing the I-divergence per bit to the problem of minimizing the un-normalized I-divergence.

Let $\mathcal{T}' \subseteq \mathcal{T}$ be some set of binary trees with 2^m leaves and define

$$\Delta := \min_{T \in \mathcal{T}'} \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)}. \quad (15)$$

Proposition 2. *We have*

$$T^* := \operatorname{argmin}_{T \in \mathcal{T}'} \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} = \operatorname{argmin}_{T \in \mathcal{T}'} \mathbb{D}(U_T \| Q_T^\Delta) \quad (16)$$

where Q_T^Δ is the weighted distribution induced by $Q \circ 2^\Delta := [q_0 2^\Delta, q_1 2^\Delta]$.

Proof: By (15), for any tree $T \in \mathcal{T}'$, we have

$$\frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} \geq \Delta \text{ with equality if } T = T^* \quad (17)$$

$$\Rightarrow \mathbb{D}(U_T \| Q_T) - \Delta \mathbb{E}_{U_T}(L) \geq 0 \text{ with equality if } T = T^* \quad (18)$$

We write the left-hand side of (18) as

$$\begin{aligned} \mathbb{D}(U_T \| Q_T) - \Delta \mathbb{E}_{U_T}(L) &= \sum_{i \in \mathcal{L}} U_T(i) \log_2 \frac{U_T(i)}{Q_T(i)} - \Delta \sum_{i \in \mathcal{L}} U_T(i) \ell_i \\ &= \sum_{i \in \mathcal{L}} U_T(i) \left[\log_2 \frac{U_T(i)}{Q_T(i)} - \log_2 2^{\Delta \ell_i} \right] \\ &= \sum_{i \in \mathcal{L}} U_T(i) \log_2 \frac{U_T(i)}{Q_T(i) 2^{\Delta \ell_i}}. \end{aligned} \quad (19)$$

Consider the path through the tree that ends at leaf i . Denote by ℓ_i^0 and ℓ_i^1 the number of times the labels 0 and 1 occur, respectively. The length of the path can be expressed as $\ell_i = \ell_i^0 + \ell_i^1$. The term $Q_T(i) 2^{\Delta \ell_i}$ can now be written as

$$\begin{aligned} Q_T(i) 2^{\Delta \ell_i} &= q_0^{\ell_i^0} q_1^{\ell_i^1} 2^{\Delta(\ell_i^0 + \ell_i^1)} \\ &= (q_0 2^\Delta)^{\ell_i^0} (q_1 2^\Delta)^{\ell_i^1} \\ &= Q_T^\Delta(i). \end{aligned} \quad (20)$$

Using (20) and (19) in (18) shows that for any binary tree $T \in \mathcal{T}'$ we have

$$\sum_{i \in \mathcal{L}} U_T(i) \log_2 \frac{U_T(i)}{Q_T^\Delta(i)} \geq 0 \text{ with equality if } T = T^* \quad (21)$$

which is the statement of the proposition. ■

Proposition 3. *Define*

$$\Delta := \min_{T \in \mathcal{C}} \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)}. \quad (22)$$

Then the optimal complete tree

$$T^* := \operatorname{argmin}_{T \in \mathcal{C}} \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} \quad (23)$$

is constructed by applying Tunstall coding to Q_T^Δ .

Proof: The proposition is a consequence of Prop. 2 and Prop. 1.i. ■

Algorithm 1.

$\hat{T} \leftarrow \operatorname{argmin}_{T \in \mathcal{C}} \mathbb{D}(U_T \| Q_T)$ *solved by Tunstall coding on Q*
repeat
 1. $\hat{\Delta} = \frac{\mathbb{D}(U_{\hat{T}} \| Q_{\hat{T}})}{\mathbb{E}_{U_{\hat{T}}}(L)}$
 2. $\hat{T} = \operatorname{argmin}_{T \in \mathcal{C}} [\mathbb{D}(U_T \| Q_T) - \hat{\Delta} \mathbb{E}_{U_T}(L)]$ *Tunstall on $Q \circ 2^{\hat{\Delta}}$*
until $\mathbb{D}(U_{\hat{T}} \| Q_{\hat{T}}) - \hat{\Delta} \mathbb{E}_{U_{\hat{T}}}(L) = 0$
 $\Delta = \hat{\Delta}, T^* = \hat{T}$

A. Iterative Algorithm

By Prop. 3, if we know the I-divergence Δ , then we can find T^* by Tunstall coding. However, Δ is not known a priori. We solve this problem by iteratively applying Tunstall coding to $Q \circ 2^{\hat{\Delta}}$, where $\hat{\Delta}$ is an estimate of Δ and by updating our estimate. This procedure is stated in Alg. 1.

Proposition 4. *Alg. 1 finds (Δ, T^*) as defined in Prop. 3 in finitely many steps.*

Proof: The proof is similar to the proof of [3, Prop. 4.1].

We first show that Δ is strictly monotonically decreasing. Let $\hat{\Delta}_i$ be the value that is assigned to $\hat{\Delta}$ in step 1. of the i th iteration and denote by \hat{T}_i the value that is assigned to \hat{T} in step 2. of the i th iteration. Suppose that the algorithm does not terminate in the i th iteration. We have

$$\begin{aligned} \hat{\Delta}_i &= \frac{\mathbb{D}(U_{\hat{T}_{i-1}} \| Q_{\hat{T}_{i-1}})}{\mathbb{E}_{U_{\hat{T}_{i-1}}}(L)} \\ &\Rightarrow \mathbb{D}(U_{\hat{T}_{i-1}} \| Q_{\hat{T}_{i-1}}) - \hat{\Delta}_i \mathbb{E}_{U_{\hat{T}_{i-1}}}(L) = 0. \end{aligned} \quad (24)$$

By step 2, we have

$$\hat{T}_i = \operatorname{argmin}_{T \in \mathcal{C}} [\mathbb{D}(U_T \| Q_T) - \hat{\Delta}_i \mathbb{E}_{U_T}(L)] \quad (25)$$

and since by our assumption the algorithm does not terminate in the i th iteration, we have

$$\begin{aligned} \mathbb{D}(U_{\hat{T}_i} \| Q_{\hat{T}_i}) - \hat{\Delta}_i \mathbb{E}_{U_{\hat{T}_i}}(L) &< 0 \\ \Rightarrow \frac{\mathbb{D}(U_{\hat{T}_i} \| Q_{\hat{T}_i})}{\mathbb{E}_{U_{\hat{T}_i}}(L)} &< \hat{\Delta}_i \\ \Rightarrow \hat{\Delta}_{i+1} &< \hat{\Delta}_i. \end{aligned} \quad (26)$$

Now assume the algorithm terminated, and let \hat{T} be the tree after termination. Because of the assignments in steps 1. and 2., the terminating condition implies that for any tree $T \in \mathcal{C}$, we have

$$\mathbb{D}(U_T \| Q_T) - \hat{\Delta} \mathbb{E}_{U_T}(L) \geq 0, \text{ with equality if } T = \hat{T}. \quad (27)$$

Consequently, we have

$$\frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} \geq \hat{\Delta}, \text{ with equality if } T = \hat{T}. \quad (28)$$

We conclude that after termination, (Δ, \hat{T}) is equal to the optimal tuple (Δ, T^*) in Prop. 3.

Finally, we have shown that Δ is strictly monotonically decreasing so that $\hat{T}_i \neq \hat{T}_j$ for all $i < j$. But there is only a

finite number of complete binary trees with 2^m leaves. Thus, the algorithm terminates after finitely many steps. ■

IV. OPTIMALITY OF COMPLETE TREES

Complete trees are not optimal in general: Consider $m = 1$ and $Q: q_0 = \frac{5}{6}, q_1 = \frac{1}{6}$. For $m = 1$, Tunstall coding constructs the (unique) complete binary tree T with 2 leaves, independent of which target vector we pass to it. The path lengths are $\ell_1 = \ell_2 = 1$. The I-divergence per bit achieved by this is

$$\frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} = \frac{-1 - \frac{1}{2} \log_2(q_0 q_1)}{1} = 0.424 \text{ bits}. \quad (29)$$

Now, we could instead use a non-complete tree T with the paths 0 and 10. In this case, I-divergence per bit is

$$\frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} = \frac{-1 - \frac{1}{2} \log_2(q_0 q_1 q_0)}{\frac{1}{2}(1+2)} = 0.37034 \text{ bits}. \quad (30)$$

In summary, for the considered example, using a complete tree is sub-optimal. We will in the following derive simple conditions on the target vector Q that guarantee that the optimal tree is complete.

A. Sufficient Conditions for Optimality

Proposition 5. *Let Q be a distribution. If $\max\{q_0, q_1\} \leq 4 \min\{q_0, q_1\}$, then the optimal tree is complete for any $m \geq 1$ and it is constructed by Alg. 1.*

Proof: According to Prop. 1.ii, the tree that minimizes $\mathbb{D}(U_T \| Q_T)$ is complete if the entries of the weighted distribution $Q \circ 2^{\Delta}$ are both less than or equal to one. Without loss of generality, assume that $q_0 \geq q_1$. Thus, we only need to check this condition for q_0 . We have

$$\begin{aligned} q_0 2^{\Delta} &\leq 1 \\ \Leftrightarrow \log_2 q_0 + \Delta &\leq 0 \\ \Leftrightarrow \Delta &\leq -\log_2 q_0. \end{aligned} \quad (31)$$

We calculate the value of Δ that is achieved by the (unique) complete tree with 2 leaves, namely

$$\Delta = \frac{\mathbb{D}(U_T \| Q_T)}{\mathbb{E}_{U_T}(L)} = -1 - \frac{1}{2} \log_2(q_0 q_1). \quad (32)$$

For each $m \geq 1$, this Δ is achieved by the complete tree with all path lengths equal to m . Substituting the right-hand side of (32) for Δ in (31), we obtain

$$\begin{aligned} -1 - \frac{1}{2} \log_2(q_0 q_1) &\leq -\log_2 q_0 \\ \Leftrightarrow 1 + \log_2(q_0 q_1)^{\frac{1}{2}} &\geq \log_2 q_0 \\ \Leftrightarrow 2\sqrt{q_0 q_1} &\geq q_0 \\ \Leftrightarrow 4q_1 &\geq q_0 \end{aligned} \quad (33)$$

which is the condition stated in the proposition. ■

B. Asymptotic Achievability for Complete Trees

Proposition 6. Denote by $T(m)$ the complete tree with 2^m leaves that is constructed by applying Alg. 1 to a target distribution Q . Then we have

$$\frac{\mathbb{D}(U_{T(m)} \| Q_{T(m)})}{\mathbb{E}_{U_{T(m)}}(L)} \leq \frac{\log_2 \frac{1}{\min\{q_0, q_1\}}}{m} \quad (34)$$

and in particular, the I-divergence per bit approaches zero as $m \rightarrow \infty$.

Proof: The expected length can be bounded by the converse of the Coding Theorem for DMS [8, p. 45] as

$$\mathbb{E}_{U_{T(m)}}(L) \geq \mathbb{H}[U_{T(m)}] = m. \quad (35)$$

Thus, we have

$$\frac{\mathbb{D}(U_{T(m)} \| Q_{T(m)})}{\mathbb{E}_{U_{T(m)}}(L)} \leq \frac{\min_{T'(m) \in \mathcal{C}} \mathbb{D}(U_{T'(m)} \| Q_{T'(m)})}{m}. \quad (36)$$

The tree $T''(m)$ that minimizes the right-hand side is found by applying Tunstall coding to Q . Without loss of generality, assume that $q_0 \geq q_1$. According to the Tunstall Lemma [8, p. 47], the induced leaf probability of a tree constructed by Tunstall coding is lower bounded as

$$Q_{T''(m)}(i) \geq 2^{-m} q_1, \text{ for each leaf } i \in \mathcal{L}. \quad (37)$$

We can therefore bound the I-divergence as

$$\begin{aligned} \mathbb{D}(U_{T''(m)} \| Q_{T''(m)}) &= \sum_{i \in \mathcal{L}} 2^{-m} \log_2 \frac{2^{-m}}{Q_{T''(m)}(i)} \\ &\leq \sum_{i \in \mathcal{L}} 2^{-m} \log_2 \frac{2^{-m}}{2^{-m} q_1} \\ &= \log_2 \frac{1}{q_1}. \end{aligned} \quad (38)$$

We can now bound the I-divergence per bit as

$$\frac{\mathbb{D}(U_{T(m)} \| Q_{T(m)})}{\mathbb{E}_{U_{T(m)}}(L)} \leq \frac{\log_2 \frac{1}{q_1}}{\mathbb{E}_{U_{T(m)}}(L)} \leq \frac{\log_2 \frac{1}{q_1}}{m}. \quad (39)$$

This proves the proposition. \blacksquare

C. Optimality of Complete Trees for Large Enough m

Proposition 7. For any target distribution Q with $q_0 < 1$ and $q_1 = 1 - q_0$, there is an m_0 such that for all $m > m_0$, the tree that minimizes

$$\frac{\mathbb{D}(U_{T(m)} \| Q_{T(m)})}{\mathbb{E}_{U_{T(m)}}(L)} \quad (40)$$

is complete.

Proof: Without loss of generality, assume that $q_0 \geq q_1$.

By Prop. 6, we have $\Delta_m \leq \frac{\log_2 \frac{1}{q_1}}{m}$. Thus, there exists an m_0 such that

$$q_1 2^{\Delta_m} \leq q_0 2^{\Delta_m} \leq q_0 2^{\frac{\log_2 \frac{1}{q_1}}{m}} \leq 1, \text{ for all } m > m_0. \quad (41)$$

Thus, for all $m \geq m_0$, both entries of $Q_{T(m)}^{\Delta_m}$ are smaller than 1. The proposition now follows by Prop. 2 and Prop. 1.ii. \blacksquare

TABLE I
COMPARISON OF V2F SOURCE CODING AND F2V DISTRIBUTION
MATCHING: $Q: q_0 = 0.615, q_1 = 0.385; m = 2$

	Tunstall on Q	Alg. 1 on Q
v2f source encoder	00 \mapsto 00	1 \mapsto 00
	01 \mapsto 01	01 \mapsto 01
	10 \mapsto 10	001 \mapsto 10
	11 \mapsto 11	000 \mapsto 11
redundancy $\frac{\mathbb{D}(Q_T \ U_T)}{m}$	0.038503	0.04176
f2v distribution matcher	00 \mapsto 00	00 \mapsto 1
	01 \mapsto 01	01 \mapsto 01
	10 \mapsto 10	10 \mapsto 001
	11 \mapsto 11	11 \mapsto 000
I-divergence per bit $\frac{\mathbb{D}(U_T \ Q_T)}{\mathbb{E}_{U_T}(L)}$	0.039206	0.037695

V. SOURCE CODING VERSUS DISTRIBUTION MATCHING

An *ideal* source encoder transforms the output of a DMS Q into a sequence of bits that are independent and uniformly distributed. Reversely, applying the corresponding decoder to a sequence of uniformly distributed bits generates a sequence of symbols that are iid according to Q . This suggests to design a f2v distribution matcher by first calculating the optimal v2f source encoder. The inverse mapping is f2v and can be used as a distribution matcher.

We illustrate by an example that this approach is sub-optimal in general. Consider the DMS Q with $Q: q_0 = 0.615, q_1 = 0.385$. We calculate the optimal binary v2f source encoder with blocklength $m = 2$ by applying Tunstall coding to Q . The resulting encoder is displayed in the 1st column of Table I. Using the source decoder as a distribution matcher results in an I-divergence per bit of 0.039206 bits. Next, we use Alg. 1 to calculate the optimal f2v matcher for Q . The resulting mapping is displayed in the 2nd column of Table I. The achieved I-divergence per bit is 0.037695 bits, which is smaller than the value obtained by using the source decoder.

In general, the decoder of an optimal v2f source encoder is a sub-optimal f2v distribution matcher and the dematcher of an optimal v2f distribution matcher is a sub-optimal v2f source encoder.

REFERENCES

- [1] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 63–86, 1996.
- [2] D. Knuth and A. Yao, *The complexity of nonuniform random number generation*. New York: Academic Press, 1976, pp. 357–428.
- [3] G. Böcherer, "Capacity-achieving probabilistic shaping for noisy and noiseless channels," Ph.D. dissertation, RWTH Aachen University, 2012. [Online]. Available: <http://www.georg-boecherer.de/capacityAchievingShaping.pdf>
- [4] G. Böcherer and R. Mathar, "Matching dyadic distributions to channels," in *Proc. Data Compression Conf.*, 2011, pp. 23–32.
- [5] R. A. Rueppel and J. L. Massey, "Leaf-average node-sum interchanges in rooted trees with applications," in *Communications and Cryptography: Two sides of One Tapestry*, R. E. Blahut, D. J. Costello Jr., U. Maurer, and T. Mittelholzer, Eds. Kluwer Academic Publishers, 1994.
- [6] G. Böcherer, "Rooted trees with probabilities revisited," Feb. 2013. [Online]. Available: <http://arxiv.org/abs/1302.0753>
- [7] B. Tunstall, "Synthesis of noiseless compression codes," Ph.D. dissertation, 1967.
- [8] J. L. Massey, "Applied digital information theory I," lecture notes, ETH Zurich. [Online]. Available: http://www.isiweb.ee.ethz.ch/archive/massey_scr/adit1.pdf