Optimal $M$-Type Quantizations of Distributions

Georg Böcherer and Bernhard C. Geiger

Abstract—The construction of $M$-type quantizations of probability distributions over $n$ elements is investigated. Two algorithms are derived that find the optimal $M$-type quantizations in terms of the variational distance and the informational divergence, respectively. Bounds in terms of resulting quantization errors are provided.

I. INTRODUCTION

Let $t$ be a target probability distribution with $n$ non-negative entries $t_1, t_2, \ldots, t_n$ that sum up to one and let $M$ be a positive integer. A distribution $\hat{t}$ is $M$-type, if each entry can be written as $\hat{t}_i = \frac{c_i}{M}$ for some non-negative integer $c_i$. We want to determine the $M$-type distribution $\hat{t}$ that best approximates the target distribution $t$. Two quality measures for approximation are considered, namely, the variational distance

$$||\hat{t} - t||_1 := \sum_{i=1}^n |\hat{t}_i - t_i|$$

and the informational divergence

$$\mathbb{D}(\hat{t}||t) := \sum_{i: i>0} \hat{t}_i \log \frac{\hat{t}_i}{t_i}.$$ 

The main contribution of this work are two efficient algorithms that find the $M$-type quantizations that minimize the variational distance (Algorithm 1) and the informational divergence (Algorithm 2). In addition to that we provide bounds on the quantization error in terms of the number of entries $n$ and the quantization precision $M$. As an example application of our results, Algorithm 1 is used in Sec. IV in the construction of a fixed-to-variable length resolution code and Algorithm 2 is used in Sec. VI as a block-to-block resolution code.

We state $M$-type quantization algorithms and error bounds for variational distance and informational divergence in Sec. III and Sec. IV respectively. We give proofs in Sec. V.

II. VARIATIONAL DISTANCE-OPTIMAL QUANTIZATION

An ad hoc method for quantization is to initialize $\hat{t}$ by rounding off $t$ and distributing the remaining masses among the indices with the largest errors. In other words, we initialize $\hat{t}_i$ by

$$\hat{t}_i = \frac{\lfloor Mt_i \rfloor}{M}, \quad i = 1, \ldots, n.$$ 

Algorithm 1. Variational distance-optimal quantization

Initialize $\hat{t}_i := \frac{1}{M} \lfloor Mt_i \rfloor$, $i = 1, \ldots, n$.
Compute $e_i := t_i - \hat{t}_i$.
Compute $L := M \cdot \sum_{i=1}^n e_i$.
repeat $L$ times
Choose $j = \min \{ \arg \max e_i \}$.
Update $\hat{t}_j := \hat{t}_j + \frac{1}{M}$.
Set $e_j := 0$.
end repeat
Return $\hat{t}$.

Clearly, after initialization, $\hat{t}$ is a sub-probability vector whose elements do not sum up to one. This gives rise to the non-negative error

$$e_i := t_i - \frac{\lfloor Mt_i \rfloor}{M} \geq 0, \quad i = 1, \ldots, n$$

which sums up to the rest mass

$$\sum_{i=1}^n e_i = \frac{L}{M}$$

for some integer $L$. Note that $L$ can be zero (for $t$ already being $M$-type) and obviously cannot exceed $M$.

Suppose, without loss of generality, that the errors are ordered, i.e., $e_i \geq e_{i+1}$ for all $i = 1, \ldots, n-1$. We distribute the remaining $L$ unit masses to the first $L$ indices corresponding to the largest errors, i.e.,

$$\hat{t}_i = \begin{cases} \frac{\lfloor Mt_i \rfloor}{M}, & i \leq L \\ \frac{\lfloor Mt_i \rfloor}{M}, & i > L. \end{cases}$$

Method (6) has the property of assigning no mass to indices $i$ with $t_i = 0$, i.e.,

$$t_i = 0 \quad \Rightarrow \quad \hat{t}_i = 0.$$ 

Secondly, method (6) guarantees that the entries of $\hat{t}$ are close to those of $t$ in the sense that, for each $i$, \[ t_i - \frac{1}{M} \leq \hat{t}_i \leq t_i + \frac{1}{M}. \]

This property also guarantees that the quantization error can be made arbitrarily small by choosing $M$ sufficiently large. Algorithm 1 formulates method (6) for arbitrary target distributions. We summarize.

Proposition 1. Let $t$ be a target distribution with $n$ entries and let $M$ be a positive integer. Let $\hat{t}$ be the $M$-type quantization of $t$ found by Algorithm 1.

1) Among all $M$-type distributions $p$, $\|p - t\|_1$ is minimized for $p = \hat{t}$.
Algorithm 2. Informational divergence-optimal quantization

Initialize $c_i \leftarrow 0$, $i = 1, \ldots, n$.
repeat $M$ times
    Choose $j = \min\{\arg\min \Delta_i(c_i + 1)\}$.
    Update $c_j \leftarrow c_j + 1$.
end repeat
Return $c$.

2) For $n \leq M$, the quantization error is bounded as $\|\hat{t} - t\|_1 \leq \frac{n}{2M}$.

Proof: The proof is given in Sec. IV-A.

This ad hoc method is essentially equivalent to one proposed by Reznik [3, 4], which was shown to be optimal in terms of the variational distance, as well as in terms of the $L_2$ and $L_\infty$ distances. Furthermore, Reznik proved the upper bound of the quantization error using a different approach [3 Eq. (18)].

It is easily verified that this bound is tight for a uniform target distribution and $M = 3n/2$.

A. Complexity

The complexity of the algorithm depends on the number $L$ of unit masses to be distributed after initialization. For $n \leq M$, the fact $|Mt_i| > Mt_i - 1$ dictates that $L \leq n - 1$.

In each of the $L$ iterations, Algorithm 1 must find the maximum error, which is of complexity $O(n)$. The algorithm terminates after at most $\min\{n - 1\}$ iterations, yielding the overall complexity $O(n^2)$. The complexity can be brought down to $O(n \log n)$ by sorting the errors descendingly and distributing the remaining masses equally over the first $L$ indices.

III. INFORMATIONAL DIVERGENCE-OPTIMAL QUANTIZATION

We now consider $M$-type quantization with respect to the informational divergence, i.e., we want to solve the problem

$$\min_{\hat{t}} \mathbb{D}(\hat{t}||t)$$

subject to $\hat{t}$ is $M$-type. (9)

A. Equivalent Problem

Recall that each entry of an $M$-type distribution can be written as $\hat{t}_i = c_i/M$ for some non-negative integer $c_i$. Writing

$$\mathbb{D}(\hat{t}||t) = \sum_{i: c_i > 0} c_i \log \frac{c_i}{t_i}$$

leads to the conclusion that Problem (9) is equivalent to

$$\min_{c_1, \ldots, c_n} \sum_{i: c_i > 0} c_i \log \frac{c_i}{t_i}$$

subject to $c_i \in \{0, 1, 2, \ldots, M\}, \quad i = 1, \ldots, n$ (11)

$$\sum_{i=1}^n c_i = M.$$ If $c^*$ is a solution of Problem (11), then $\hat{t}^* = c^*/M$ is a solution of Problem (9).

B. Algorithm

To solve problem (11), we write the objective function as a telescoping sum

$$\sum_{i: c_i > 0} c_i \log \frac{c_i}{t_i} = \sum_{i=1}^n \sum_{k=1}^{c_i} \left( k \log \frac{k}{t_i} - (k-1) \log \frac{k-1}{t_i} \right)$$

$$= \sum_{i=1}^n \sum_{k=1}^{c_i} \Delta_i(k).$$ (12)

An allocation $c$ can be obtained by initially assigning the zero vector 0 to $c$ and then successively incrementing the elements of $c$. After $M$ iterations, the constraint $\sum_i c_i = M$ is fulfilled and $c$ is a valid allocation.

The cost for increasing the $j$th element from $c_j$ to $c_j + 1$ is $\Delta_j(c_j + 1)$. Algorithm 2 finds an allocation in a greedy manner: In each iteration, the element $c_i$ with the smallest cost $\Delta_i(c_i + 1)$ is increased by 1. We summarize the properties of Algorithm 2

Proposition 2. Let $t$ be a target distribution with $n$ entries. Let $\hat{t}$ be the $M$-type quantization of $t$ found by Algorithm 2.

1) Among all $M$-type distributions $p$, $\mathbb{D}(p||t)$ is minimized for $p = \hat{t}$.

2) For $n \leq M$, the quantization error is bounded as $\mathbb{D}(\hat{t}||t) \leq \frac{n}{2M}$ where $\mu := \min_{i: t_i > 0} t_i$.

Proof: The proof is given in Sec. IV-B.

Reznik briefly sketched an evaluation of his algorithm (similar to Algorithm 1) in terms of the informational divergence [3]. However, instead of bounding the informational divergence from above by the $\chi^2$-distance, he presented an approximation, which yields the quantization error bound stated above.

We strive to mention that, in contrary to the bound on the variational distance presented in Prop. 1 the bound on the informational divergence is not tight in general. This is because the used inequality $\log(x + 1) \leq x$ is tight only for $x = 0$, i.e., for very small differences between $\hat{t}_i$ and $t_i$ (for large $M$).

The increment in the $k$th iteration of Algorithm 2 does not depend on $M$. This means that the algorithm not only calculates the optimal $M$-type quantization, but actually all optimal $k$-type quantizations for $k = 1, 2, \ldots, M$. We state this property as a corollary of Prop. 2

Corollary 1. Let $c$ be the pre-allocation calculated by Algorithm 2 in the $k$th iteration and define

$$\hat{t} := \left( \frac{c_1}{k}, \ldots, \frac{c_n}{k} \right)^T.$$ Among all $k$-type distributions $p$, $\mathbb{D}(p||t)$ is minimized for $p = \hat{t}$. 
C. Complexity

Algorithm 2 must find the minimum of a vector with \( n \) elements in each iteration, which is of complexity \( \mathcal{O}(n) \). The algorithm terminates after \( M \) iterations, so the overall complexity is \( \mathcal{O}(n M) \). The complexity could be further reduced to \( \mathcal{O}(M \log n) \) by keeping the list of increments \( \Delta_i(c_i + 1) \) sorted.

IV. PROOFS

A. Proof of Proposition 1

Statement 1): To prove that Algorithm 1 is optimal among all algorithms satisfying (8) is straightforward: With the first \( L \) indices being assigned the rest mass, the variational distance computes to

\[
\|\hat{t} - t\|_1 = \sum_{i=1}^{L} \left( \frac{1}{M} - e_i \right) + \sum_{i=L+1}^{n} e_i.
\]

But this is minimized if the first \( L \) indices correspond to the \( L \) largest errors. It thus remains to show that the optimal algorithm satisfies (8).

We prove this by contradiction: Suppose that \( |\hat{t}_i - t_i| \geq \frac{1}{M} \) for some \( i \). Thus, for some \( k \in \mathbb{N} = \{1, 2, \ldots\} \), either \( \hat{t}_i = \frac{\lfloor M t_j \rfloor}{M} - k \) or \( \hat{t}_i = \frac{\lfloor M t_j \rfloor + k}{M} \). We treat only the first case, the second follows by proceeding along the same lines. Recognize that

\[
\hat{t}_i = \frac{\lfloor M t_j \rfloor - k}{M} \Rightarrow \exists j: \exists m \in \mathbb{N}: \hat{t}_j = \frac{\lfloor M t_j \rfloor + m}{M} \tag{14}
\]

because the entries of \( \hat{t} \) must sum up to one. Second, let \( \hat{t}^o \) be such that \( \hat{t}^o = \hat{t}_i + \frac{1}{M}, \hat{t}^o_j = \hat{t}_j - \frac{1}{M}, \) and \( \hat{t}^o = \hat{t}_k \) for all \( k \neq i, j \). We calculate

\[
\|\hat{t} - t\|_1 - \|\hat{t}^o - t\|_1 = t_i - \hat{t}_i - t_j + \hat{t}_j + \hat{t}^o_t - t_j - |\hat{t}^o_t - t_j| = \frac{1}{M} + \hat{t}_j - t_j - |\hat{t}^o_t - t_j| \tag{15}
\]

For \( m > 1 \) one obtains \( \hat{t}^o_t > t_j \) and (15) evaluates to \( \frac{2}{M} > 0 \). For \( m = 1 \), continuing with (15) yields

\[
\|\hat{t} - t\|_1 - \|\hat{t}^o - t\|_1 = \frac{1}{M} + \hat{t}_j + \hat{t}^o_t - 2t_j = \frac{2\lfloor M t_j \rfloor + 2}{M} - 2t_j > 0. \tag{16}
\]

We conclude that an optimal algorithm cannot violate (8); this proves optimality of Algorithm 1.

Statement 2): We now prove the upper bound. By the rest mass from (5), the empirical mean error per entry is \( \frac{1}{M n} \). Assuming without loss of generality that the errors \( e_i \) are ordered, we can bound the mean error from below and from above as

\[
\frac{1}{L} \sum_{i=1}^{L} e_i \geq \frac{L}{M n} \geq \frac{1}{n - L} \sum_{i=L+1}^{n} e_i \tag{17}
\]

where lower and upper bound coincide if all errors have the same magnitude, i.e., if \( e_i = L / M n \) for all \( i \). We can now bound the variational distance from above as a function of \( L \):

\[
\|\hat{t} - t\|_1 = \frac{L}{M n} + \sum_{i=L+1}^{n} e_i \leq \frac{L}{M n} + \frac{L^2}{M n} + \frac{(n - L)}{M n} = \frac{2L(n - L)}{M n} \tag{18}
\]

The right-hand side is maximized for \( L = n/2 \), which completes the proof.

B. Proof of Proposition 2

Statement 1): We need the following two lemmas.

Lemma 1. For each \( i \), if \( k > \ell \) then \( \Delta_i(k) > \Delta_i(\ell) \), i.e., the increment functions are strictly monotonically increasing.

Proof: We interpret the increment function \( \Delta_i \) as defined on the set of real numbers greater than 1 and calculate

\[
\frac{\partial}{\partial x} \Delta_i(x) = \log \frac{x}{x-1} > 0. \tag{19}
\]

Lemma 2. Let \( c^* \) be an optimal allocation. Let \( c \) be a pre-allocation with \( \sum_i c_i < M \) and \( c_i \leq c^*_i \) for \( i = 1, \ldots, n \). Take

\[
j \in \arg\min_i \Delta_i(c_i + 1). \tag{20}
\]

Then for some optimal allocation \( \hat{c} \) we have

\[
c_j + 1 \leq \hat{c}_j \tag{21}
\]

\[
c_i \leq \hat{c}_i, \quad i = 1, \ldots, n. \tag{22}
\]

Proof: Suppose we have

\[
c_j + 1 > c^*_j. \tag{23}
\]

Since \( c_j \leq c^*_j \) by assumption, (23) implies

\[
c_j + 1 = c^*_j + 1. \tag{24}
\]

Since \( \sum_i c_i < M \) and \( \sum_i c^*_i = M \), there must be at least one \( \ell \neq j \) with

\[
c^*_\ell \geq c_\ell + 1. \tag{25}
\]

By decreasing \( c^*_\ell \) by one and increasing \( c^*_j \) by one, the change of the objective function is \( \Delta_j(c^*_j + 1) - \Delta_\ell(c^*_\ell) \). We bound this change as follows.

\[
\Delta_j(c^*_j + 1) - \Delta_\ell(c^*_\ell) \overset{(a)}{\leq} \Delta_j(c^*_j + 1) - \Delta_\ell(c_\ell + 1) \overset{(b)}{=} \Delta_j(c_j + 1) - \Delta_\ell(c_\ell + 1) \overset{(c)}{\leq} 0 \tag{26}
\]

where (a) follows by (25) and Lemma 1, (b) follows by (24), and (c) follows by the definition of \( j \) in (20). We have to consider two cases. First, suppose we have strict inequality in either (26) or (28). Then the objective function is decreased, which contradicts the assumption that \( c^* \) is optimal. Thus, the supposition (23) is false and the statements of the lemma hold.
for $\tilde{c} = c^*$. Second, suppose we have equality both in (26) and (28). In this case, define the allocation

$$\tilde{c}_\ell = c_\ell^* - 1, \quad \tilde{c}_j = c_j^* + 1, \quad \tilde{c}_i = c_i^* \text{ for } i \neq j, \ell. \quad (29)$$

Equality in (26)–(28) implies optimality of $\tilde{c}$. By (24) and (25), we can verify that $\tilde{c}$ fulfills the statements of the lemma. This concludes the proof.

By Lemma 2, there is an optimal allocation $\tilde{c}$ such that in each iteration of Algorithm 2 we have

$$c_i \leq \tilde{c}_i, \quad i = 1, \ldots, n. \quad (30)$$

After termination of Algorithm 2 we have

$$M = \sum_i c_i \leq \sum_i \tilde{c}_i = M. \quad (31)$$

Statements (30) and (31) can only be true simultaneously if $c_i = \tilde{c}_i$ for all $i = 1, \ldots, n$. Consequently, the constructed allocation $c$ is optimal.

Statement 2: We now show the quantization error bound. Let $\hat{t}_{VD}$ be the $M$-type quantization obtained by Algorithm 1, i.e., $\hat{t}_{VD}$ is optimal w.r.t. the variational distance. Obviously,

$$D(\hat{t}|t) \leq D(\hat{t}_{VD}|t) \quad (32)$$

where $\hat{t}$ is obtained via Algorithm 2. This upper bound can further be loosened by employing Pearson’s $\chi^2$-distance

$$D(\hat{t}_{VD}|t) \leq \chi^2(\hat{t}_{VD}|t) := \sum_{i=1}^n \frac{(\hat{t}_{VD,i} - t_i)^2}{t_i} \quad (33)$$

where the inequality is due to $\log(x + 1) \leq x$.

By (7), indices $i$ for which $t_i = 0$ need not be considered in the sum above. One can thus bound $t_i$ from below by $\mu$ and write

$$\chi^2(\hat{t}_{VD}|t) \leq \frac{1}{\mu} \sum_{i=1}^n |\hat{t}_{VD,i} - t_i| |\hat{t}_{VD,i} - t_i| \leq \frac{1}{\mu} \underbrace{|\hat{t} - t|}_\leq \leq \frac{1}{\mu M} \|\hat{t} - t\|_1. \quad (34)$$

The corresponding upper bound from Proposition 1 completes the proof.

V. ACKNOWLEDGMENT

The authors thank Rana Ali Amjad for pointing out Corollary 1.

REFERENCES

[1] G. Böcherer, “Optimal non-uniform mapping for probabilistic shaping,” in Proc. Int. ITG Conf. on Systems, Communication and Coding (SCC), Munich, Jan. 2013, pp. 1–6.
[2] G. Böcherer and R. A. Amjad, “Fixed-to-variable length resolution coding for target distributions,” in Proc. IEEE Inf. Theory Workshop (ITW), 2013.
[3] Y. Reznik, “An algorithm for quantization of discrete probability distributions,” in Data Compression Conference (DCC), Snowbird, UT, Mar. 2011, pp. 333–342.
[4] Y. A. Reznik, V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, and B. Girod, “Fast quantization and matching of histogram-based image features,” in Proc. SPIE 7798, Applications of Digital Image Processing XXXIII, San Diego, Aug. 2010.