entropy

MDPI

Article

Greedy Algorithms for Optimal Distribution Approximation

Bernhard C. Geiger * and Georg Böcherer

Institute for Communications Engineering, Technical University of Munich, Munich 80290, Germany; georg.boecherer@tum.de

* Correspondence: geiger@ieee.org; Tel.: +49-89-289-23452

Academic Editor: Raúl Alcaraz Martínez

Received: 14 June 2016; Accepted: 11 July 2016; Published: 18 July 2016

Abstract: The approximation of a discrete probability distribution $\mathbf{t}$ by an M-type distribution $\mathbf{p}$ is considered. The approximation error is measured by the informational divergence $\mathbb{D}(\mathbf{t}\|\mathbf{p})$, which is an appropriate measure, e.g., in the context of data compression. Properties of the optimal approximation are derived and bounds on the approximation error are presented, which are asymptotically tight. A greedy algorithm is proposed that solves this M-type approximation problem optimally. Finally, it is shown that different instantiations of this algorithm minimize the informational divergence $\mathbb{D}(\mathbf{p}\|\mathbf{t})$ or the variational distance $\|\mathbf{p}-\mathbf{t}\|_1$.

Keywords: distribution approximation; finite precision; informational divergence; greedy algorithm

1. Introduction

In this work, we consider finite precision representations of probabilistic models. Suppose the original model, or *target distribution*, has n non-zero mass points and is given by $\mathbf{t} := (t_1, \dots, t_n)$. We wish to approximate it by a distribution $\mathbf{p} := (p_1, \dots, p_n)$ of which each entry is a rational number with a fixed denominator. In other words, for every i, $p_i = c_i/M$ for some non-negative integer $c_i \leq M$. The distribution $\mathbf{p}$ is called an *M-type distribution*, and the positive integer $M \geq n$ is the *precision* of the approximation. The problem is non-trivial, since computing the numerator c_i by rounding Mt_i to the nearest integer in general fails to yield a distribution.

M-type approximations have many practical applications, e.g., in political apportionments, M seats in a parliament need to be distributed to n parties according to the result of some vote $\mathbf{t}$. This problem led, e.g., to the development of *multiplier methods* [1]. In communications engineering, example applications are finite precision implementations of probabilistic data compression [2], distribution matching [3], and finite-precision implementations of Bayesian networks [4,5]. In all of these applications, the M-type approximation $\mathbf{p}$ should be close to the target distribution $\mathbf{t}$ in the sense of an appropriate error measure. Common choices for this approximation error are the variational distance and the informational divergences:

$$\|\mathbf{p}-\mathbf{t}\|_1 := \sum_{i=1}^{n} |p_i - t_i| \tag{1a}$$

$$\mathbb{D}(\mathbf{p}\|\mathbf{t}) := \sum_{i:\, p_i>0} p_i \log \frac{p_i}{t_i} \tag{1b}$$

$$\mathbb{D}(\mathbf{t}\|\mathbf{p}) := \sum_{i:\, t_i>0} t_i \log \frac{t_i}{p_i} \tag{1c}$$

where log denotes the natural logarithm.

Variational distance and informational divergence Equation (1b) have been considered by Reznik [6] and Böcherer [7], respectively, who presented algorithms for optimal M-type approximation and developed bounds on the approximation error. In a recent manuscript [8], we extended the existing works on Equation (1a,b) to target distributions with infinite support ($n = \infty$) and refined the bounds from [6,7].

In this work, we focus on the approximation error Equation (1c). It is an appropriate cost function for data compression [9] (Theorem 5.4.3) and seems appropriate for the approximation of parameters in Bayesian networks (see Section 4). Nevertheless, to the best of the authors' knowledge, the characterization of M-type approximations minimizing $\mathbb{D}(\mathbf{t}\|\mathbf{p})$ has not received much attention in literature so far.

Our contributions are as follows. In Section 2, we present an efficient greedy algorithm to find M-type distributions minimizing Equation (1c). We then discuss in Section 3 the properties of the optimal M-type approximation and bound the approximation error Equation (1c). Our bound incorporates a reverse Pinsker inequality recently suggested in [10] (Theorem 7). The algorithm we present is an instance of a greedy algorithm similar to *steepest ascent hill climbing* [11] (Chapter 2.6). As a byproduct, we unify this work with [6–8] by showing that also the algorithms optimal w.r.t. variational distance Equation (1a) and informational divergence Equation (1b) are instances of the same general greedy algorithm, see Section 2.

2. Greedy Optimization

In this section, we define a class of problems that can be optimally solved by a greedy algorithm. Consider the following example:

Example 1. *Suppose there are n queues with jobs, and you have to select M jobs minimizing the total time spent. A greedy algorithm suggests to select successively the job with the shortest duration, among the jobs that are at the front of their queues. If the jobs in each queue are ordered by increasing duration, then this greedy algorithm is optimal.*

We now make this precise: Let M be a positive integer, e.g., the number of jobs that have to be completed, and let $\delta_i \colon \mathbb{N} \to \mathbb{R}$, $i = 1, \ldots, n$, be a set of functions, e.g., $\delta_i(k)$ is the duration of the k-th job in the i-th queue. Let furthermore $\mathbf{c}_0 := (c_{1,0}, \ldots, c_{n,0}) \in \mathbb{N}_0^n$ be a *pre-allocation*, representing a constraint that has to be fulfilled (e.g., in the i-th queue at least $c_{i,0}$ jobs have to be completed) or a chosen initialization. Then, the goal is to minimize

$$U(\mathbf{c}) := \sum_{i=1}^{n} \sum_{k_i=c_{i,0}+1}^{c_i} \delta_i(k_i) \tag{2}$$

i.e., to find a *final allocation* $\mathbf{c} := (c_1, \ldots, c_n)$ satisfying $\|\mathbf{c}\|_1 = M$ and, for every i, $c_i \geq c_{i,0}$. A greedy method to obtain such a final allocation is presented in Algorithm 1. We show in Appendix A.1. that this algorithm is optimal if the functions δ_i satisfy certain conditions:

Algorithm 1: Greedy Algorithm

Initialize $k_i = c_{i,0}$, $i = 1, \ldots, n$.
repeat $M - \|\mathbf{c}_0\|_1$ times
  Compute $\delta_i(k_i + 1)$, $i = 1, \ldots, n$.
  Compute $j = \min \operatorname{argmin}_i \delta_i(k_i + 1)$. // (choose one minimal element) Update $k_j \leftarrow k_j + 1$.
end repeat
Return $\mathbf{c} = (k_1, \ldots, k_n)$.

Proposition 1. *If the functions $\delta_i(k)$ are non-decreasing in k, Algorithm 1 achieves a global minimum $U(\mathbf{c})$ for a given pre-allocation $\mathbf{c}_0$ and a given M.*

Remark 1. *The minimum of $U(\mathbf{c})$ may not be unique.*

Remark 2. *If a function $f_i\colon \mathbb{R} \to \mathbb{R}$ is convex, the difference $\delta_i(k) = f_i(k) - f_i(k-1)$ is non-decreasing in k. Hence, Algorithm 1 also minimizes*

$$U(\mathbf{c}) = \sum_{i=1}^{n} f_i(c_i). \tag{3}$$

Remark 3. *Note that the functions $\delta_i(k)$ need not be non-negative, i.e., in the view of Example 1, jobs may have negative duration. The functions $\delta_i(k)$ are non-negative, though, if $f_i\colon \mathbb{R} \to \mathbb{R}$ in Remark 2 is convex* and *non-decreasing.*

Remark 2 connects Algorithm 1 to steepest ascent hill climbing [11] (Chapter 2.6) with fixed step size and a constrained number of M steps.

We now show that instances of Algorithm 1 can find M-type approximations $\mathbf{p}$ minimizing each of the cost functions in Equation (1). Noting that $p_i = c_i/M$ for some non-negative integer c_i, we can rewrite the cost functions as follows:

$$\|\mathbf{p} - \mathbf{t}\|_1 = \frac{1}{M} \sum_{i=1}^{n} |c_i - Mt_i| \tag{4a}$$

$$\mathbb{D}(\mathbf{p}\|\mathbf{t}) = \frac{1}{M} \left(\sum_{i\colon c_i>0} c_i \log \frac{c_i}{t_i} \right) - \log M \tag{4b}$$

$$\mathbb{D}(\mathbf{t}\|\mathbf{p}) = \log M - H(\mathbf{t}) - \sum_{i\colon t_i>0} t_i \log c_i \tag{4c}$$

where $H(\cdot)$ denotes entropy in nats.

Ignoring constant terms, these cost functions are all instances of Remark 2 for convex functions $f_i\colon \mathbb{R} \to \mathbb{R}$ (see Table 1). Hence, the three different M-type approximation problems set up by Equation (1) can all be solved by instances of Algorithm 1, for a trivial pre-allocation $\mathbf{c}_0 = \mathbf{0}$ and after taking M steps. The final allocation $\mathbf{c}$ simply defines the M-type approximation by $p_i = c_i/M$.

For variational distance optimal approximation, we showed in [8] (Lemma 3), that every optimal M-type approximation satisfies $p_i \geq \lfloor Mt_i \rfloor / M$, hence one may speed up the algorithm by pre-allocating $c_{i,0} = \lfloor Mt_i \rfloor$. We furthermore show in Lemma 1 below that the support of the optimal M-type approximation in terms of Equation (1c) equals the support of $\mathbf{t}$ (if $M \geq n$). Assuming that $\mathbf{t}$ is positive, one can pre-allocate the algorithm with $c_{i,0} = 1$. We summarize these instantiations of Algorithm 1 in Table 1.

Table 1. Instances of Algorithm 1 Optimizing Equation (1).

Cost	$f_i(x)$	$\delta_i(k)$	$c_{i,0}$	References
$\|\mathbf{p} - \mathbf{t}\|_1$	$\|x - Mt_i\|$	$\|k - Mt_i\| - \|k - 1 - Mt_i\|$	$\lfloor Mt_i \rfloor$	[6,8]
$\mathbb{D}(\mathbf{p}\|\mathbf{t})$	$x \log(x/t_i)$	$k \log \frac{k}{k-1} + \log(k-1) - \log t_i$	0	[7,8]
$\mathbb{D}(\mathbf{t}\|\mathbf{p})$	$-t_i \log x$	$t_i \log((k-1)/k)$	$\lceil t_i \rceil$	This work

This list of instances of Algorithm 1 minimizing information-theoretic or probabilistic cost functions can be extended. For example, the χ^2-divergences $\chi^2(\mathbf{t}||\mathbf{p})$ and $\chi^2(\mathbf{p}||\mathbf{t})$ can also be minimized, since the functions inside the respective sums are convex. However, Rényi divergences of orders $\alpha \neq 1$ cannot be minimized by applying Algorithm 1.

3. M-Type Approximation Minimizing $\mathbb{D}(\mathbf{t}\|\mathbf{p})$

As shown in the previous section, Algorithm 1 presents a minimizer of the problem $\min_{\mathbf{p}} \mathbb{D}(\mathbf{t}\|\mathbf{p})$ if instantiated according to Table 1. Let us call this minimizer $\boldsymbol{t}^{\mathrm{a}}$. Recall that $\mathbf{t}$ is positive and that

$M \geq n$. The support of $\boldsymbol{t}^{\mathrm{a}}$ must contain the support of $\mathbf{t}$, since otherwise $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) = \infty$. Note further that the costs $\delta_i(k)$ are negative if $t_i > 0$ and zero if $t_i = 0$; hence, if $t_i = 0$, the index i cannot be chosen by Algorithm 1, thus also $t_i^{\mathrm{a}} = 0$. This proves:

Lemma 1. *If $M \geq n$, the supports of $\mathbf{t}$ and $\boldsymbol{t}^{\mathrm{a}}$ coincide, i.e., $t_i = 0 \Leftrightarrow t_i^{\mathrm{a}} = 0$.*

The assumption that $\mathbf{t}$ is positive and that $M \geq n$ hence comes without loss of generality. In contrast, neither variational distance nor informational divergence Equation (1b) require $M \geq n$: As we show in [8], the M-type approximation problem remains interesting even if $M < n$.

Based on Lemma 1, the following example explains why the optimal M-type approximation does not necessarily result in a "small" approximation error:

Example 2. *Let $\mathbf{t} = (1-\varepsilon, \frac{\varepsilon}{n-1}, \ldots, \frac{\varepsilon}{n-1})$ and $M = n$, hence by Lemma 1, $\boldsymbol{t}^{\mathrm{a}} = \frac{1}{n}(1, 1, \ldots, 1)$. It follows that $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) = \log n - H(\mathbf{t})$, which can be made arbitrarily close to $\log n$ by choosing a small positive ε.*

In Table 1 we made use of [8] (Lemma 3), which says that every $\mathbf{p}$ minimizing the variational distance $\|\mathbf{p} - \mathbf{t}\|_1$ satisfies $p_i \geq \lfloor Mt_i \rfloor / M$, to speed up the corresponding instance of Algorithm 1 by proper pre-allocation. Initialization by rounding is not possible when minimizing $\mathbb{D}(\mathbf{t}\|\mathbf{p})$, as shown in the following two examples:

Example 3. *Let $\mathbf{t} = (17/20, 3/40, 3/40)$ and $M = 20$. The optimal M-type approximation is $\mathbf{p} = (8/10, 1/10, 1/10)$, hence $p_1 < \lfloor Mt_1 \rfloor / M$. Initialization via rounding off fails.*

Example 4. *Let $\mathbf{t} = (0.719, 0.145, 0.088, 0.048)$ and $M = 50$. The optimal M-type approximation is $\mathbf{p} = (0.74, 0.14, 0.08, 0.04)$, hence $p_1 > \lceil Mt_1 \rceil / M$. Initialization via rounding up fails.*

To show that informational divergence vanishes for $M \to \infty$, assume that $M > 1/t_i$ for all i. Since the variational distance optimal approximation $\boldsymbol{t}^{\mathrm{vd}}$ satisfies $t_i^{\mathrm{vd}} \geq \lfloor Mt_i \rfloor / M$ for every i, $\boldsymbol{t}^{\mathrm{vd}}$ has the same support as $\mathbf{t}$, which ensures that $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}}) < \infty$. By similar arguments as in the proof of [8] (Proposition 4), we obtain

$$\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) \leq \mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}}) \leq \log\left(1 + \frac{n}{2M}\right) \overset{M\to\infty}{\longrightarrow} 0. \tag{5}$$

Note that this bound is *universal*, i.e., it prescribes the same convergence rate for every target distribution with n mass points.

We now develop an upper bound on $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}})$ that holds for every M. To this end, we first approximate $\mathbf{t}$ by a distribution $\mathbf{t}^*$ in $\mathcal{P}_M := \{\mathbf{p} \colon \forall i \colon p_i \geq 1/M, \|\mathbf{p}\|_1 = 1\}$ that minimizes $\mathbb{D}(\mathbf{t}\|\mathbf{t}^*)$. If $\mathbf{t}^*$ is unique, then it is called the *reverse I-projection* [12] (Section I.A) of $\mathbf{t}$ onto $\mathcal{P}_M$. Since $\mathbf{t}^* \in \mathcal{P}_M$, its variational distance optimal approximation $\boldsymbol{t}^{\mathrm{vd}}$ has the same support as $\mathbf{t}$, which allows us to bound $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}})$ by $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}})$.

Lemma 2. *Let $\mathbf{t}^* \in \mathcal{P}_M$ minimize $\mathbb{D}(\mathbf{t}\|\mathbf{t}^*)$. Then,*

$$t_i^* := \frac{t_i}{\nu(M)} + \left(\frac{1}{M} - \frac{t_i}{\nu(M)}\right)^+ \tag{6}$$

where $\nu(M)$ is such that $\|\mathbf{t}^*\|_1 = 1$, and where $(x)^+ := \max\{0, x\}$.

Proof. See Appendix A.2. □

Let $\mathcal{K} := \{i : t_i < \nu(M)/M\}$, $k := |\mathcal{K}|$, and $T_{\mathcal{K}} := \sum_{i\in\mathcal{K}} t_i$. The parameter $\nu(M)$ must scale the mass $(1 - T_{\mathcal{K}})$ such that it equals $(M-k)/M$, i.e., we have

$$\nu(M) = \frac{1 - T_{\mathcal{K}}}{1 - \frac{k}{M}}. \tag{7}$$

If, for all i, $t_i > 1/M$, then $\mathbf{t} \in \mathcal{P}_M$, hence $\mathbf{t}^* = \mathbf{t}$ is feasible and $\nu(M) = 1$. One can show that $\nu(M)$ decreases with M.

Proposition 2 (Approximation Bounds)**.**

$$\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) \leq \log \nu(M) + \frac{\log(2)}{2}\left(1 - \nu(M)\left(1 - \frac{n}{M}\right)\right) \tag{8}$$

Proof. See Appendix A.3. □

The first term on the right-hand side of Equation (8) accounts for the error caused by first approximating $\mathbf{t}$ by $\mathbf{t}^*$ (in the sense of Lemma 2). The second term accounts for the additional error caused by the M-type approximation of $\mathbf{t}^*$ and incorporates the reverse Pinsker inequality [10] (Theorem 7). If $M > t_i$ for every i, hence $\mathbf{t} \in \mathcal{P}_M$, then $\nu(M) = 1$ and the bound simplifies to

$$\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) \leq \log(2)\frac{n}{2M}. \tag{9}$$

For M sufficiently large, Equation (8) thus yields better results than Equation (5), which approximates to $n/(2M)$. Moreover, for M sufficiently large, our bound Equation (8) is uniform, i.e., it prescribes the same convergence rate for every target distribution with n mass points. We illustrate the bounds for an example in Figure 1.

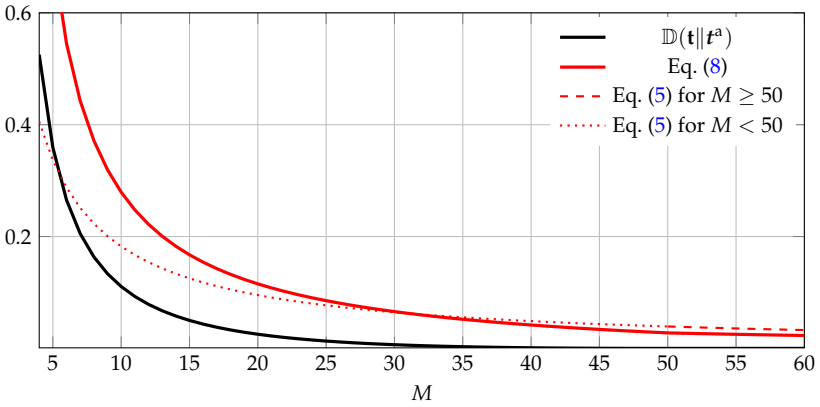

Figure 1. Evaluating the bounds Equations (5) and (8) for $\mathbf{t} = (0.48, 0.48, 0.02, 0.02)$. Note that Equation (5) is a valid bound only for $M \geq 50$, i.e., where the curve is dashed.

4. Applications and Outlook

Arithmetic coding uses a probabilistic model to compress a source sequence. Applying Algorithm 1 with cost Equation (1c) to the empirical distribution of the source sequence provides an M-type distribution as a probabilistic model. The parameter M can be choosen small for reduced complexity. Another application of Algorithm 1 can be found in [3], which considers the problem of generating length-M sequences according to a desired distribution. Since a length-M sequence has an M-type empirical distribution, the Reference [3] applies Algorithm 1 with cost Equation (1b) to pre-calculate the M-type approximation of the desired distribution.

Algorithm 1 can also be used to calculate the M-type approximation of Markov models, i.e., approximating the transition matrix $\mathbf{T}$ of an n-state, irreducible Markov chain with invariant distribution vectors $\boldsymbol{\mu}$ by a transition matrix $\mathbf{P}$ containing only M-type probabilities.

Generalizing Equation (1c), the approximation error can be measured by the informational divergence rate [13]

$$\overline{\mathbb{D}}(\mathbf{T}\|\mathbf{P}) := \sum_{i,j=1}^{n} \mu_i T_{ij} \log \frac{T_{ij}}{P_{ij}} = \sum_{i=1}^{n} \mu_i \, \mathbb{D}(\mathbf{t}_i\|\mathbf{p}_i). \tag{10}$$

The optimal M-type approximation is found by applying the instance of Algorithm 1 to each row separately, and Lemma 1 ensures that the transition graph of $\mathbf{P}$ equals that of $\mathbf{T}$, i.e., the approximating Markov chain is irreducible. Future work shall extend this analysis to hidden Markov models and should investigate the performance of these algorithms in practical scenarios, e.g., speech processing with finite-precision arithmetic.

Another possible application is the approximation of Bayesian network parameters. The authors of [4] approximated the true parameters using a stationary multiplier method from [14]. Since rounding probabilities to zero led to bad classification performance, they replaced zeros in the approximating distribution afterwards by small values. This in turn led to the problem that probabilities that are in fact zero, were approximated by a non-zero probability. We believe that these problems can be removed by instantiating Algorithm 1 for cost Equation (1c). This automatically prevents approximating non-zero probabilities with zeros and vice-versa, see Lemma 1.

Finally, for approximating Bayesian network parameters, recent work suggests rounding *log-probabilities*, i.e., to approximate $\log t_i$ by $\log p_i = -c_i/M$ for a non-negative integer c_i [5]. Finding an optimal approximation that corresponds to a true distribution is equivalent to solving

$$\begin{aligned} &\min \; d(\mathbf{t},\mathbf{p}) \\ &\text{s.t. } \|e^{-\mathbf{c}}\|_{1/M} = 1 \end{aligned}$$

where $d(\cdot,\cdot)$ denotes any of the considered cost functions Equation (1). If $M = 1$ and $d(\mathbf{t},\mathbf{p}) = \mathbb{D}(\mathbf{t}\|\mathbf{p})$ using the binary logarithm, the constraint translates to the requirement that $\mathbf{t}$ is approximated by a complete binary tree. Then, the optimal approximation is the Huffman code for $\mathbf{t}$.

Acknowledgments: The work of Bernhard C. Geiger was partially funded by the Erwin Schrödinger Fellowship J 3765 of the Austrian Science Fund. The work of Georg Böcherer was partly supported by the German Ministry of Education and Research in the framework of an Alexander von Humboldt Professorship. This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.

Author Contributions: Bernhard C. Geiger and Georg Böcherer conceived this study, derived the results, and wrote the manuscript. Specifically, Bernhard C. Geiger proved Proposition 1 and Lemmas 3, 5 and 6, and Georg Böcherer proved Lemmas 2 and 4. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix. Proofs

Appendix A.1. Proof of Proposition 1

Since a pre-allocation only fixes a lower bound for $U(\mathbf{c})$, w.l.o.g. we assume that $\mathbf{c}_0 = \mathbf{0}$ and thus $\mathbf{c} \in \mathbb{N}_0^n$ with $\|\mathbf{c}\|_1 = M$. Consider the set $\mathcal{D} := \{\delta_i(k_i)\colon k_i \in \mathbb{N}, i = 1,\dots,n\}$ and assume that the (not necessarily unique) set $\mathcal{D}_M$ consists of M smallest values in $\mathcal{D}$, i.e., $|\mathcal{D}_M| = M$ and

$$\forall d \in \mathcal{D}_M, d' \in \mathcal{D} \setminus \mathcal{D}_M\colon \quad d \leq d'. \tag{A1}$$

Clearly, $U(\mathbf{c})$ cannot be smaller than the sum over all elements in $\mathcal{D}_M$. Since the δ_i are non-decreasing, there exists at least one final allocation $\mathbf{c}$ that takes successively the first c_i values from each queue i, i.e., $\mathcal{D}_M = \{\delta_1(1),\dots,\delta_1(c_1),\dots,\delta_n(1),\dots,\delta_n(c_n)\}$ satisfies Equation (A1). This shows that the lower bound induced by Equation (A1) can actually be achieved.

We prove the optimality of Algorithm 1 by contradiction: Assume that Algorithm 1 finishes with a final allocation $\tilde{\mathbf{c}}$ such that $U(\tilde{\mathbf{c}})$ is strictly larger than the (unique) sum over all elements in

(non-unique) $\mathcal{D}_M$. Hence, $\tilde{\mathbf{c}}$ must exchange at least one of the elements in $\mathcal{D}_M$ for an element that is strictly larger. Thus, by the properties of the functions δ_i and Algorithm 1, there must be indices ℓ and m such that $\tilde{c}_\ell > c_\ell$, $\tilde{c}_m < c_m$, and $\delta_\ell(\tilde{c}_\ell) \geq \delta_\ell(c_\ell + 1) > \delta_m(c_m) \geq \delta_m(\tilde{c}_m)$. At each iteration of the algorithm, the current allocation at index m satisfies $k_m \leq \tilde{c}_m < c_m$. Since $\delta_m(c_m) < \delta_\ell(c_\ell + 1)$, $\delta_\ell(c_\ell + 1)$ can never be a minimal element, and hence is not chosen by Algorithm 1. This contradicts the assumption that Algorithm 1 finishes with a $\tilde{\mathbf{c}}$ such that $U(\tilde{\mathbf{c}})$ is strictly larger than the sum of $\mathcal{D}$'s M smallest values. □

Appendix A.2. Proof of Lemma 2

The problem finding a $\mathbf{t}^* \in \mathcal{P}_M$ minimizing $\mathbb{D}(\mathbf{t}\|\mathbf{t}^*)$ is equivalent to finding an optimal point of the problem:

$$\underset{\mathbf{p}\in\mathbb{R}^n_{>0}}{\text{minimize}} \quad -\sum_{i=1}^{n} t_i \log p_i \tag{A2a}$$

$$\text{subject to} \quad \frac{1}{M} - p_i \leq 0, \quad i = 1, 2, \ldots, n \tag{A2b}$$

$$-1 + \sum_{i=1}^{n} p_i = 0. \tag{A2c}$$

The Lagrangian of the problem is

$$L(\mathbf{p}, \boldsymbol{\lambda}, \nu) = -\sum_{i=1}^{n} t_i \log p_i + \sum_{i=1}^{n} \lambda_i \left(\frac{1}{M} - p_i\right) + \nu \left(-1 + \sum_{i=1}^{n} p_i\right). \tag{A3}$$

By the *Karush–Kuhn–Tucker* (KKT) conditions [15] (Chapter 5.5.3), a feasible point $\mathbf{t}^*$ is optimal if, for every $i = 1, \ldots, n$,

$$\lambda_i \geq 0 \tag{A4a}$$

$$\lambda_i \left(\frac{1}{M} - t_i^*\right) = 0 \tag{A4b}$$

$$\frac{\partial}{\partial p_i} L(\mathbf{p}, \boldsymbol{\lambda}, \nu)|_{\mathbf{p}=\mathbf{t}^*} = -\frac{t_i}{t_i^*} - \lambda_i + \nu = 0. \tag{A4c}$$

By Equation (A2b), we have $t_i^* \geq 1/M$. If $t_i^* > 1/M$, then $\lambda_i = 0$ by Equation (A4b) and $t_i^* = t_i/\nu$ by Equation (A4c). Thus

$$t_i^* = \frac{t_i}{\nu} + \left(\frac{1}{M} - \frac{t_i}{\nu}\right)^+ \tag{A5}$$

where ν is such that $\sum_{i=1}^{n} t_i^* = 1$. □

Appendix A.3. Proof of Proposition 2

Reverse I-projections admit a Pythagorean inequality [12] (Theorem 1). In other words, if $\mathbf{p}$ is a distribution, $\mathbf{p}^*$ its reverse I-projection onto a set $\mathcal{S}$, and $\mathbf{q}$ any distribution in $\mathcal{S}$, then

$$\mathbb{D}(\mathbf{p}\|\mathbf{q}) \geq \mathbb{D}(\mathbf{p}\|\mathbf{p}^*) + \mathbb{D}(\mathbf{p}^*\|\mathbf{q}). \tag{A6}$$

For the present scenario, we can show an even stronger result:

Lemma 3. *Let* $\mathbf{t}$ *be the target distribution, let* $\mathbf{t}^*$ *be as in Lemma 2, and let* $\boldsymbol{t}^{\mathrm{vd}}$ *be the variational distance optimal M-type approximation of* $\mathbf{t}^*$*. Then,*

$$\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}}) = \mathbb{D}(\mathbf{t}\|\mathbf{t}^*) + \nu\,\mathbb{D}(\mathbf{t}^*\|\boldsymbol{t}^{\mathrm{vd}}). \tag{A7}$$

Proof.

$$\begin{aligned}
\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}}) &= \sum_{i=1}^{n} t_i \log \frac{t_i}{t_i^{\mathrm{vd}}} && \text{(A8)}\\
&= \sum_{i=1}^{n} t_i \log \frac{t_i t_i^*}{t_i^{\mathrm{vd}} t_i^*} && \text{(A9)}\\
&= \sum_{i=1}^{n} t_i \log \frac{t_i}{t_i^*} + \sum_{i=1}^{n} t_i \log \frac{t_i^*}{t_i^{\mathrm{vd}}} && \text{(A10)}\\
&\overset{(a)}{=} \mathbb{D}(\mathbf{t}\|\mathbf{t}^*) + \nu \sum_{i\notin\mathcal{K}} \frac{t_i}{\nu} \log \frac{t_i^*}{t_i^{\mathrm{vd}}} && \text{(A11)}\\
&\overset{(b)}{=} \mathbb{D}(\mathbf{t}\|\mathbf{t}^*) + \nu\,\mathbb{D}(\mathbf{t}^*\|\boldsymbol{t}^{\mathrm{vd}}) && \text{(A12)}
\end{aligned}$$

Here, (a) follows because for $i \in \mathcal{K}$, $t_i^* = 1/M$ and thus, the M-type approximation minimizing the variational distance satisfies $t_i^{\mathrm{vd}} = 1/M$; furthermore, (b) is because for $i \notin \mathcal{K}$, $t_i^* = t_i/\nu$. □

We now bound the summands in Lemma 3.

Lemma 4. *In the setting of Lemma 3,*

$$\mathbb{D}(\mathbf{t}^*\|\boldsymbol{t}^{\mathrm{vd}}) \le \log(2)\|\mathbf{t}^* - \boldsymbol{t}^{\mathrm{vd}}\|_1. \tag{A13}$$

Proof. We first employ a *reverse Pinsker inequality* from [10] (Theorem 7), stating that

$$\mathbb{D}(\mathbf{t}^*\|\boldsymbol{t}^{\mathrm{vd}}) \le \frac{1}{2}\frac{r\log r}{r-1}\|\mathbf{t}^* - \boldsymbol{t}^{\mathrm{vd}}\|_1 \tag{A14}$$

where $r := \sup_{i:\, t_i^*>0} \frac{t_i^*}{t_i^{\mathrm{vd}}}$. Furthermore, since for variational distance optimal approximations we always have $|t_i^* - t_i^{\mathrm{vd}}| < 1/M$ [8] (Lemma 3), we can bound

$$r < \frac{t_i^{\mathrm{vd}} + \frac{1}{M}}{t_i^{\mathrm{vd}}} \le 2 \tag{A15}$$

since $t_i^{\mathrm{vd}} \ge \lfloor M t_i^* \rfloor / M \ge 1/M$. Since the factor $\frac{r\log r}{r-1}$ increases in r, the bound Equation (A13) follows by substituting r in Equation (A14) by 2. □

Lemma 5. *In the setting of Lemma 3,*

$$\mathbb{D}(\mathbf{t}\|\mathbf{t}^*) \le \log \nu. \tag{A16}$$

Proof.

$$\begin{aligned}
\mathbb{D}(\mathbf{t}\|\mathbf{t}^*) &= \sum_{i=1}^{n} t_i \log \frac{t_i}{t_i^*} && \text{(A17)}\\
&= \sum_{i\notin\mathcal{K}} t_i \log \frac{\nu t_i}{t_i} + \sum_{i\in\mathcal{K}} t_i \log M t_i && \text{(A18)}\\
&\overset{(a)}{\le} (1 - T_{\mathcal{K}}) \log \nu + \sum_{i\in\mathcal{K}} t_i \log \nu && \text{(A19)}\\
&= \log \nu && \text{(A20)}
\end{aligned}$$

where (a) is because for $i \in \mathcal{K}$, $Mt_i \leq \nu$. $\square$

To bound $\|\mathbf{t}^* - \boldsymbol{t}^{\mathrm{vd}}\|_1$, we present

Lemma 6. *Let $\mathbf{p}^*$ be a sub-probability distribution with $m \leq M$ masses and total weight $1-T$, and let $\mathbf{p}^{\mathrm{vd}^*}$ be its variational distance optimal M-type approximation using $J \leq M$ masses. Then,*

$$\|\mathbf{p}^* - \mathbf{p}^{\mathrm{vd}^*}\|_1 \leq \frac{m}{2M} + \frac{(M - MT - J)^2}{2mM}. \tag{A21}$$

Note that for $J = M$ we recover [8] (Lemma 4).

Proof. Assume first that either $\forall i$: $p_i^* \geq p_i^{\mathrm{vd}^*}$ or $\forall i$: $p_i^* \leq p_i^{\mathrm{vd}^*}$. Note that this is possible since $\mathbf{p}^*$ and $\mathbf{p}^{\mathrm{vd}^*}$ are sub-probability distributions, summing to $1-T$ and J/M, respectively. Then, $\|\mathbf{p}^* - \mathbf{p}^{\mathrm{vd}^*}\|_1 = |1 - T - J/M|$ which satisfies this bound. This can be seen by rearranging Equation (A21) such that J only appears on the left-hand side; the maximizing J (not necessarily integer) then satisfies Equation (A21) with equality.

We thus remain to treat the case where after rounding off all indices, $1 \leq L \leq M-1$ masses remain and we have

$$\sum_{i=1}^{m} p_i^* - \frac{\lfloor Mp_i^* \rfloor}{M} =: \sum_{i=1}^{m} e_i = 1 - T - \frac{J-L}{M} =: g(L). \tag{A22}$$

The variational distance is minimized by distributing the L masses to L indices $i \in \mathcal{L}$ with the largest errors e_i, hence

$$\|\mathbf{p}^* - \mathbf{p}^{\mathrm{vd}^*}\|_1 = \sum_{i \in \mathcal{L}} \left(\frac{1}{M} - e_i\right) + \sum_{i \notin \mathcal{L}} e_i \tag{A23}$$

$$\overset{(a)}{\leq} \frac{L}{M} - \frac{L}{n} g(L) + \frac{n-L}{n} g(L) \tag{A24}$$

where (a) follows because for $i \in \mathcal{L}, j \notin \mathcal{L}$, $e_i \geq e_j$. This is maximized for $L = \frac{n-(M-MT-J)}{2}$ (not necessarily integer), which after inserting yields the upper bound. $\square$

Proof of Bound in Proposition 2. We start by bounding the informational divergence $\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}})$ by the informational divergence between $\mathbf{t}$ and the variational distance optimal approximation $\boldsymbol{t}^{\mathrm{vd}}$ of its reverse I-projection $\mathbf{t}^*$ onto $\mathcal{P}_M$:

$$\mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{a}}) \leq \mathbb{D}(\mathbf{t}\|\boldsymbol{t}^{\mathrm{vd}}) \tag{A25}$$

$$\overset{(a)}{=} \mathbb{D}(\mathbf{t}\|\mathbf{t}^*) + \nu\, \mathbb{D}(\mathbf{t}^*\|\boldsymbol{t}^{\mathrm{vd}}) \tag{A26}$$

$$\overset{(b)}{\leq} \log \nu + \nu \log(2) \|\mathbf{t}^* - \boldsymbol{t}^{\mathrm{vd}}\|_1 \tag{A27}$$

$$\overset{(c)}{\leq} \log \nu + \nu \log(2) \frac{n-k}{2M} \tag{A28}$$

$$\overset{(d)}{\leq} \log \nu + \nu \log(2) \frac{n - M + \frac{M}{\nu}}{2M} \tag{A29}$$

$$= \log \nu + \frac{\log(2)}{2} \left(1 - \nu \left(1 - \frac{n}{M}\right)\right) \tag{A30}$$

where

- (a) is due to Lemma 3,
- (b) is due to Lemmas 4 and 5,
- (c) is due to Lemma 6 with $m = n-k$, $1-T = 1-k/M$, and $J = M-k$, and

(d) follows by bounding k from below via Equation (7)

$$k = \frac{M}{\nu}(\nu - 1 + T_{\mathcal{K}}) \geq \frac{M}{\nu}(\nu - 1) = M - \frac{M}{\nu}. \tag{A31}$$

□

References

1. Dorfleitner, G.; Klein, T. Rounding with multiplier methods: An efficient algorithm and applications in statistics. *Stat. Pap.* **1999**, *40*, 143–157.
2. Rissanen, J.; Langdon, G.G. Arithmetic coding. *IBM J. Res. Dev.* **1979**, *23*, 149–162.
3. Schulte, P.; Böcherer, G. Constant Composition Distribution Matching. *IEEE Trans. Inf. Theory* **2016**, *62*, 430–434.
4. Drużdżel, M.J.; Oniśko, A. Are Bayesian Networks Sensitive to Precision of Their Parameters? In Proceedings of the International IIS'08 Conference, Intelligent Information Systems XVI, Zakopane, Poland, 16–18 June 2008; pp. 35–44.
5. Tschiatschek, S.; Pernkopf, F. On Bayesian Network Classifiers with Reduced Precision Parameters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 774–785.
6. Reznik, Y. An Algorithm for Quantization of Discrete Probability Distributions. In Proceedings of the 2011 Data Compression Conference (DCC), Snowbird, UT, USA, 29–31 March 2011; pp. 333–342.
7. Böcherer, G. Optimal Non-Uniform Mapping for Probabilistic Shaping. In Proceedings of the 9th International ITG Conference on Systems, Communications and Coding (SCC), Munich, Germany, 21–24 January 2013; pp. 1–6.
8. Böcherer, G.; Geiger, B.C. Optimal Quantization for Distribution Synthesis. **2016**, arXiv:1307.6843v4.
9. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Interscience: Hoboken, NJ, USA, 2006.
10. Verdú, S. Total variation distance and the distribution of relative information. In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 9–14 February 2014; pp. 499–501.
11. Michalewicz, Z.; Fogel, D.B. *How to Solve It: Modern Heuristics*, 2nd ed.; Springer: Berlin, Germany, 2004.
12. Csiszár, I.; František, M. Information Projections Revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1474–1490.
13. Rached, Z.; Alajaji, F.; Campbell, L.L. The Kullback–Leibler divergence rate between Markov sources. *IEEE Trans. Inf. Theory* **2004**, *50*, 917–921.
14. Heinrich, L.; Pukelsheim, F.; Schwingenschlögl, U. On stationary multiplier methods for the rounding of probabilities and the limiting law of the Sainte-Laguë divergence. *Stat. Decis.* **2005**, *23*, 117–129.
15. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.