Approximate Bayesian Inference by Adaptive Quantization of the Hypothesis Space

Mathias Johansson[,†]

Signals and Systems Group, Uppsala University, Sweden,
e-mail: mj@signal.uu.se
Dirac Research AB, Sweden, e-mail: mj@dirac.se

Abstract. We introduce a method for making approximate Bayesian inference based on quantizing the hypothesis space and repartitioning it as observations become available. The method relies on approximating an optimal inference by using a probability distribution for quantized intervals of the unknown quantity, and by adjusting the intervals so as to obtain higher resolution in regions of higher probability, and vice versa.

We repartition the hypothesis space adaptively with the aim of maximizing the mutual information between the approximate distribution and the exact distribution. It is shown that this approach is equivalent to maximizing the entropy of the approximate distribution, and we provide low-complexity algorithms for approximating multi-dimensional posterior distributions with tunable complexity/performance.

The resulting quantized distribution for a one-dimensional case can be visualized as a histogram where each bar has equal area, but in general unequal width. The method can be used to provide adaptive quantization of arbitrary data sequences, or to approximate the posterior expectation of for instance some loss function by summing over a pre-specified number of terms.

Keywords: Bayesian Inference, Information Theory, Approximation, Quantization
PACS: 02.50.Tt, 89.70.+c, 02.50.Cw

1. INTRODUCTION

In this paper we present a method for approximate Bayesian inference in cases where the unknown quantities vary in an unknown manner but where the outcomes at different times are observed.

The method relies on approximating an optimal inference by using a probability distribution for quantized intervals of the unknown quantity, and by adapting the quantization so as to obtain higher resolution in regions of higher probability. The probability distribution is partitioned into K bins. After a block of data is observed, the posterior probability for each bin is computed by the use of Laplace's rule of succession. The total probability in each bin is then spread out uniformly over the individual values within the bin. Based on this posterior probability distribution, the boundaries of the K bins are adjusted so as to maximize the mutual information between the quantized distribution[1] and the unquantized distribution. As we shall see, this approach is equivalent to max-

[1] In this paper, whenever we speak of a quantized distribution we really mean a continuous-valued distribution over discrete intervals of the variable of interest. It is not the probabilities that are quantized, but rather the variables for which the probability is calculated.

imizing the entropy of the approximate distribution. For the one-dimensional case, the resulting quantized distribution can be regarded as a histogram with K bars of equal area, but in general of unequal width. Using this strategy, the posterior quantized distribution will increase the resolution in regions of high probability and decrease it in low-intensity regions.

In the following example taken from mobile communications we provide a motivating application for the method.

Example 1. *Consider the problem of scheduling transmissions to users in a mobile communications system. A controller wishes to schedule the use of a communications channel for T time slots ahead, but then faces the problem that the channel quality and the arrival rates into each buffer is unknown. Focusing here only on the arrival rates, a possible approach for handling the uncertainty regarding the number of bits entering the buffer would be to assign a probability distribution based only on the maximum entropy principle, as discussed in [1]. For instance, this is a valid approach if the controller has information about the average arrival rate in each buffer. However, as time evolves the controller can monitor the arrival rates and thus learn any patterns in the arrival rates by the use of Bayes' rule. Assuming that the statistics of the arrival rates do not change considerably during a certain period, we could use Laplace's rule of succession to obtain the probability p_k for an influx of size k bits,*

$$p_k = \frac{n_k + 1}{N + K} \ , \tag{1}$$

where n_k is the number of times over the N most recent observations that the influx consisted of k bits, and K is the number of possible influx sizes. But if the possible data rates vary over a large interval, say from 0 *bits/second to* 1 *megabit/second, K would be so large that the posterior distribution p_k would be uniform*[2] *for all practical purposes (since the number of observations N would then typically be much smaller than K).*

Instead, it could prove useful to partition the interval of possible influxes into a smaller set of regions, or bins, and apply the rule of succession on this smaller set of possibilities. For improved performance we should let the bin widths be adapted based on incoming data. Then the bins should spread out and become wide in regions where little activity is observed, and become denser in the rate interval of frequent observations. Thus, high fidelity is attained where it is suggested by the data, and less attention is paid to atypical rate regions. Within each bin, the probability for individual values is assigned by the principle of indifference. The expectation of any function of the arrival rates can then be obtained by a simple summation over the quantized posterior distribution and the function. □

[2] By uniform, we here refer to the fact that the majority of all possible outcomes will be equally likely, although the distribution will have occasional peaks. When we say that a distribution is close to uniform, we mean this in the sense that the entropy of the distribution is close to that of a uniform one (i.e. $\log K$).

2. MAXIMIZING THE MUTUAL INFORMATION

We here show that maximizing the mutual information between an approximate posterior distribution and an exact posterior is equivalent to maximizing the entropy of the approximate distribution. Let K be the number of bins to use in the approximation, and $i_{min} \leq i < i_{max}$ be the lower and upper bounds on the unquantized variable i between which we want to approximate $p(i \mid DI)$ (where D is the observed data and I our omnipresent background information). Denoting the mutual information between the quantized and the exact distributions $\mathscr{I}(k,i)$ and writing $p(k) = p(k \mid DI)$ for the posterior probability for obtaining an observation in bin k, and $p(i) = p(i \mid DI)$ for the posterior probability for obtaining the exact value i, we now prove the following theorem.

Theorem 1. *The optimum approximation to an exact distribution $p(i)$ for a quantity i, in terms of maximum mutual information between $p(i)$ and an approximate distribution $p(k)$ for quantized intervals (bins) k of the same underlying variable, is obtained when the bin boundaries of the latter distribution are adjusted so that the resulting distribution for k has maximum entropy.*

Proof. The mutual information between the distribution for the quantized variable k and the distribution for the unquantized variable i is given by

$$\begin{aligned}
\mathscr{I}(k,i) &= H(k) - H(k \mid i) && (2)\\
&= \sum_{k=1}^{K} \int_{i_{min}}^{i_{max}} p(ik) \log p(k \mid i) di - \sum_{k=1}^{K} p(k) \log p(k) && (3)\\
&= \sum_{k=1}^{K} \int_{i_{min}}^{i_{max}} p(ik) \log p(k \mid i) di - \sum_{k=1}^{K} \int_{i_{min}}^{i_{max}} p(i \mid k) p(k) \log p(k) di && (4)\\
&= \sum_{k=1}^{K} \int_{i_{min}}^{i_{max}} p(ik) \log \frac{p(k \mid i)}{p(k)} di = -\sum_{k=1}^{K} \int_{i \in \text{bin } k} p(i \mid k) p(k) \log p(k) di && (5)\\
&= -\sum_{k=1}^{K} p(k) \log p(k) \, , && (6)
\end{aligned}$$

where (4) follows from (3) by using the fact that $\int_{i_{min}}^{i_{max}} p(i \mid k) di = 1$ We obtain the second equality in (5) by noting that given knowledge of i we know in which bin k the observation lies, i.e. $p(k \mid i) = 1$ or $p(k \mid i) = 0$ depending on whether i is in bin k or not. Since $p(i \mid k)$ integrates to unity we finally have (6) from (5). The theorem can also be obtained directly from (2) by proving that $H(k \mid i) = 0$. (Given i, there is no uncertainty concerning which is the corresponding bin k.)

□

Thus, in order to obtain a quantized distribution which is as similar in information content to the unquantized distribution as possible, we should adjust the bin sizes to obtain equal probability mass in each bin.

3. ANOTHER RATIONALE FOR MAXIMUM ENTROPY DISTRIBUTIONS?

We phrased the theorem above in terms of an approximate distribution based on repartitioning the hypothesis space into a smaller number of intervals. But note that the proof only requires that there is a known mapping between k and i. It may be a 1-to-1 mapping, or an n-to-1 mapping from i to k, of any kind, and the result is still that we should use a maximum entropy distribution. If there is no ambiguity of which mapped value k that a given exact value i corresponds to (i.e. there is a deterministic mapping between the two), then the proof goes through.

This means that if we are to approximate a distribution with a simpler one (maybe using exactly the same hypothesis space, i.e. without quantization) under arbitrary constraints on the probability distribution, for instance in the form of known mean values, then the distribution which looses the least amount of information from the exact distribution is the one with maximum entropy subject to the imposed constraints. For example, wishing to approximate the distribution for a variable x using only first and second moments, we should use a Gaussian distribution regardless of the exact distribution's shape (and regardless of whether we know that distribution or not!).

Note that this provides an explicit motivation for using maximum entropy distributions: they approximate any known or unknown probability distribution with minimum information loss regardless of the 'underlying' distribution. We can thus *select* any testable information (any information which can be expressed as a function of the probability distribution) that we find convenient to work with and determine the 'best' approximation to *any* distribution (and again, it does not matter whether the exact distribution is known or unknown) by simply maximizing the entropy under the constraints imposed by this information.

The most common rationale for using maximum entropy distributions takes as its starting point that we know certain testable information, and then argues that by using the distribution with maximum entropy, we should add the fewest number of assumptions. Here, on the other hand, we see that it is also motivated, regardless of the reason for how we chose the constraints, to pick any subset of known testable information, use it in a maximum entropy assignment and thus obtain an approximate distribution which achieves minimum information loss under those constraints[3]

4. MAXIMIZING THE ENTROPY

Assume that we observe N samples of data before updating the bin sizes. In the multivariate case, each sample is a vector-valued observation specifying which bin the observation corresponds to. Within bin k we obtain n_k observations, and we have K bins in total. Assuming that the underlying causal mechanisms which determine the outcomes

[3] Of course, if there are relevant constraints that are not included in the approximation there may be a performance loss as compared to using the exact distribution. We should always strive for including the most relevant constraints.

are stationary over the N observations and the coming period of N observations, and ignoring possible time-dependencies, the probability for a future observation in bin k is

$$p_k = \frac{n_k + 1}{N + K} \tag{7}$$

according to Laplace's rule of succession (see Ch. 18 of [2] for an excellent discussion on this rule).

Now, how we choose to adjust the bin boundaries so as to obtain equal probability for all bins (and thus maximum entropy of the approximate distribution) depends on how flexible we allow the geometry of the bins to become. We will here prioritize simplicity over performance, and propose a method that at each step increases the entropy although it does not guarantee to actually maximize it. The algorithm works by halving bins in the *split* step, and combining two neighboring bins in the *expansion* step.

Split. In an n-dimensional case, the bin with the highest probability is halved in the one of the n directions that has the steepest probability gradient. In a 2-dimensional case, the bin can thus be halved either vertically or horizontally.

Expansion. The expansion step, the combination of one bin with a neighboring one, is slightly more complicated. In order to facilitate this step, we will use a multidimensional tree representation of bins and their probabilities. The example in Figure 1 shows the principle for the one-dimensional case, and Figure 2 shows the 2-dimensional case. Higher dimensions follow in a similar way. Each leaf represents a bin and has a probability attached to it as well as the bin boundaries. Since the bins are all obtained by halving the bin represented by the node directly above it in the tree, the bin boundaries do not have to be stored but can be computed from knowledge of the depth of the corresponding leaf in the tree and the minimum and maximum possible values that the variable can take in any direction. We only allow leaves connected to the same node at the preceding depth to be combined into one bin. This simplifies the algorithm, but could in principle be avoided for higher flexibility.

Only bins, i.e. leaves, at the same depth and connected to the same node are allowed to be combined. Out of the possible bins to be combined, the pair with the lowest sum probability is chosen. If there are several candidates, the pair with lowest probability gradient is chosen.

A split is always followed by an expansion to guarantee that the number of bins is held constant. The computational complexity and the rate of change can be adjusted by choosing the number of split-expansion steps to take after each block of observations. Full details of the algorithm cannot be given here due to lack of space.

5. COMPUTING POSTERIOR EXPECTATIONS

In order to compute the expectation of some function of the unknown quantities, we need to determine the probability for an individual value within an arbitrary bin k. Assume that the volume of bin k is w_k, i.e. the bin covers exactly a volume of that size of the underlying (vector-valued) quantity i. Then our task reduces to distributing

FIGURE 1. One-dimensional example of how a tree is built from a succession of splits. The left side shows an axis representing possible values of an unknown quantity divided into 2, 3, 4, and 5 bins. The right column depicts the corresponding tree. Since each leaf is obtained by halving the leaf above it, the bin widths can be computed from the depth of the leaf and the global minimum and maximum of the unknown quantity. Information about the probability for each bin must be stored with the corresponding leaf.

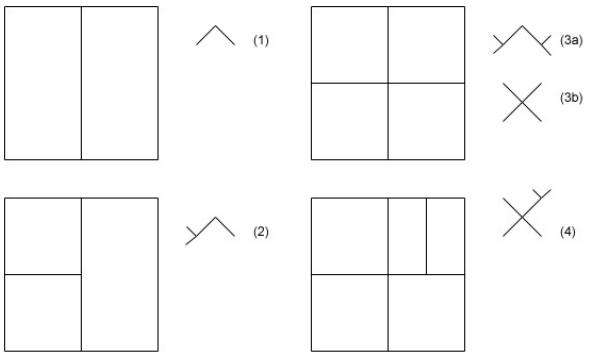

FIGURE 2. This shows how a tree is created in a 2-dimensional case. Here, the horizontal and vertical axes of the left sides correspond to 2 different variables. Each node in the tree now has the possibility to have up to four nodes, two for each dimension. (1) shows the tree when the horizontal part of the 2-dimensional plane is split into two halves. In (2), the left bin is further split in the vertical direction, generating two new leaves from the previous left leaf. In (3a), also the right leaf of depth 1 is halved in the vertical direction, creating two new leaves on that node. (3b) is an alternative representation of the same situation. (4) shows a horizontal split of the top right bin.

the probability p_k over the volume of size w_k. In order to assume anything else than a uniform distribution within the bin we would require some information which is not indifferent between the different values inside the volume. Here, we shall keep our solution general and therefore assume information indifference between the different values. Then the principle of indifference requires that we distribute the probability as

$$p_i = p_k / w_k \qquad i \in \text{bin } k \,. \tag{8}$$

Given the approximate posterior distribution p_k, what is the expectation of some function $f(\cdot)$ of the n-dimensional unquantized variable? The expectation of i given

the N most recent data is obtained *before* repartitioning the bins (because the statistics were collected based on the previous partition, not on the new one) by summing over all K bins the probability for that bin multiplied by the n-dimensional mean within that bin, i.e.

$$\langle i \rangle = \sum_{k=1}^{K} p_k \frac{i_{k\downarrow} + i_{k\uparrow}}{2} \tag{9}$$

where we define $i_{k\uparrow}$ and $i_{k\downarrow}$ as the (n-dimensional) upper and lower limit of bin k, respectively.

Similarly, the posterior expectation for an arbitrary function $f(i)$ is given by

$$\begin{aligned} \langle f(i) \rangle &= \sum_{k=1}^{K} p_k \int_{i_{k\downarrow}}^{i_{k\uparrow}} p(i \mid k) f(i) di \\ &= \sum_{k=1}^{K} \frac{p_k}{w_k} \int_{i_{k\downarrow}}^{i_{k\uparrow}} f(i) di \,, \end{aligned} \tag{10}$$

where the second equality was obtained by noting that $p(i \mid k) = 1/w_k$. (Given which bin we are in, each value within the bin is equally likely and has a probability equal to the inverse of the bin volume.)

6. EXAMPLE: A TWO-VALUED ALTERNATING SEQUENCE

We here study the performance of the proposed adaptive approximate inference for a one-dimensional case with $N = 100$ samples per block of data. In this case, the algorithm from Section 4 was not used. Our focus here is on the principles, not on the exact algorithm used. Thus, we repartitioned the bins according to an algorithm described in [3] (Ch. 8), sweeping through the entire data range once after each block and repartitioning the bin limits to attain close to maximum entropy after each block. For multidimensional cases, this algorithm is prohibitively complex since it has to sweep through all possible data values.

The data were generated so that each data block consists of 50 samples taking the value $i = 1$ and 50 samples of value $i = 7$, i.e. there are only two values and they occur with equal frequency. An approximate inference is carried out on the interval of integers between 0 and 100 using $K = 6$ bins, and an initial uniform partition over the integer interval 0...100. Figure 3 shows the probabilities for each bin after each of the first five updates. The resulting repartitioning of the bins was obtained as:

Block 1:	0	3	6	9	12	14	100
Block 2:	0	1	2	6	7	8	100
Block 3:	0	1	2	7	8	54	100
Block 4:	0	1	2	7	8	54	100
Block 5:	0	1	2	7	8	54	100

The bins quickly concentrate around $i = 1$ and $i = 7$, the only bins where any activity is registered, leaving larger implausible values nearly unattended. After the first update

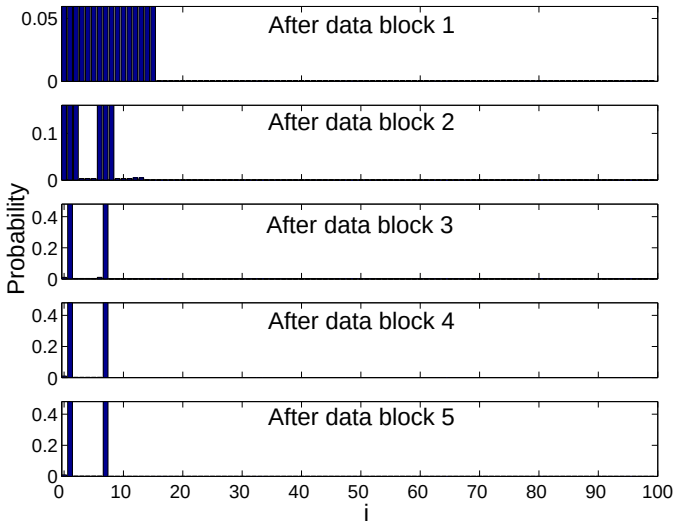

FIGURE 3. The evolution of the probabilities in each bin based on an adaptively quantized hypothesis space in an example where each block of $N = 100$ samples contained only two values, $i = 1$ and $i = 7$, occurring with exactly the same frequency. The convergence is quick and nearly all attention is focused around the two observed values.

the expectation of i becomes 9.9, after the second and the later updates the expectation is between 4 and 5, near the arithmetic mean $(7+1)/2 = 4$ of the sequence.

7. CONCLUSIONS

We have shown that when there is a known n-to-1 (for arbitrary n) mapping between some variable i and some other arbitrarily chosen variable k, the distribution for k that approximates the probability distribution for i with the least amount of information loss is the distribution that maximizes the entropy for k under any chosen constraints. We provided a simple adaptation rule for changing bin sizes in the general multidimensional case according to incoming data with tunable complexity and performance. Due to the limited space here, a performance study of this algorithm will be offered elsewhere. Likewise, the important problem of adapting not only the bin boundaries but also the number of bins will be investigated in future work. Finally, we would like to point out that time dependencies can be inferred by treating variables at different time delays as additional dimensions. All ideas given above are thus applicable also in the case of time-dependencies.

REFERENCES

1. M. Johansson and M. Sternad,"Resource allocation under uncertainty using the maximum entropy principle", *IEEE Transactions on Information Theory*, (To appear).
2. E. T. Jaynes, *Probability Theory - The Logic of Science*, Cambridge University Press, (2003).
3. M. Johansson, *Resource Allocation Under Uncertainty – Applications in Mobile Communications*, Ph.D. Thesis, Uppsala University, Signals and Systems Group, (2004).