Optimal Mutual Information Quantization is NP-complete

Brendan Mumey* Tomáš Gedeon†

February 20, 2003

Abstract

We consider the computational complexity of quantizing a joint (X, Y) probability distribution in order to maximize the mutual information of the quantization. We show that, in general, this problem is NP-complete via reductions from various forms of the PARTITION problem.

1 Introduction

In recent years, the problem of optimal quantization of the mutual information between two random variables has received increased attention. We note the application to discovery of neural coding scheme in early sensory system of cricket [3, 4] as well as the central role of quantization of mutual information in closely related optimization scheme named "Information bottleneck" [11, 12, 10, 9]. The latter method has been applied to data mining problems, coding theory and computational biology.

Given two discrete random variables X and Y the mutual information $I(X, Y)$ is defined by

$$I(X, Y) = \sum_{x,y} p(x, y) \lg \frac{p(x, y)}{p(x)p(y)}.$$

The mutual information is always nonnegative and achieves its minimal value 0 if the random variables X and Y are independent. The value $I(X, Y)$ tells us, roughly, how much we can learn X by observing Y, and vice versa.

In the application to neural coding, one models the input to the system as a random variable X and the output as a random variable Y. To simplify the situation, as well as to acknowledge inherent discreteness of collected data, X and Y are assumed to be discrete random variables.

Since noise is omnipresent in neural processing, sensory systems must be robust. As such, they must represent similar stimuli in a similar way and then react to similar internal representations of the outside stimuli in a consistent way. This leads to a conclusion that individual input and output patterns are not important for understanding neural function, but rather classes of input stimuli and classes of output patterns and their correspondence is the key. Therefore we are led to a problem of optimal assignment of input and output patterns to classes, where the optimality is judged on how much original mutual information is still present between the collection of classes.

More specifically, we will consider two different quantization methods. First is *one-sided quantization* and the other is *joint quantization*. We discuss briefly these two approaches.

In a one-sided quantization [2, 7] we quantize (cluster) only one variable, let us say Y, to a space of finitely many abstract classes, Y_N, where N represents the number of classes. This number is fixed. A particular

*Department of Computer Science, Montana State University, Bozeman, MT 59717-3880, USA
†Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717-0240, USA

quantization of Y to Y_N is determined by the choice of conditional probability $p(Y_N|Y)$, which then determines probability distribution of Y_N by

$$p(Y_N) = p(Y_N|Y)p(Y).$$

We would like to think of conditional probability $p(Y_N|Y)$ as a stochastic map. Every element $p(y_n|y)$ gives a probability that an element y belongs to class y_n. The collection $\{p(z|y)\}_{z\in Y_N, y\in Y}$ will be called a *quantizer*. To distinguish the quantizer from other probabilities in future calculations, we will use notation $q(y_n|y)$ instead of $p(y_n|y)$.

The problem of optimal one-sided quantization is to find quantizer $q(y_n|y)$ which minimizes the function

$$D_I(Y, Y_N) = I(X, Y) - I(X, Y_N).$$

Mutual information between X and quantized variable Y_N then captures maximum of the original mutual information between X and Y. Note that since $I(X, Y)$ does not depend on $q(y_n|y)$ this problem is equivalent to

$$\max_{q\in\Delta} I(X, Y_N), \tag{1}$$

where Δ is a constraint space representing the fact that collection $q(y_n|y)$ is a conditional probability, and

$$I(X, Y_N) = \sum_{x,y,y_n} q(y_n|y)p(x, y) \lg \frac{\sum_y q(y_n|y)p(x, y)}{p(x)\sum_y q(y_n|y)p(y)}.$$

The optimal quantizer $q(y_n|y)$ induces a coding scheme from $X \to Y_N$ by $p(y_n|x) = \sum_y q(y_n|y)p(y|x)$ which is the most informative approximation of the original relation $p(x|y)$ for a fixed size N of the reproduction Y_N.

In joint quantization [5] we fix sizes M and N of the reproduction spaces $X_M = \{x_1, \ldots, x_m\}$ for X and $Y_N = \{y_1, \ldots, y_n\}$ for Y and define joint quantization using two sets of quantizers, $q(x_m|x)$ and $q(y_n|y)$. The function to be minimized is

$$I(X, Y) - I(X_M, Y_N).$$

This is equivalent to the optimization problem

$$\max_{q(x_m|x)\in\Delta_M, q(y_n|y)\in\Delta_N} I(X_M, Y_N), \tag{2}$$

where

$$\begin{aligned} I(X_M, Y_N) &= \sum_{x_m, y_n} p(x_m, y_n) \lg \frac{p(x_m, y_n)}{p(x_m)p(y_n)} \\ &= \sum_{x_m, y_n, x, y} q(x_m|x)q(y_n|y)p(x, y) \lg \frac{\sum_{x,y} q(x_m|x)q(y_n|y)p(x, y)}{\sum_x q(x_m|x)p(x)\sum_y q(y_n|y)p(y)}. \end{aligned}$$

2 Results

2.1 Bounds on the Optimal Quantizer

It has been shown [7] that the optimal quantizer in one-sided quantization is generically deterministic, that is, the probability $q(y_n|y)$ is 1 or 0. In this case the responses associated with class z are $\mathcal{Y}_z = \{y|q(y_z|y) = 1\}$.

Lemma 2.1 *Consider a one-sided quantization into N classes, i.e. the optimization problem,*

$$max_{q\in\Delta} I(X, Y_N).$$

Assume $|X| > N$. Then $I(X, Y_N) \leq \lg N$.

Furthermore, if $p(x, y)$ is given by a diagonal matrix, then the maximum $I(X, Y_N)$ is achieved at a unique quantizer $q(y_n|y)$ with the property that

$$\sum_{q(y_j|y)=1} p(x, y) = \frac{1}{N}$$

for all classes y_j with $j = 1, \ldots, N$.

Proof. The first result follows from inequality [8, 1]

$$I(X, Y_N) \leq \min\{H(X), H(Y_N)\}, \tag{3}$$

the fact that we assume $|X| > N$ and $H(Y_N) \leq \lg N$, since Y_N is a discrete variable of N symbols.

Now the second part. The optimal quantizer is deterministic and so $q(y_j|y)$ is either zero or one for each class y_j and y. Therefore we can write without confusion for each y either $y \in y_j$ or $y \notin y_j$. Since $p(x, y)$ is diagonal, we can write $\bar{x} \in y_j$ if and only if $\bar{y} \in y_j$, where $\bar{y}$ is the unique element with $p(\bar{x}, \bar{y}) \neq 0$. Another consequence of diagonality of $p(x, y)$ is that

$$p(x, y_j) = \sum_y p(x, y)q(y_j|y) = p(x)$$

if $x \in y_j$ and $p(x, y_j) = 0$ if $x \notin y_j$. We compute

$$\begin{aligned}
I(X, Y_N) &= \sum_{y_N} \sum_x p(x, y_N) \lg \frac{p(x, y_N)}{p(x)p(y_N)} \\
&= \sum_j \sum_{x\in y_j} p(x, y_j) \lg \frac{p(x, y_j)}{p(x)p(y_j)} \\
&= \sum_j \sum_{x\in y_j} p(x) \lg \frac{1}{p(y_j)} \\
&= -\sum_j p(y_j) \lg p(y_j) \\
&= H(Y_N)
\end{aligned}$$

The entropy $H(Y_N)$ is a strictly concave function and has a unique maximum $p(y_j) = 1/N$ for all j. The value at the maximum is $\lg N$. This together with the above calculation gives

$$\sum_{y\in y_j} p(x, y) = \sum_{y\in y_j} p(x, y)q(1|y) = \sum_{x\in y_j} p(x, y_j) = p(y_j) = 1/N$$

Lemma 2.2 *Consider a joint quantization problem (equation 2) with $M = N = k \geq 2$ Then $I(X_k, Y_k) \leq \lg k$ bits with equality achieved only when there exists a quantizations $q(x_m|x)$ and $q(y_n|y)$ with a class labeling such that $p(x_i, y_i) = \frac{1}{k}$ and $p(x_i, y_j) = 0$ for $i \neq j$.*

Proof:

We use the estimate (3) again to get $\max I(X_k, Y_k) \leq \min(H(X_k), H(Y_k))$. Since $I(X_k, Y_k) = H(X_k) - H(Y_k|X_k) = H(Y_k) - H(X_k|Y_k)$ the equality is achieved when $H(Y_k|X_k) = H(X_k|Y_k) = 0$ and $H(X_k) = H(Y_k) = \lg k$. A random variable with k outcomes achieves a maximum entropy of $\lg k$ bits only if each outcome is equally likely and so $p(x_m) = p(y_n) = \frac{1}{k}$. Substituting these equalities into (3) yields

$$I(X_k, Y_k) = \sum_{m,n} p(x_m, y_n) \lg[k^2 p(x_m, y_n)] = 2 \lg k + \sum_m \sum_n p(x_m, y_n) \lg p(x_m, y_n),$$

since $\sum_{m,n} p(x_m, y_n) = 1$. Given the constraints $\sum_n p(x_m, y_n) = \frac{1}{k}$ and the fact that $x \lg x$ is convex, the inner sum is maximized exactly when $p(x_m, y_n) = \frac{1}{k}$ for one value of n and the remaining probabilities are 0. Given the constraints $\sum_m p(x_m, y_n) = \frac{1}{k}$ it follows that if $p(x_m, y_n) = \frac{1}{k}$, then $p(x_i, y_n) = p(x_m, y_j) = 0$ for all $i \neq m$ and $j \neq m$. By permuting the class labels appropriately the lemma is proven.

2.2 Decision Problems for Quantization

We define the following parametrized versions of the quantization problem:

N-QUANT: Given a joint (X, Y) probability distribution, decide if there exists a quantization Y_N of Y with mutual information at least s.

$M \cdot N$-QUANT: Given a joint (X, Y) probability distribution, decide if there exists a joint product quantization (X_M, Y_N) with mutual information at least s.

Recall the PARTITION problem: Given a set of real numbers $\{r_1, r_2, \ldots, r_n\}$, decide if there is a set of indices S such that $\sum_{i \in S} r_i = \sum_{i \notin S} r_i$. PARTITION is a well known NP-complete problem [6]. The general problem of finding a balanced k-way partition is also NP-complete. We call this generalization the k-PARTITION problem: Find index sets $S_1, \ldots S_k$ such that for any $u \neq v$, $\sum_{i \in S_u} r_i = \sum_{i \in S_v} r_i$.

Theorem 2.3 *$k \cdot k$-QUANT and k-QUANT problems are NP-complete for $k \geq 2$.*

Proof:

We prove the statement by giving a reduction from k-PARTITION to both $k \cdot k$-QUANT and k-QUANT.

Let $R = \{r_1, r_2, \ldots, r_n\}$ be an instance of k-PARTITION. Let $r = \sum r_i$. We consider the joint diagonal distributions of the form

$$p(X, Y) = \text{diag}(r_1/r, r_2/r, \ldots, r_n/r) \tag{4}$$

(note $|X| = |Y| = n$, non-diagonal entries are 0). We choose $s = \lg k$. We will show that R can be k-partitioned if and only there exists a joint quantization (or a one-sided quantization) of $p(X, Y)$ with $\lg k$ bits of mutual information. The only if direction is easy: Suppose there are index sets $S_1, \ldots S_k$ such that for any $u \neq v$, $\sum_{i \in S_u} r_i = \sum_{i \in S_v} r_i$. For one-sided quantization, we define classes y_u by $y_i \in y_u$ if and onli if $i \in S_u$. For joint quantization we define, in addition, classes x_u by $x_i \in x_u$ if and only if $i \in S_u$. Clearly, $p(x_u) = p(y_u) = \frac{1}{k}$ and thus $I(X_k, Y_k) = \lg k$ and $I(X, Y_k) = \lg k$. For the other direction we first consider joint quantization. Suppose that (X_k, Y_k) is a joint quantization that achieves a mutual information of $\lg k$ bits. By Lemma 2.2, there exists such that $p(x_u, y_u) = \frac{1}{k}$ and $p(x_u, y_v) = 0$ for $u \neq v$. Since the underlying (X, Y) distribution is diagonal, there is a permutation of labels of classes such that $x_i \in x_u$ if and only if $y_i \in y_u$, for all classes u. We define index set S_u by setting $i \in S_u$ if and only if $x_i \in x_u$ and $y_i \in y_u$. Then $\sum_{i \in S_u} r_i/r = p(x_u, y_u) = 1/k$ for all classes u. Multiplying both sides of this equation by r shows that S yields the desired partition of R. Now we consider one sided quantization. Suppose that Y_k is a one-sided quantization that achieves a mutual information of $\lg k$ bits. By Lemma 2.1, there exists classes y_u, $u = 1, \ldots, k$, such that $\sum_{y \in y_u} p(x, y) = \frac{1}{k}$. We define index set S_u by setting $i \in S_u$ if and only if $y_i \in y_u$.

Then $\sum_{i \in S_u} r_i / r = \sum_{y \in y_u} p(x, y) = 1/k$. Multiplying both sides of this equation by r shows that S yields the desired partition of R.

Corollary 2.4 *$M \cdot N$-QUANT is NP-complete for $m, n \geq 2$.*

Proof:

Suppose $M < N$ (the case $M > N$ can argued symmetrically). The $M \cdot M$ quantization problem used in the above lemma can be encoded in an $M \cdot N$ quantization problem by adding $N - M$ "dummy" outcomes y_k to Y that occur with probability 0. As above, the maximum possible mutual information is $\lg M$ bits which is achievable if and only if there is a balanced partition of X into M classes of equal probability. By adding the dummy outcomes, we guarantee a trivial way of including $N - M$ new Y classes to the remaining M classes in Y_N that are identical to the classes in X_M.

3 Discussion

Mention ideas for PTAS, etc. Implications for existing algorithms...

References

[1] Thomas Cover and Jay Thomas. *Elements of Information Theory*. Wiley Series in Communication, New York, 1991.

[2] Alexander G. Dimitrov and John P. Miller. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

[3] Alexander G. Dimitrov, John P. Miller, Zane Aldworth, and Tomáš Gedeon. Non-uniform quantization of neural spike sequences through an information distortion measure. In James Bower, editor, *Computational Neuroscience: Trends in Research 2001*. Elsevier, 2001.

[4] Alexander G. Dimitrov, John P. Miller, Tomáš Gedeon, Zane Aldworth, and Albert Parker. Analysis of neural coding using quantization with an information-based distortion measure. *Network: Computation in Neural Systems*, 2003.

[5] Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. In *Proc. Seventeenth Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2001.

[6] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the theory of NP-completeness*. W. H. Freeman and Co., 1979.

[7] Tomáš Gedeon, Albert Parker, and Alexander G. Dimitrov. Information distortion and neural coding. *Canadian Applied Math. Q.*, 2003.

[8] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[9] Elad Schneidman, Noam Slonim, Naftali Tishby, Rob R. de Ruyter van Steveninck, and William Bialek. Analyzing neural codes using the information bottleneck method. *preprint*, 2001.

[10] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 617–623. MIT Press, 2000.

[11] Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. In *Proceedings of The 37th annual Allerton conference on communication, control and computing*. University of Illinios, 1999.

[12] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. LANL preprint, http://arXiv.org/abs/physics/0004057, 2000.