EXPLOITING AN INTERPLAY BETWEEN NORMS TO ANALYZE SCALAR QUANTIZATION SCHEMES

Parimal Parag, Jean-Francois Chamberland

Department of Electrical and Computer Engineering, Texas A&M University

ABSTRACT

Quantization is intrinsic to several data acquisition systems. This process is especially important in distributed settings, where observations must first be compressed before they are disseminated. There have been many practical successes in the area of quantization, including the acclaimed Lloyd-Max algorithm. This article adopts a different perspective and it explores quantization at a fundamental level, seeking to identify classes of problems for which efficient quantization is possible. The focus is primarily on positive random variables of unbounded support, where severe degradation may occur. Established properties of Banach spaces are exploited, together with the boundedness of probability measures, to prove that efficient quantization schemes necessarily exist in the fine-quantization regime. The results are algorithmic in nature and provide bounds on the number of bits necessary to achieve a desired level of performance.

Index Terms— Quantization, Norm

1. INTRODUCTION

Remote sensing and distributed information systems are becoming increasingly popular as means of gathering pertinent data. Potential applications for these devices include detection, estimation and control. Instrumental to the success of such systems are efficient methods to process, compress and transmit information. In particular, quantization plays a central role in the design of many information systems. The literature on quantization is vast, and it contains a multitude of schemes tailored to different application scenarios [1]. The celebrated Lloyd-Max algorithm and its many variants point to the importance of quantization in practical settings.

A prevalent approach in the design of quantizers is to begin with a simple function and then improve performance through a two-step iterative procedure, alternating between quantization regions and their representatives. Another possible line of research is to study the properties of a class of quantizers such as uniform quantizers [2]. Interesting results have also been reported on the existence of universal quantizers for families of random variables with bounded support [3].

An important characteristic underlying many previous contributions in this area is that, for problems of statistical inference, most of the information present in an observation appears to be contained within the first few bits of quantized data [4]. In this article, we revisit scalar quantization and focus on the basic properties of quantized signals. Our goal is to offer new insights about the quantization process, thereby providing further evidence that efficient quantization is possible for large classes of distributions. Furthermore, we derive bounds on the number of quantization bits needed to achieve a desired level of performance. We are especially interested in the quantization of random variables with unbounded support, as it can lead to detrimental errors. Such situations arise, for instance, in decentralized detection where agents must transmit compressed versions of empirical log-likelihood ratios.

The motivation for this article can be explained as follows. When considering distributed systems, analysis is often greatly simplified by assuming that agents can exchange raw observations. While this may not be true in practice, if efficient quantization is possible then this approximation is justified. This is especially pertinent for packetized systems where headers are substantial; in such situations, it is reasonable to use a correspondingly large number of quantization bits per observation in the payload.

2. QUANTIZATION AND SIMPLE FUNCTIONS

Suppose X is a non-negative random variable on space $(\Omega, \mathcal{F}, P)$ and assume that X is in L_1. Then, we can write the mean of X as

$$\mathrm{E}[X] = \int_{\Omega} X(\omega) dP(\omega) = \|X\|_1 < \infty.$$

Alternatively, let μ represent the probability measure induced by X on $\mathbb{R}$, with $\mu(S) = \Pr(X^{-1}(S))$. We can express the expected value of X as $\mathrm{E}[X] = \int x d\mu(x) = \|X\|_1$.

A standard argument from analysis states that X is the pointwise limit of a monotonic increasing sequence of non-negative simple functions [5]. One possible construction for this argument is reviewed below. For $b > 0$, we can partition the range of X into 2^{2b} intervals. The first $2^{2b} - 1$ intervals have length 2^{-b} and are given by

$$A_{b,k} = \left\{ \omega \in \Omega : X \in \left[\frac{k-1}{2^b}, \frac{k}{2^b} \right) \right\} \quad k = 1, \ldots, 2^{2b} - 1.$$

The last interval is $A_{b,2^b} = \left\{ \omega \in \Omega : X \in \left[2^b - 2^{-b}, \infty \right) \right\}$. Consider the functions defined by

$$X_{2b}(\omega) = \sum_{k=1}^{2^{2b}} \frac{k-1}{2^b} \mathbb{1}_{A_{b,k}}(\omega),$$

This material is based upon work supported, in part, by the National Science Foundation (NSF) under Grants No. 0747363. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect NSF's views.

then the sequence $\{X_{2b}\}_{b=1}^{\infty}$ converges pointwise to X as b tends to infinity. If X is bounded, the convergence is uniform.

Since X is in L_1, we can applying Lebesgue's dominated convergence theorem [5] to the approximation error $(X - X_{2b})$; this yields

$$\begin{aligned}\lim_{b\to\infty} \|X - X_{2b}\|_1 &= \lim_{b\to\infty} \int_0^{\infty} |X(\omega) - X_{2b}(\omega)| dP(\omega) \\ &= \int_0^{\infty} \lim_{b\to\infty} |X(\omega) - X_{2b}(\omega)| dP(\omega) = 0.\end{aligned}$$

By construction, X_{2b} admits a $2b$-bit representation. This fact implies that a non-negative random variable X with finite mean can be quantized to arbitrary precision with respect to the L_1-norm. This is very encouraging because it establishes the existence of a good quantization scheme for random variable X and distribution μ.

While we know that a sequence of simple quantization schemes can lead to arbitrary small error, the mathematical justification described above does not specify bounds on the number of bits necessary to achieve a prescribed level of accuracy. Indeed, for any given b, there exists a random variable X that produces a large residual error. In this work, we wish to derive a set of conditions that provides insight on how large b should be to achieve a quantization performance of the form $\|X - \tilde{X}\|_1 < \delta$, where $\tilde{X}$ is the quantized version of X. To accomplish this task, we exploit relations between norms and we construct an iterative, greedy quantization procedure.

Assume that X is a non-negative random variable in L_2. First, we note that X is also in L_1 because $L_1 \subseteq L_2$ for spaces with bounded measures. Thus, we have

$$\begin{aligned}\|X\|_1 &= \mathrm{E}[X] = \int_0^{\infty} x d\mu(x) < \infty \\ \|X\|_2^2 &= \mathrm{E}\left[X^2\right] = \int_0^{\infty} x^2 d\mu(x) < \infty.\end{aligned}$$

An alternative expression for $\mathrm{E}[X^p]$ is presented below. This latter expression is closely related to arguments found in the subsequent sections of this article.

Lemma 2.1: Suppose X is a non-negative random variable in L_p, where $p > 1$, then we can write

$$\mathrm{E}\left[X^p\right] = \int_0^{\infty} p x^{p-1} \Pr(X > x) dx.$$

Proof: We find a characterization of $\mathrm{E}[X^p]$ in terms of the complementary cumulative distribution function of X,

$$\begin{aligned}\mathrm{E}\left[X^p\right] &= \int_{\Omega} X^p dP = \int_{\Omega} \int_0^X p x^{p-1} dx dP \\ &= \int_{\Omega} \int_0^{\infty} p x^{p-1} \mathbb{1}_{\{X > x\}} dx dP \\ &= \int_0^{\infty} \int_{\Omega} p x^{p-1} \mathbb{1}_{\{X > x\}} dP dx \\ &= \int_0^{\infty} p x^{p-1} \Pr(X > x) dx.\end{aligned}$$

Under Fubini's theorem, the non-negativity of X is sufficient to warrant changing the order of integration. ■

Corollary 2.2: If X is a non-negative random variable in L_1, then $\|X\|_1 = \int \Pr(X > x) dx$.

3. ONE-BIT QUANTIZATION

From this point forward, we assume that X is a non-negative random variable in L_2. Consider the quantization function defined by

$$\mathcal{Q}_c(X) = \begin{cases} c & X > c \\ 0 & X \leq c. \end{cases}$$

This scheme admits a one-bit representation. The residual error associated with this function is

$$\begin{aligned}Y &= X - \mathcal{Q}_c(X) = X - c\mathbb{1}_{\{X > c\}} \\ &= X\mathbb{1}_{\{X \leq c\}} + (X - c)\mathbb{1}_{\{X > c\}}.\end{aligned}$$

We emphasize that error Y is non-negative and satisfies

$$\|Y\|_1 = \|X\|_1 - c \Pr(X > c). \tag{1}$$

Looking at this equation and the expression for $\|X\|_1$ given in Corollary 2.2, it appears that $c \Pr(X > c)$ is an important quantity in bounding the difference between $\|X\|_1$ and $\|Y\|_1$.

Using a parallel progression with respect to L_2, we obtain

$$\begin{aligned}\|Y\|_2^2 &= \|X\|_2^2 - c \int_c^{\infty} (2x - c) d\mu(x) \\ &\leq \|X\|_2^2 - c^2 \Pr(X > c).\end{aligned}$$

In this case, the function $c^2 \Pr(X > c)$ plays a major role in establishing a relation between $\|X\|_2^2$ and $\|Y\|_2^2$.

To obtain bounds on the L_1-norm of X, we introduce two chief quantities,

$$s_1 = \sup_{x>0} x \Pr(X > x), \quad s_2 = \sup_{x>0} x^2 \Pr(X > x). \tag{2}$$

It is valuable to first study properties of s_1 and s_2.

Proposition 3.1: Suppose X is a random variable in L_2 and let s_1, s_2 be as defined above. The following inequalities hold,

$$s_1 \leq \min\{\|X\|_1, \|X\|_2\}, \qquad s_2 \leq \|X\|_2^2.$$

Proof: We initiate this demonstration by examining the simplest inequality. For any $c \geq 0$, we have

$$\begin{aligned}c \Pr(X > c) &= c \int_c^{\infty} d\mu(x) \\ &\leq \int_c^{\infty} x d\mu(x) \leq \int_0^{\infty} x d\mu(x) = \|X\|_1.\end{aligned}$$

Thus $s_1 \leq \|X\|_1$, as desired. A similar derivation leads to the second inequality. For any $c \geq 0$, we can write

$$\begin{aligned} c^2 \Pr(X > c) &= c^2 \int_c^\infty d\mu(x) \\ &\leq \int_c^\infty x^2 d\mu(x) \leq \int_0^\infty x^2 d\mu(x) = \|X\|_2^2. \end{aligned}$$

Therefore, we gather that $s_2 \leq \|X\|_2^2$.

The last inequality is slightly more involved. First, we find an upper bound for $\|X\|_1$. For any $z > 0$, we can write

$$\begin{aligned} \|X\|_1 &= \int_0^\infty \Pr(X > x)dx \\ &= \int_0^z \Pr(X > x)dx + \int_z^\infty \Pr(X > x)dx \\ &\leq \int_0^z \Pr(X > x)dx + \frac{1}{z}\int_z^\infty x\Pr(X > x)dx \\ &= \int_0^z \left(1 - \frac{x}{z}\right)\Pr(X > x)dx + \frac{1}{z}\int_0^\infty x\Pr(X > x)dx \\ &\leq \int_0^z \left(1 - \frac{x}{z}\right)\min\left\{1, \frac{s_1}{x}\right\}dx + \frac{1}{2z}\|X\|_2^2. \end{aligned}$$

Suppose $s_1 > \|X\|_2$. Choosing $z = \|X\|_2$ above, we get

$$\|X\|_1 \leq \int_0^z \left(1 - \frac{x}{z}\right)dx + \frac{1}{2z}\|X\|_2^2 = \|X\|_2.$$

However, this leads to a contradiction because we have already established that $s_1 \leq \|X\|_1$. As such, we conclude that $s_1 \leq \|X\|_2$. This completes the proof of the proposition. ■

As a simple corollary to this proof, we obtain the following bound on the L_1-norm of X. In particular, if $\|X\|_2$ and s_1 are small, then $\|X\|_1$ must also be small.

Corollary 3.2: If X is a non-negative random variable in L_2, then the L_1-norm of X is bounded by

$$\|X\|_1 \leq s_1 \log\left(\frac{s_1^2 + \|X\|_2^2}{2s_1^2}\right) + s_1.$$

Proof: From our previous derivation, we have

$$\|X\|_1 \leq \int_0^z \left(1 - \frac{x}{z}\right)\min\left\{1, \frac{s_1}{x}\right\}dx + \frac{1}{2z}\|X\|_2^2.$$

and $s_1 \leq \|X\|_2$. Then, for $z \geq s_1$, we get

$$\begin{aligned} \|X\|_1 &\leq \int_0^{s_1} \left(1 - \frac{x}{z}\right)dx + \int_{s_1}^z \left(1 - \frac{x}{z}\right)\frac{s_1}{x}dx + \frac{\|X\|_2^2}{2z} \\ &= s_1 \log z - s_1 \log s_1 + \frac{s_1^2 + \|X\|_2^2}{2z}. \end{aligned}$$

This bounding family is parametrized by z and it is valid over $[s_1, \infty)$. It therefore holds for $z = (s_1^2 + \|X\|_2^2)/2s_1$. The corollary is obtained by substituting this optimal value for z in the expression above. ■

The next set of results presented below hints at the fact that quantization offers diminishing returns as the number of levels increases.

Lemma 3.3: Suppose X is a non-negative random variable in L_2. Let $Y = X - c\mathbb{1}_{\{X>c\}}$ for some $c > 0$. Then, $s_1 \geq t_1$ and $s_2 \geq t_2$, where

$$t_1 = \sup_{x>0} x\Pr(Y > x), \qquad t_2 = \sup_{x>0} x^2 \Pr(Y > x)$$

and s_1, s_2 are as defined in (2).

Proof: For any $\omega \in \Omega$, we have $X(\omega) \geq Y(\omega)$. This implies that $\Pr(X > x) \geq \Pr(Y > x)$ and, consequently,

$$\begin{aligned} x\Pr(X > x) &\geq x\Pr(Y > x) \\ x^2\Pr(X > x) &\geq x^2\Pr(Y > x). \end{aligned}$$

We immediately conclude that $s_1 \geq t_1$ and $s_2 \geq t_2$. ■

We pursue our analysis by deriving various bounds for $\|X\|_1$ in terms of s_1, s_2 and $\|X\|_2$. These upper bounds can subsequently be employed to provide performance guarantees in terms of the number of bits required to achieve a certain performance. The first step in establishing these results is to provide a bound for $\Pr(X > x)$.

Lemma 3.4: Assume X is non-negative and in L_2. Let s_1, s_2 be as defined above. For any $x > 0$, we can write

$$\Pr(X > x) \leq \min\left\{1, \frac{s_1}{x}, \frac{s_2}{x^2}\right\}. \qquad (3)$$

Proof: The first component of the inequality is trivial, as $\Pr(X > x) \leq 1$. To obtain the second component, it suffices to notice that $x\Pr(X > x) \leq s_1$ for any $x > 0$. Dividing both sides of this inequality by x leads to the desired result. The third component is obtained in a similar manner. Since $x^2\Pr(X > x) \leq s_2$, we gather that $\Pr(X > x) \leq s_2/x^2$ for any $x > 0$. Collecting these results, we get (3). ■

Proposition 3.5: If X is a non-negative random variable in L_2, then $\|X\|_1 \leq 2\sqrt{s_2}$.

Proof: Using Lemma 2.1 and Lemma 3.4, we can write

$$\begin{aligned} \|X\|_1 &= \int_0^\infty \Pr(X > x)dx \leq \int_0^\infty \min\left\{1, \frac{s_2}{x^2}\right\}dx \\ &= \int_0^{\sqrt{s_2}} dx + \int_{\sqrt{s_2}}^\infty \frac{s_2}{x^2}dx = 2\sqrt{s_2}. \end{aligned}$$

The inequality holds because neglecting one of the arguments of the minimization can only lead to a looser upper bound. ■

A different bound can be derived by accounting for the third argument in (3). This, on the other hand, leads to an additional constraint.

Proposition 3.6: If X is a non-negative random variable in L_2 and $s_1^2 < s_2$, then

$$\|X\|_1 \leq 2s_1 - 2s_1 \log s_1 + s_1 \log(s_2).$$

Proof: Applying Lemma 2.1 and Lemma 3.4, we have

$$\begin{aligned} \|X\|_1 &= \int_0^\infty \Pr(X > x)dx \leq \int_0^\infty \min\left\{1, \frac{s_1}{x}, \frac{s_2}{x^2}\right\}dx \\ &= \int_0^{s_1} dx + \int_{s_1}^{s_2/s_1} \frac{s_1}{x}dx + \int_{s_2/s_1}^\infty \frac{s_2}{x^2}dx \\ &= 2s_1 - 2s_1 \log s_1 + s_1 \log s_2. \end{aligned}$$

We note that, when $s_1^2 < s_2$, this bound is tighter than the one contained in Proposition 3.5. ■

4. GREEDY QUANTIZATION PROCEDURE

In this section, we seek to derive a bound on the number of bits necessary to approximate random variable X such that the L_1-norm of the residual quantization error is less than $\delta > 0$. To achieve this goal, we employ an iterative, greedy algorithm that reduces the norm of the quantization error at every step.

Consider the following procedure. First, we let $\epsilon_1 > 0$ and $\epsilon_2 = \delta^2/4$. We initialize the algorithm with $X_0 = X$. At time instant i, if $\sup x \Pr(X_i > x) > \epsilon_1$, then we select a value of c_i such that $c_i \Pr(X_i > c_i) > \epsilon_1$. Also, we define the binary random variable b_i by

$$b_i = \mathbb{1}_{\{X_i > c_i\}}. \tag{4}$$

We can write the residual quantization error recursively,

$$X_{i+1} = X_i - c_i \mathbb{1}_{\{X_i > c_i\}} = X_i - b_i c_i. \tag{5}$$

From (1), we gather that $\|X_{i+1}\|_1 \leq \|X_i\|_1 - \epsilon_1$. When either $\|X_i\|_1 < \delta$ or $\sup x \Pr(X_i > x) \leq \epsilon_1$, this phase of the iteration process is terminated.

At this point, two cases are possible. If $\|X_i\|_1 < \delta$, then we have achieved our objective. On the other hand, if $\|X_i\|_1 \geq \delta$, then we start choosing thresholds as to reduce the L_2-norm of X_i. Specifically, if $\sup x^2 \Pr(X_i > x) > \epsilon_2$, then we pick a value of c_i such that $c_i^2 \Pr(X_i > c_i) > \epsilon_2$. For these steps, the binary digit b_i is still defined through (4) and the residual error is given by (5). This second phase of the iterative algorithm stops when $\|X_i\|_1 < \delta$.

Theorem 4.1: The algorithm described above will terminate in at most $M + N$ steps, where

$$M = \left\lfloor \frac{\|X\|_1}{\epsilon_1} \right\rfloor, \qquad N = \left\lfloor \frac{\|X\|_2^2}{\epsilon_2} \right\rfloor.$$

Proof: This result is somewhat straightforward, in view of our previous discussion. At every step during the first phase of this iteration procedure, the reduction in L_1-norm is at least ϵ_1. It follows that the maximum number of steps for this phase of the algorithm is M. Similarly, the second phase will end in at most N iterations, as the L_2-norm of the residual error decreases by at least ϵ_2 with every additional step. Note that if, at any point, there does not exist a value of c_i such that $c_i^2 \Pr(X_i > c_i) > \epsilon_2$ then

$$s_2 = \sup_{x>0} x^2 \Pr(X > x) \leq \delta^2/4$$

and $\|X_i\| < \delta$ by Proposition 3.5. Implicit to this argument is the fact that s_1 and s_2 are non-increasing over time. Collecting these facts, we conclude that the overall procedure must end in at most $M + N$ steps. ■

One of the irony associated with this procedure is that our bound is tightest when ϵ_1 is large. In the limit, as ϵ_1 tends to infinity, the upper bound simply becomes

$$N = \left\lfloor \frac{\|X\|_2^2}{\epsilon_2} \right\rfloor.$$

Yet from an engineering point of view, it makes sense to first try to quantize with respect to the L_1-norm; minimizing the L_1-norm of the quantization error is the intended goal of the algorithm. Unfortunately, we have very little insight to share on how to resolve this apparent dichotomy.

On the positive side, having a universal bound on the number of bits required to achieve a certain performance is a very valuable result. For completeness, we note that the value of the quantized observation is given by $\tilde{X} = \sum_i b_i c_i$. Also, we emphasize that the proposed quantization procedure relies on perfect knowledge of the distribution of X. Obtaining an accurate distribution for X may be a monumental undertaking. This task is intimately linked to system modeling [6], a topic beyond the scope of this article.

5. DISCUSSION

Quantization can be especially difficult for distributions with unbounded support, as this process may lead to very large quantization errors. However, we showed that if a nonnegative random variable belongs to L_2, then the L_1-norm of the quantization error is bounded. Furthermore, we proved that, under these circumstances, efficient quantization is possible; the number of bits required to achieved a desired level of performance is finite. Thus, this study promotes analysis frameworks where agents are theoretically able to exchange raw observations. This is especially fitting for problems where the objective criterion is continuous with respect to the topology induced by the L_1-norm and where data is transmitted using a packetized infrastructure. In this sense, this article provides new supporting evidence for a popular assumption that is often found in the literature.

6. REFERENCES

[1] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, October 1998.

[2] H. V. Poor, “Fine quantization in signal detection and estimation,” *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 960–972, September 1988.

[3] Z.-Q. Luo, “Universal decentralized estimation in a bandwidth constrained sensor network,” *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2210–2219, June 2005.

[4] J.-F. Chamberland and V. V. Veeravalli, “Decentralized detection in sensor networks,” *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407–416, February 2003.

[5] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1986.

[6] D. MacKay, *Information Theory, Inference and Learning Algorithms*. New York: Cambridge University Press, 2002.