Optimal One-Bit Quantization

Alessandro Magnani Arpita Ghosh Robert M. Gray [*]
Information Systems Laboratory, Stanford University
Stanford, CA 94305-9510
alem, arpitag, rmgray@stanford.edu

Abstract

We consider the problem of finding the optimal one-bit quantizer for symmetric source distributions, with the Euclidean norm as the measure of distortion. For fixed rate quantizers, we prove that for (symmetric) monotonically decreasing source distributions with ellipsoidal level curves, the centroids of the optimal 1-bit quantizer must lie on the major axis of the ellipsoids. Under the same assumptions on the source distribution, the centroids of the optimal one-bit *variable-rate* quantizer lie on one of the axes of the ellipsoid. If further, the source distribution $f(x)$ is log-concave in x, the optimal 1-bit fixed-rate quantizer is unique and symmetric about the origin. (The Gaussian is an example of a distribution that satisfies all these conditions.) Under a further set of conditions on the source distributions, we show that there is a threshold below which the optimal fixed rate and variable rate quantizer are the same.

1 Introduction

We consider the problem of describing the optimal one-bit quantizer for both the fixed rate and variable rate case. By one-bit quantizer, we mean a quantizer with two codewords. We restrict ourselves to the Euclidean distance as a measure of distortion.

Although the problem appears to be simple, it is in fact quite challenging, and counter-intuitive. For example, it is not true that if the source distribution is symmetric, there is an optimal 1-bit quantizer with symmetric centroids. The following source distribution (Figure 1) from [AW82] demonstrates this:

$$f(x) = \begin{cases} -|x|/3 + 5/12 & |x| < 1 \\ (7-|x|)/72 & 1 \leq |x| < 7 \\ 0 & 7 \leq |x| \end{cases} \tag{1}$$

For this distribution, the optimal centroids are located at -1 and 3, with a distortion of 2.61. In contrast, the optimal symmetric quantizer has a distortion of 2.74.

[*]This work is supported in part by a Stanford Graduate Fellowship.

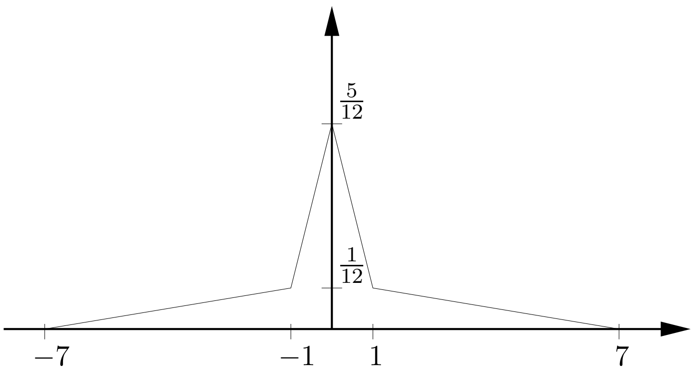

Figure 1: Optimal centroids need not be symmetric for symmetric distribution.

For a source distribution on $\mathbf{R}$, Kieffer [Kie83] showed that if the source density is log-concave, then (for the Euclidean distance), there is only one locally optimal k bit quantizer. This result applied for the specific case $k = 1$ guarantees the convergence of the Lloyd algorithm to the globally optimal 1-bit quantizer for a source with a log-concave density in $\mathbf{R}$.

The rest of the paper is organized as follows. In Section 2, we present some results for the one-bit quantization problem for symmetric distributions in higher dimensions. In Section 3, we present some results which relate the optimal fixed rate 1-bit quantizer to the optimal variable-rate 1-bit quantizer. In this case, variable rate merely means that the objective is a Lagrangian combining the distortion and the entropy of the code.

2 Fixed rate optimal quantizer

In this section, we first prove a condition on the centroids of an optimal (1-bit) quantizer for a symmetric source distribution. We then use this to prove a result for a specific class of distributions, which include, for example, the Gaussian and uniform distribution on an ellipsoid in $\mathbf{R}^n$.

Theorem 1. *For a source distribution $f : \mathbf{R}^n \to \mathbf{R}$ such that $f(x) = f(-x), x \in \mathbf{R}^n$, a necessary condition for codewords C_1 and C_2 to be optimal is that $C_1 = \alpha C_2$, $\alpha \in \mathbf{R}$.*

Proof. Consider a given pair of codewords C_1 and C_2; the Lloyd condition [GG92] implies that the boundary between the two codecells must be a hyperplane. Let this hyperplane be $a^T x = b$. Denote the region $a^T x > b$ by R_1, and $a^T x \leq b$ by R_2 (Figure 2). For the codewords to be Lloyd optimal (for the Euclidean distance), C_1 must be the centroid of R_1, and C_2 must be the centroid of R_2.

Construct the hyperplane $a^T x = -b$. Denote by R_3 the region $a^T x \leq -b$. By symmetry of f, we have

$$\int_{a^T x \leq -b} x f(x) dx = - \int_{a^T x > b} x f(x) dx,$$

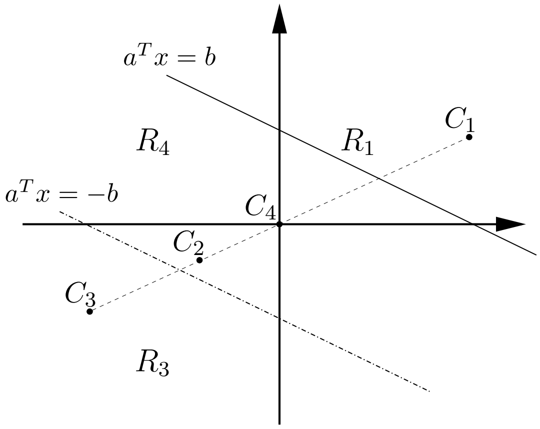

Figure 2: Line through the Lloyd optimal centroids passes through the origin.

i.e., the centroid of R_3, $C_3 = -C_1$. For the region $R_4 = \{x : a^T x \geq -b, a^T x \leq b\}$, by symmetry again, the centroid C_4 is 0. By construction C_2 must be a linear combination of C_3 and C_4. Therefore $C_2 = -\alpha C_3 + \beta C_4 = \alpha C_1$ for some $\alpha,\ \beta \in \mathbf{R}$.

■

We will use this result to prove the following theorem.

Theorem 2. *Consider a source with a distribution $f(x) = g(x^T \Sigma x)$, where $x \in \mathbf{R}^n$, $\Sigma \in \mathbf{R}^{n \times n}$ is a diagonal matrix with distinct non-negative entries, $g : \mathbf{R} \to \mathbf{R}$ is strictly decreasing, and $\int_{\mathbf{R}^n} |x| f(x) dx < +\infty$.*

The codewords for the optimal 1-bit quantizer for such a source lie on the axis corresponding to the largest Σ_{ii}.

Proof. Consider a pair of given codewords C_1 and C_2. From Theorem 1, the codewords must be of the form $C_1 = \alpha a,\ C_2 = \beta a$ where $\alpha,\ \beta \in \mathbf{R}$ and $\|a\| = 1$. The boundary between the regions is a separating hyperplane, given by $a^T x = b$, where we may assume without loss of generality that $b \geq 0$.

Change coordinates $x = Q\bar{x}$ where Q is orthogonal and satisfies $a = Qe_1$ (e_i is the vector with 1 in the ith position, and 0 everywhere else).

In addition we want to choose Q so that $(Q^T \Sigma Q)_{ij} = 0$ unless either $i = 1$, $j = 1$ or $i = j$. To do this let's call k_i for $i = 1, \ldots, n-1$, a set of orthonormal vectors which together with a form a basis for $\mathbf{R}^n$. Define $B = [k_1 \cdots k_{n-1}]$ and let $v_1, \cdots, v_{n-1}$ be a set of orthonormal eigenvectors of $B^T \Sigma B$. It is easy to check that if we set

$$Q = [a\ Bv_1 \cdots Bv_{n-1}],$$

then $Q^{-1} = Q^T$; also $(Q^T \Sigma Q)_{ij} = 0$ unless either $i = 1$, $j = 1$ or $i = j$ by construction.

For the pair of codewords to be Lloyd-optimal, the codewords must be centroids of their codecells. Specifically, then C_1 must be the centroid of the region $a^T x \geq b$. Therefore in the new set of coordinates we should have that

$$\frac{\int_{a^T \bar{x} > b} \bar{x}_i g(\bar{x}^T Q^T \Sigma Q \bar{x}) d\bar{x}}{\int_{a^T \bar{x} > b} g(\bar{x}^T Q^T \Sigma Q \bar{x}) d\bar{x}} = 0 \qquad (2)$$

for $i = 2, \ldots, n$. Since $\int_{\mathbf{R}^n} |x| f(x) dx < +\infty$, we can use Fubini's theorem to rewrite (2) as

$$\frac{\int_b^{+\infty} \left(\int_{\mathbf{R}^{n-1}} \bar{x}_i g(\bar{x}^T Q^T \Sigma Q \bar{x}) d\bar{x}_2 \cdots d\bar{x}_n \right) d\bar{x}_1}{\int_{a^T \bar{x} > b} g(\bar{x}^T Q^T \Sigma Q \bar{x}) d\bar{x}} = 0, \qquad (3)$$

for $i = 2, \ldots, n$.

Fix i and for $\bar{x} = [\bar{x}_1 \cdots \bar{x}_i \cdots \bar{x}_n]$ define $\tilde{x} = [\bar{x}_1 \cdots -\bar{x}_i \cdots \bar{x}_n]$. Given the choice of Q we have that

$$\bar{x}^T Q^T \Sigma Q \bar{x} - \tilde{x}^T Q^T \Sigma Q \tilde{x} = 4(Q^T \Sigma Q)_{1i} \bar{x}_1 \bar{x}_i.$$

Since over the range of integration $\bar{x}_1 > 0$ and g is monotonic, we have that

$$g(\bar{x}^T Q^T \Sigma Q \bar{x}) - g(\tilde{x}^T Q^T \Sigma Q \tilde{x})$$

is either strictly less or greater than 0 for every value of $\bar{x}_1 > 0$ and all $\bar{x}_j$ for $j = 2, \ldots, n$ unless $(Q^T \Sigma Q)_{1i} = 0$. Since

$$\int_b^{+\infty} \left(\int_{\mathbf{R}^{n-1}} \bar{x}_i g(\bar{x}^T Q^T \Sigma Q \bar{x}) d\bar{x}_2 \cdots d\bar{x}_n \right) d\bar{x}_1 =$$
$$\int_b^{+\infty} \int_0^{+\infty} \left(\int_{\mathbf{R}^{n-2}} \bar{x}_i (g(\bar{x}^T Q^T \Sigma Q \bar{x}) - g(\tilde{x}^T Q^T \Sigma Q \tilde{x})) d\bar{x}_2 \cdots d\bar{x}_{i-1} d\bar{x}_{i+1} \cdots d\bar{x}_n \right) d\bar{x}_i d\bar{x}_1,$$

if $(Q^T \Sigma Q)_{1i}$ is not zero, the ith component of the centroid is not 0 and therefore the Lloyd optimality condition is not satisfied. From this we conclude that the codewords can be optimal only if $(Q^T \Sigma Q)_{1i} = 0$ for $i = 2, \ldots, n$.

We will show, by contradiction, that such a Q can be found only if $a = e_i$.

Since

$$(Q^T \Sigma Q)_{1i} = a^T \Sigma B v_{i-1} = 0, \quad i = 2, \ldots, n,$$

and the v_i are a basis for $\mathbf{R}^{n-1}$, we must have $a^T \Sigma B = 0$.

Suppose that $a \neq e_i$. Then a must have at least two non-zero components, say a_1 and a_2. Choose $k_1 = [a_1 \quad -a_2 \ 0 \ldots 0]/\sqrt{a_1^2 + a_2^2}$. Then $(a^T \Sigma B)_1 = a^T \Sigma k_1 = a_1 a_2 (\Sigma_{11} - \Sigma_{22}) \neq 0$. So we have a contradiction, *i.e.*, $a = e_i$ necessarily. This shows that the codewords must lie on one of the unit vectors.

We will now proceed to show that in fact, the codewords lie on the unit vector corresponding to the largest Σ_{ii}.

Since the optimal codewords must be aligned with one of the axes, we find the optimal quantizer by finding the optimal codewords for each axis i and choosing the best of these n quantizers, *i.e.*, the one with the smallest distortion, where D_i is

$$D_i = \min_{c_1, c_2} \int (\min_{j=1,2} (x_i - c_j)^2 + x_1^2 + \ldots + x_{i-1}^2 + x_{i+1}^2 + \ldots x_n^2) g(x^T \Sigma x) dx_1 \ldots dx_n.$$

Change variables so that $\Sigma^{1/2} x = y$, the distortion is

$$D_i = \frac{1}{\det(\Sigma^{1/2})} \left(\int (\sum_{k=1}^{n} \frac{y_k^2}{\Sigma_{kk}} g(\|y\|^2) dy + \frac{1}{\Sigma_{ii}} \min_{c_1, c_2} \int \min_{j=1,2} (-2 y_i \Sigma_{ii}^{1/2} c_j + \Sigma_{ii} c_j^2) g(\|y\|^2) dy \right). \qquad (4)$$

Observe that when we replace i by j, the first term remains the same. Also, observe that

$$\min_{c_1,c_2}\int\min_{j=1,2}(-2y_i\Sigma_{ii}^{1/2}c_j+\Sigma_{ii}c_j^2)g(\|y\|^2)dy=\min_{c_1,c_2}\int\min_{j=1,2}(-2y_k\Sigma_{kk}^{1/2}c_j+\Sigma_{kk}c_j^2)g(\|y\|^2)dy. \tag{5}$$

From this, we can see that D_i is smallest for i such that Σ_{ii} is largest, *i.e.*, the codewords for the optimal quantizer lie along the axis corresponding to the largest Σ_{ii}. ■

We now prove the following corollary:

Corollary 1. *If, in addition, the density f is a log-concave function of x, the global optimal one-bit quantizer is unique and symmetric about the origin.*

Proof. By Theorem 2, the optimal quantizer lies along the axis corresponding to the largest Σ_{ii}. The problem of finding the codewords is now a one-dimensional quantization problem, since

$$\begin{aligned}D &= \min_{c_1,c_2}\int(\min_{i=1,2}(x_1-c_i)^2+x_2^2+\ldots+x_n^2)f(x)dx\\ &= \int(x_2^2+\ldots+x_n^2)f(x)dx+\min_{c_1,c_2}\int(\min_{i=1,2}(x_1-c_i)^2f_1(x_1)dx\end{aligned}$$

where $f_1(x_1)$ is the marginal of f along x_1.
Since the density f is log-concave, $f_1(x_1)=\int_{\mathbf{R}^{n-1}}f(x)dx_2\ldots dx_n$ is a log-concave function of x_1 (see, for example, [BV03]). Using the result of Kieffer [Kie83] we conclude that the globally optimal 1-bit quantizer is unique. Moreover since we know that along this axis there is local minimum symmetric about the origin we conclude that this is the global minimum. ■

Remarks:

- The matrix Σ need not be diagonal; all we need is that it be positive semidefinite and symmetric. In this case we can perform an eigenvalue decomposition of $\Sigma=Q^T\Lambda Q$, where Q is an orthogonal matrix, and change coordinates by Q to obtain the form in Theorem 2.

- Note also that the requirement that Σ has distinct (diagonal) entries is for convenience of proof. If Σ has two entries, say Σ_{ii} and Σ_{jj} which are equal, then it can be seen from the proof that the Lloyd condition can be satisfied also by codewords that are a linear combination of e_i and e_j. However, the distortion for these codewords will be the same, and in this case also, it is sufficient to check along the n axes e_i to find the globally optimal quantizer.

  The globally optimal quantizer is still found along a combination of th e_i associated with the biggest Σ_{ii}.

- The proof can be extended to decreasing functions. In fact the proof only requires that there are two non-trivial intervals I_1 and I_2 such that $g(I_1) > g(I_2)$. Since f has to be integrable this property comes directly from the fact that the function g is decreasing.

- It should be clear that g can also be strictly increasing on a ellipsoidal set and 0 outside of it.

3 Variable-rate 1-bit quantizer

Here we consider the problem of the optimal 1-bit variable rate quantizer. Recall that by variable rate, we mean that the objective function is now a linear combination of the distortion and the entropy of the code. Specifically, suppose Q is a quantizer, specified by (two) codewords and codecells. We will denote by $D_f(Q)$ the distortion for this quantizer (note that this is the objective function when we consider the fixed rate quantizer). The objective function for the variable rate quantizer is

$$D_\lambda(Q) = D_f(Q) + \lambda H(Q),$$

where $\lambda \geq 0$, and $H(Q)$ is the *entropy* corresponding to Q, *i.e.*, the entropy of the 1-bit random variable with probabilities computed from the codecells. Note that $0 \leq H(Q) \leq 1$.

Many of the results for fixed-rate quantizers can be extended to variable rate quantizers as well. In particular, we first state the following lemma, proved by Linder and Gyorgy in [GL03].

Lemma. *The boundary between the codecells of an optimal 1-bit variable-rate quantizer for an absolutely continuous source distribution is a hyperplane orthogonal to the line joining the codewords.*

In fact, it is enough for f to be absolutely continuous on its support, provided the support is a compact set. For the rest of this report we will assume that f satisfies this condition.

Note that for the variable-rate quantizer, the separating hyperplane need not pass through the midpoint of the two codewords.

Using the fact that the boundary is a separating hyperplane, we can conclude that the result of Theorem 1 holds also for variable rate quantizers and such source distributions, since given a codecell, the codeword must still be the centroid of the codecell for Lloyd optimality for the variable rate quantizer.

Since the result of Theorem 1 still holds, we can use the same arguments as in Theorem 2 to prove the following, slightly weaker result:

Theorem 3. *Consider a source with a distribution $f(x) = g(x^T \Sigma x)$, where $x \in \mathbf{R}^n$, $\Sigma \in \mathbf{R}^{n\times n}$ is a diagonal matrix with distinct non-negative entries, $g : \mathbf{R} \to \mathbf{R}$ is strictly decreasing, and $\int_{\mathbf{R}^n} |x| f(x) dx < +\infty$.*

The codewords for the optimal variable-rate 1-bit quantizer for such a source must lie on one of the coordinate axes.

We do not repeat the proof of the theorem. Note that in this case, we cannot conclude that the quantizer must lie on the axis corresponding to the largest Σ_{ii}, because of the additional entropy term in the objective function.

The optimal quantizers for the fixed-rate and the variable rate cases are related by a *threshold property*: Under certain conditions on f, there exists a λ^* such that for $\lambda \leq \lambda^*$, the fixed rate quantizer is also optimal for the variable rate quantizer. We make this precise in the following theorem.

Theorem 4. *Let f be a generic source distribution satisfying the following properties:*

- *There is exactly one globally optimal fixed-rate quantizer for f.*
- *The Hessian of the distortion D exists at the optimal, and $\nabla^2 D \succeq 0$.*
- *The gradient of the entropy is zero for the optimal fixed rate quantizer.*
- *The distortion at the global optimum is strictly less than the infimum over all locally (non-global) optimal quantizers.*

Then there exists λ^* such that for all $\lambda < \lambda^*$, the globally optimal fixed rate quantizer is also (globally) optimal for the variable-rate case.

Proof. A 1-bit variable-rate quantizer can be completely described by the position of the centroids (C_1, C_2) the normal a of the separating hyperplane and a point b on the hyper plane.

Let $x = [C_1, C_2]^T$ and $y = [a, b]^T$. We call the distortion for the variable rate $D_\lambda(x, y) = D(x, y) + \lambda H(y)$, where $D(x, y)$ is the fixed rate distortion. Let $[x^*, y^*]$ be the unique globally optimal fixed rate quantizer.

Since $\nabla D(x, y)|_{[x^*,y^*]} = 0$, $\nabla H(y)|_{y^*} = 0$, and $\nabla^2 D(x, y)|_{[x^*,y^*]} > 0$, we can find $\hat{\lambda}$ and δ such that for $0 < \|y^* - y\| < \delta$, all x and $\lambda < \hat{\lambda}$,

$$D_\lambda(x, y) > D_\lambda(x^*, y^*).$$

For $\|y^* - y\| > \delta$ and all x, let $D(x^*, y^*) - \inf(D(x, y)) = \beta$, $\beta < 0$. Let $\tilde{\lambda} < \beta$.

For $\|y^* - y\| > \delta$, $\lambda < \tilde{\lambda}$ and all x we have

$$D_\lambda(x, y) \geq D(x, y) > D(x^*, y^*) + \tilde{\lambda} > D_\lambda(x^*, y^*)$$

So by picking $\lambda^* = \min(\tilde{\lambda}, \hat{\lambda})$ the result follows. ■

Remarks.

- It appears that this proof can be extended to the optimal k-bit quantizer as well (the quantizer is now described by a larger number of parameters; suitably modify the vectors x and y).
- Similar results can be proved even with relaxing the hypothesis of a unique global minimum. For example

Theorem 5. *Let f be a generic source distribution satisfying the following properties:*

- *There is exactly one point b_0 such that for every a with norm 1 there exists a globally optimal quantizer with y of the form $y = (ab_0)$.*
- *The Hessian of the distortion D with respect to b exists for at all the globally optimal points, and is positive definite.*
- *The gradient of the entropy is zero for the optimal fixed rate quantizer.*
- *The distortion at a global optimum is strictly less than the infimum over all locally (non-global) optimal quantizers.*

Then there exists λ^* such that for all $\lambda < \lambda^*$, a globally optimal fixed rate quantizer is also (globally) optimal for the variable-rate case.

Proof. We call the distortion for the variable rate $D_\lambda(x,y) = D(x,y) + \lambda H(y)$, where $D(x,y)$ is the fixed rate distortion.

Let $[x^*, y^*]$ be a generic globally optimal fixed rate quantizer.

Since $\nabla D(x,y)|_{[x^*,y^*]} = 0$, $\nabla H(y)|_{y^*} = 0$, the Hessian with respect to b is positive definite, we can find $\hat{\lambda}$ and δ such that for all y with $0 < \|b_0 - b\| < \delta$, all x, and all $\lambda < \hat{\lambda}$,

$$D_\lambda(x,y) > D_\lambda(x^*, y^*).$$

For y such that $\|b_0 - b\| > \delta$ and all x, let $D(x^*, y^*) - \inf(D(x,y)) = \beta$, $\beta < 0$. Let $\tilde{\lambda} < \beta$.

For all y with $\|b_0 - b\| > \delta$, $\lambda < \tilde{\lambda}$ and all x we have

$$D_\lambda(x,y) \geq D(x,y) > D(x^*, y^*) + \tilde{\lambda} > D_\lambda(x^*, y^*)$$

So by picking $\lambda^* = \min(\tilde{\lambda}, \hat{\lambda})$ the result follows. ■

It appears from the proof that this result can be extended to a finite set of distinct globally optimal quantizers. We omit the proof and the statement for brevity.

References

[AW82] E. Abaya and G. Wise. Some remarks on optimal quantization. *Proc. Conference on Information Sciences and Systems*, March 1982.

[BV03] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003. Available at `www.stanford.edu/~boyd/cvxbook.html`.

[GG92] R. Gray and A. Gersho. *Vector Quantization and Signal Compression.* Kluwer, 1992.

[GL03] A. Gyorgy and T. Linder. Codecell convexity in optimal entropy-constrained vector quantization. *IEEE Transactions on Information Theory*, 49(7):1821–1828, July 2003.

[Kie83] J. C. Kieffer. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *IEEE Transactions on Information Theory*, IT-29(1):42–47, January 1983.