Jmspeex' Journal of Dubious Theoretical Results

July 15, 2020

Abstract

This is a log of theoretical calculations and approximations that are used in some of the Daala code. Some approximations are likely to be too coarse, some assumptions may not correspond to the observable universe and some calculations may just be plain wrong. You have been warned.

Part I
Relationship Between λ and Q in RDO

When using a high-rate scalar quantizer, the distortion is given by

$$D = \frac{Q^2}{12} ,$$

where Q is the quantizer's interval between two levels (not the maximum error like in some other work). The rate required to code the quantized values (assuming round-to-nearest) is

$$R = -\log_2 Q + C$$

where C is a constant that does not depend on Q. Starting from a known λ we want to find the quantization interval Q that minimizes the rate-distortion curve, so

$$\begin{aligned}
\frac{\partial}{\partial Q}\left(D + \lambda R\right) &= 0 \\
\frac{\partial}{\partial Q}\left(\frac{Q^2}{12} - \lambda \log_2 Q - \lambda C\right) &= 0 \\
\frac{Q}{6} - \frac{\lambda}{Q \log 2} &= 0 \\
Q &= \sqrt{\frac{6\lambda}{\log 2}}
\end{aligned}$$

Or, if Q is known, then

$$\lambda = \frac{Q^2 \log 2}{6}$$

Quantization threshold

When we have a value between 0 and 1 and consider whether to round up or down, we can compute the optimal decision threshold x for which the RD cost for the decision is equal

$$x^2 + \lambda R_0 = (1 - x)^2 + \lambda R_1 ,$$

where R_0 and R_1 are the costs for coding a zero and a one, respectively. Solving for x we have

$$\begin{aligned} x^2 + \lambda R_0 &= (1-x)^2 + \lambda R_1 \\ x^2 + \lambda R_0 &= x^2 - 2x + 1 + \lambda R_1 \\ 2x &= 1 + \lambda (R_1 - R_0) \\ x &= \frac{1}{2} + \frac{\lambda \Delta R}{2} , \end{aligned}$$

where $\Delta R = R_1 - R_0$. In other words, it's like round-to-nearest, but with an additional bias of $\lambda \Delta R/2$ towards zero.

Part II

Rate-Distortion Analysis of a Quantized Laplace Distribution

Here we assume that the quantization step size has already been taken into account and that σ is the normalized standard deviation of a DCT coefficient. The post-quantization distribution of a Laplace-distributed variable with non-zero quantization threshold θ is:

$$p(n) = \begin{cases} 1 - r^\theta & n = 0 \\ r^\theta (1-r) r^{n-1} & n > 0 \end{cases}$$

where $\theta = \frac{1}{2}$ for round-to-nearest and $r = e^{-\sqrt{2}/\sigma}$. The entropy (rate) R of the quantized Laplace distribution is:

$$R = \overbrace{r^\theta}^{\text{sign}} + \overbrace{H\left(r^\theta\right)}^{\text{non-zero}} + \overbrace{\frac{r^\theta H(r)}{1-r}}^{\text{tail}}$$

$$\begin{aligned} D(r) &= -\log r \left(\int_0^\theta x^2 r^x dx + \sum_{k=1}^{\infty} \int_{\theta-1}^{\theta} x^2 r^k r^x dx \right) \\ &= I(r,\theta) - I(r,0) + \frac{r}{1-r} \left(I(r,\theta) - I(r,\theta-1) \right) \end{aligned}$$

where

$$\begin{aligned} I(r,x) &= -\log r \int x^2 r^x dx \\ &= r^x \frac{2x \log r - x^2 \log^2 r - 2}{\log^2 r} + C \end{aligned}$$

When σ is much smaller than the quantization step size (everything quantizes to zero), then the distortion is simply σ^2 and when σ is very large (flat distribution), then the distortion is that of a scalar quantizer: $1/12$. So in the general case we can approximate with

$$D = \min\left(\sigma^2, \frac{1}{12}\right)$$

which tends to overestimate D in the region where σ^2 is close to $1/12$. Assuming high-rate RDO, we have

$$\lambda = \frac{Q^2 \log(2)}{6} = \frac{\log(2)}{6} .$$

The total RD-cost (expressed as a rate) becomes

$$R + \frac{D}{\lambda} = r^{\theta} + H\left(r^{\theta}\right) + \frac{r^{\theta} H\left(r\right)}{1-r} + \min\left(\frac{6\sigma^2}{\log(2)}, \frac{1}{2\log(2)}\right)$$

This cost function can be approximated by the (smoother) cost function

$$C = \frac{1}{2}\log_2\left(1 + (6.33\sigma)^2\right)$$

Part III
PVQ Distortion

Let $\mathbf{X} = g\mathbf{z}$ and $\hat{\mathbf{X}} = \hat{g}\hat{\mathbf{z}}$ be the unquantized and quantized coefficient vector, respectively. The quantization distortion is

$$\begin{aligned} D &= \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\mathbf{X} - \hat{\mathbf{X}}\right) \\ &= \mathbf{X}^T\mathbf{X} + \hat{\mathbf{X}}^T\hat{\mathbf{X}} - 2\hat{\mathbf{X}}^T\mathbf{X} \\ &= g^2 + \hat{g}^2 - 2g\hat{g}\mathbf{z}\hat{\mathbf{z}} \\ &= (g-\hat{g})^2 + 2g\hat{g} - 2g\hat{g}\mathbf{z}\hat{\mathbf{z}} \\ &= (g-\hat{g})^2 + g\hat{g}\left(2 - 2\mathbf{z}\hat{\mathbf{z}}\right) \ . \end{aligned} \tag{1}$$

Let D_z be the distance between $\mathbf{z}$ and $\hat{\mathbf{z}}$,

$$\begin{aligned} D_z &= (\mathbf{z} - \hat{\mathbf{z}})^T (\mathbf{z} - \hat{\mathbf{z}}) \\ &= 2 - 2\mathbf{z}^T\hat{\mathbf{z}} \ . \end{aligned} \tag{2}$$

We can then rewrite (1) as

$$D = (g-\hat{g})^2 + g\hat{g}D_z \ , \tag{3}$$

which separates the gain quantization from the quantization of the unit vector $\mathbf{z}$.

Part IV
Distortion from theta PVQ

Let the normalized theta-PVQ vector be

$$\mathbf{z} = \begin{bmatrix} \cos\theta \\ \mathbf{x}\sin\theta \end{bmatrix} \ ,$$

where $\mathbf{x}$ is the unit-vector coded with the PVQ quantizer, the distortion D between the unquantized $\mathbf{z}$ and its quantized version $\hat{\mathbf{z}}$ is

$$\begin{aligned} D_z &= (\mathbf{z} - \hat{\mathbf{z}})^T (\mathbf{z} - \hat{\mathbf{z}}) \\ &= \left(\cos\theta - \cos\hat{\theta}\right)^2 + \left(\mathbf{x}\sin\theta - \hat{\mathbf{x}}\sin\hat{\theta}\right)^T \left(\mathbf{x}\sin\theta - \hat{\mathbf{x}}\sin\hat{\theta}\right) \\ &= \cos^2\theta - 2\cos\theta\cos\hat{\theta} + \cos^2\hat{\theta} + \sin^2\theta + \sin^2\hat{\theta} - 2\sin\theta\sin\hat{\theta}\mathbf{x}^T\hat{\mathbf{x}} \\ &= 2 - 2\cos\theta\cos\hat{\theta} - 2\sin\theta\sin\hat{\theta}\mathbf{x}^T\hat{\mathbf{x}} \ . \end{aligned} \tag{4}$$

Using the identity (2), we can then rewrite (4) as

$$\begin{aligned} D_z &= 2 - 2\cos\theta\cos\hat{\theta} - \sin\theta\sin\hat{\theta}\,(2 - D_x) \\ &= 2 - 2\cos\left(\theta - \hat{\theta}\right) + \sin\theta\sin\hat{\theta}D_x \\ &= D_\theta + \sin\theta\sin\hat{\theta}D_x \ , \end{aligned} \tag{5}$$

where $D_\theta = 2 - 2\cos\left(\theta - \hat{\theta}\right)$ is the mean square error due to quantizing θ. So essentially, the total error is the sum of the error due to quantization of θ and the error in the PVQ quantization assuming a radius that's the geometric mean of the quantized and unquantized radius.

Putting (5) into (3), we obtain

$$D = (g - \hat{g})^2 + g\hat{g}\left(D_\theta + \sin\theta\sin\hat{\theta}D_x\right) \ . \tag{6}$$

Part V
Biorthogonality and quantization noise

A biorthogonal transform defined as

$$\mathbf{H}\mathbf{G}^T = I$$

where the columns of $\mathbf{G}$ are the analysis basis functions and the columns of $\mathbf{H}$ are the columns of the synthesis basis functions. We define the diagonal matrix $\mathbf{S}$ such that the diagonal elements $s_{i,i} = \sqrt{\mathbf{h}_i^T\mathbf{h}_i}$ are the magnitudes of the synthesis basis functions. For an input vector $\mathbf{x}$, the quantization process can be modeled as adding an uncorrelated noise vector $\mathbf{n}$ such that the reconstruction $\mathbf{y}$ is

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{S}^{-1}\left(\mathbf{S}\mathbf{G}^T\mathbf{x} + \mathbf{n}\right) \\ &= \mathbf{x} + \mathbf{H}\mathbf{S}^{-1}\mathbf{n} \end{aligned}$$

The distortion due to quantization is

$$\begin{aligned} D &= (\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y}) \\ &= \left(\mathbf{H}\mathbf{S}^{-1}\mathbf{n}\right)^T\mathbf{H}\mathbf{S}^{-1}\mathbf{n} \\ &= \mathbf{n}^T\left(\mathbf{S}^T\right)^{-1}\mathbf{H}^T\mathbf{H}\mathbf{S}^{-1}\mathbf{n} \\ &= \mathbf{n}^T\mathbf{R}\mathbf{n} \end{aligned}$$

where $\mathbf{R} = \left(\mathbf{S}^{-1}\right)^T\mathbf{H}^T\mathbf{H}\mathbf{S}^{-1}$. Rewriting D as a summation, we have

$$D = \sum_i \sum_j r_{i,i} n_i n_j$$

This is different from the orthonormal case where

$$D = \sum_i n_i^2$$

due to $\mathbf{R}$ being the identity matrix (also known as Parseval's theorem).

Since the noise is uncorrelated and since the diagonal of $\mathbf{R}$ is equal to 1, then the expectation of the noise power is $E\{D\} = E\{\mathbf{n}^T\mathbf{n}\}$, like in the orthonormal case. This shows that multiplying each transformed coefficient x_i by the magnitude of the corresponding synthesis function $s_{i,i}$ prior to quantization is sufficient to obtain the same *average* noise behaviour. In practice, it *may* be possible

to do some clever trick in the quantization search to obtain a smaller distortion than the orthonormal case, partially compensating for the entropy cost of the biorthogonal transform.

The average distortion we find also supports the use of the squared magnitude of the synthesis basis function in the computation of the coding gain.

Part VI
Temporal RDO

Let's assume two sequences of N samples each being quantized with a resolution Q. One sequence is constant, while the other isn't. The total distortion will be

$$D = \frac{2NQ^2}{12} .$$

If we assume that we can skip encoding of the constant sequence at no cost and shift b bits away from the variable sequence to the constant one, it costs only b/N bits per sample on the variable sequence and we get a distortion

$$\begin{aligned} D &= N\frac{\left(2^{-b}Q\right)^2}{12} + N\frac{\left(2^{b/N}Q\right)^2}{12} \\ &= \frac{NQ^2}{12}\left(2^{-2b} + 2^{2b/N}\right) \end{aligned}$$

We solve for $\partial D/\partial b = 0$ to minimize distortion:

$$\begin{aligned} \frac{\partial D}{\partial b} = \frac{NQ^2}{12}\left(-2b\log 2 2^{-2b} + \frac{2b\log 2}{N}2^{2b/N}\right) &= 0 \\ \frac{2b2^{2b/N}\log 2}{N} &= 2b2^{-2b}\log 2 \\ \frac{2^{2b/N}}{N} &= 2^{-2b} \end{aligned}$$

Taking the base-2 log on both side:

$$\begin{aligned} 2b/N - \log_2 N &= -2b \\ \frac{2b\,(N+1)}{N} &= \log_2 N \\ b &= \frac{N\log_2 N}{2\,(N+1)} \end{aligned}$$

Slowly varying sequence

Let's see what happens with a slowly varying sequence rather than a constant one...

Part VII
Motion Compensation RDO

Ideally, we'd want the MC RDO to consider the final rate and distortion after quantization. This is hard to do with PVQ, so let's do it by assuming a scalar quantizer instead. For each DCT coefficient

x_i there are two choices: either the coefficient quantizes to zero, or it doesn't. If it does, then the RD cost Z_i is

$$Z_i = x_i^2 + \lambda r_{i,0} \tag{7}$$

where $r_{i,0}$ is the rate for coding a zero. Assuming the coefficients are Laplace-distributed, we have

$$p(x_i) = K_i \exp\left(-\sqrt{2}\,|x_i|/\sigma_i\right) \tag{8}$$

where K_i is a normalization constant. Assuming a high rate, then the rate is

$$r_i = -\log_2 p(x_i) = r_{i,0} + \frac{\sqrt{2}\,|x_i|}{\sigma_i \log 2} \tag{9}$$

Considering that $r_{i,0} = -\log K_i/\log 2$ does not depend on the value x_i, it will be constant for all MV candidates, so it is safe to ignore it. The only term remaining is what the distortion should be for the case where the coefficient is encoded to a non-zero value. Since we don't want the distortion to oscillate, then a good choice is simply the average $Q^2/12$. So the RD-cost for a non-zero coefficient is

$$N_i = \frac{Q^2}{12} + \lambda \frac{\sqrt{2}\,|x_i|}{\sigma_i \log 2} \tag{10}$$

Also, since the ideal λ is approximately $Q^2 \log 2/6$, we can rewrite the above as

$$N_i = \lambda\left(\frac{\sqrt{2}\,|x_i|}{\sigma_i \log 2} + d\right) \tag{11}$$

where $d = 1/(2\log 2) = 0.72$, which we might want to tune.

The final RD-cost for a coefficient i is simply the min of the zero and non-zero case, so

$$C_i = \min\left[Z_i, N_i\right] \tag{12}$$

$$= \min\left[x_i^2, \lambda\left(\frac{\sqrt{2}\,|x_i|}{\sigma_i \log 2} + d\right)\right] \tag{13}$$

Considering an entire block, all we need to do is sum the cost of each coefficient, along with the cost of coding the motion vector, so

$$C_{total} = \sum_i C_i + \lambda rate_{mv} \tag{14}$$

The key equation here is (13). At high rate, the right side term N_i will dominate, but since it also includes a λfactor, the cost function should never have the effect of completely ignoring the cost of coding the MV, even at very high bitrate. On the other hand, at very low bit-rate when we skip very often, then the RDO cost tends toward plain MSE. The N_i term looks like SATD, except that each coefficient is weighted by the inverse of its standard deviation. This may be why pure SATD has been found to work well in many codecs.