Scalar quantization to a signed integer

Kalle Rutanen

March 4, 2009

1 Introduction

This paper discusses the scalar quantization of a real number range $[-1, 1]$ to a p-bit signed integer range $[-2^{p-1}, 2^{p-1} - 1]$. Our approach is to list a set of requirements for the quantization, and then find conversion functions that fulfill these requirements. Our result will be that the conversion functions to both directions are given by:

$$\begin{aligned} f(i) &= \mathrm{clamp}\left(\frac{i}{2^{p-1} - 0.5}, -1, 1\right) \\ g(x) &= \left[\mathrm{clamp}\left(x(2^{p-1} - 0.5) + \frac{\mathrm{sign}(x)}{2}, -(2^{p-1} - 1), 2^{p-1} - 1\right)\right] \end{aligned}$$

The notation will be as follows. If x is a real number, $[x]$ denotes the nearest integer towards zero (*truncation*). The set of integers is denoted by $\mathbb{Z}$, and the set of real numbers is denoted by $\mathbb{R}$. If a function f is defined from set A to set B, and $B' \subset B$, then $f^{-1}[B'] = \{a \in A : f(a) \in B'\}$. If A is a subset of $\mathbb{R}$, then its measure is denoted by $m(A)$. The cardinality of a set A is denoted by $|A|$.

$$\mathrm{sign}(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

2 General scalar quantization problem

We will base our approach to a more general problem of quantization between arbitrary intervals. Let

$$\begin{aligned} I &= [i_{\min}, i_{\max}] \subset \mathbb{Z}, \\ X &= [x_{\min}, x_{\max}] \subset \mathbb{R}, \\ \Delta x &= x_{\max} - x_{\min}, \\ \Delta i &= i_{\max} - i_{\min}. \end{aligned}$$

The problem is to define two functions f and g:

$$\begin{aligned} f: \quad I \to X: \quad x = f(i) \\ g: \quad X \to I: \quad i = g(x) \end{aligned}$$

such that they satisfy the following constraints:

1. The set $P = \{g^{-1}[\{i\}] : i \in I\}$ forms a partition of X.
2. Each set in P is an interval.
3. $\forall i \in I : g(f(i)) = i$.
4. The supremum error between $f(g(x))$ and x is minimal.
5. f preserves order.
6. g preserves order.
7. $g(x_{\min}) = i_{\min}$
8. $g(x_{\max}) = i_{\max}$

The error minimization requirement implies that the measures of all preimages of the singular subsets of I must equal some $C \in \mathbb{R}$. Because these measures must sum to $m(X)$, it must hold that:

$$|I|C = m(X).$$

Thus we can compute C by

$$C = \frac{m(X)}{|I|} = \frac{\Delta x}{\Delta i + 1}.$$

Furthermore, the error minimization implies that for all i $f(i)$ must map to center of the interval $g^{-1}[\{i\}]$. A function pair that fulfills the requirements is given by:

$$\begin{aligned}
f(i) &= x_{\min} + (i + 0.5 - i_{\min})C \\
&= x_{\min} + \frac{i + 0.5 - i_{\min}}{\Delta i + 1}\Delta x \\
g(x) &= i_{\min} + \operatorname{clamp}\left(\left\lfloor \frac{x - x_{\min}}{C} \right\rfloor, 0, \Delta i\right) \\
&= i_{\min} + \operatorname{clamp}\left(\left\lfloor \frac{x - x_{\min}}{\Delta x}(\Delta i + 1) \right\rfloor, 0, \Delta i\right)
\end{aligned}$$

The g function is well-known in the field of quantization and is called the *mid-rise* quantizer. Let us now prove that the requirements really are fulfilled.

Preservation of order

$$\begin{array}{rcl}
i & < & k \\
& \Leftrightarrow & \\
i + 0.5 - i_{\min} & < & k + 0.5 - i_{\min} \\
& \Leftrightarrow & \\
\dfrac{i + 0.5 - i_{\min}}{\Delta i + 1} & < & \dfrac{k + 0.5 - i_{\min}}{\Delta i + 1} \\
& \Leftrightarrow & \\
\dfrac{i + 0.5 - i_{\min}}{\Delta i + 1}\Delta x & < & \dfrac{k + 0.5 - i_{\min}}{\Delta i + 1}\Delta x \\
& \Leftrightarrow & \\
x_{\min} + \dfrac{i + 0.5 - i_{\min}}{\Delta i + 1}\Delta x & < & x_{\min} + \dfrac{k + 0.5 - i_{\min}}{\Delta i + 1}\Delta x \\
& \Leftrightarrow & \\
f(i) & < & f(k)
\end{array}$$

g is order-preserving too, but we leave out the proof as trivial (it would be similar to that for f).

End-point requirements

$$\begin{aligned}
g(x_{\min}) &= i_{\min} + \text{clamp}\left(\left\lfloor \frac{x_{\min} - x_{\min}}{\Delta x}(\Delta i + 1)\right\rfloor, 0, \Delta i\right) \\
&= i_{\min} + \text{clamp}\,(0, \Delta i) \\
&= i_{\min} \\
g(x_{\max}) &= i_{\min} + \text{clamp}\left(\left\lfloor \frac{x_{\max} - x_{\min}}{\Delta x}(\Delta i + 1)\right\rfloor, 0, \Delta i\right) \\
&= i_{\min} + \text{clamp}\,(\lfloor \Delta i + 1 \rfloor, 0, \Delta i) \\
&= i_{\min} + \Delta i \\
&= i_{\max}
\end{aligned}$$

g is a left-inverse of f

$$\begin{aligned}
g(f(i)) &= i_{\min} + \text{clamp}\left(\left\lfloor \frac{x_{\min} + \frac{i+0.5-i_{\min}}{\Delta i+1}\Delta x - x_{\min}}{\Delta x}(\Delta i + 1)\right\rfloor, 0, \Delta i\right) \\
&= i_{\min} + \text{clamp}\left(\left\lfloor \frac{i + 0.5 - i_{\min}}{\Delta i + 1}(\Delta i + 1)\right\rfloor, 0, \Delta i\right) \\
&= i_{\min} + \text{clamp}\,(\lfloor i + 0.5 - i_{\min} \rfloor, 0, \Delta i) \\
&= i_{\min} + \text{clamp}\,(i - i_{\min} + \lfloor 0.5 \rfloor, 0, \Delta i) \\
&= i_{\min} + \text{clamp}\,(i - i_{\min}, 0, \Delta i) \\
&= \text{clamp}\,(i, i_{\min}, i_{\max}) \\
&= i
\end{aligned}$$

It follows from this that g is a surjection and f is an injection.

Other requirements

For each i, $g^{-1}[\{i\}]$ is a half-open interval. Thus, the functions f and g as defined satisfy all the requirements and represent a solution to the problem.

3 Signed integer quantization

Let us try to apply the generalized solution to our practical problem. In this case $I = [-2^{p-1}, 2^{p-1} - 1] \subset \mathbb{Z}$ and $X = [-1, 1] \subset \mathbb{R}$. We wish to add two more requirements for the solution: $f(-i) = -f(i)$ and $g(-x) = -g(x)$. That is, the functions must be antisymmetric. In particular, this implies that $f(0) = 0$ and $g(0) = 0$. Clearly this can't be fulfilled because I is not symmetric. The solution is to remove the value -2^{p-1} from the integer interval. While this may sound bad, you don't actually lose anything important: for a p-bit signed integer in two's complement form it holds that $-(-(2^{p-1})) = -2^{p-1}$, that is, this value has no negation. This means, for example, that you can't correctly flip the sign of a signed integer sound signal if it contains the value -2^{p-1}. This is why this value value should never be used. Let us take away that problematic value from our interval, and obtain a symmetric integer interval. Let us then convert between a real range $[-1, 1]$ and an integer range $[-N, N]$, where $N = 2^{p-1} - 1$. Then

$$\begin{aligned}
g(x) &= -N + \operatorname{clamp}\left(\left\lfloor \frac{x-(-1)}{2}(2N+1) \right\rfloor, 0, 2N\right) \\
&= -N + \operatorname{clamp}\left(\left\lfloor \frac{x+1}{2}(2N+1) \right\rfloor, 0, 2N\right) \\
&= \operatorname{clamp}\left(-N + \left\lfloor \frac{x+1}{2}(2N+1) \right\rfloor, -N, 2N-N\right) \\
&= \operatorname{clamp}\left(\left\lfloor -N + \frac{x+1}{2}(2N+1) \right\rfloor, -N, N\right) \\
&= \operatorname{clamp}\left(\left\lfloor \frac{-2N+(x+1)(2N+1)}{2} \right\rfloor, -N, N\right) \\
&= \operatorname{clamp}\left(\left\lfloor \frac{-2N+x(2N+1)+2N+1)}{2} \right\rfloor, -N, N\right) \\
&= \operatorname{clamp}\left(\left\lfloor \frac{x(2N+1)+1}{2} \right\rfloor, -N, N\right) \\
&= \operatorname{clamp}\left(\left\lfloor x(N+0.5)+\frac{1}{2} \right\rfloor, -N, N\right)
\end{aligned}$$

To check antisymmetry, let us first assume that the term in the floor function is not an integer. Then it holds that $1 - \lceil x \rceil = - \lfloor x \rfloor$.

$$\begin{aligned}
g(-x) &= \operatorname{clamp}\left(\left\lfloor -x(N+0.5) + \frac{1}{2} \right\rfloor, -N, N\right) \\
&= \operatorname{clamp}\left(-\left\lceil x(N+0.5) - \frac{1}{2} \right\rceil, -N, N\right) \\
&= \operatorname{clamp}\left(1-\left\lceil x(N+0.5) + \frac{1}{2} \right\rceil, -N, N\right) \\
&= \operatorname{clamp}\left(-\left\lfloor x(N+0.5) + \frac{1}{2} \right\rfloor, -N, N\right) \\
&= -\operatorname{clamp}\left(\left\lfloor x(N+0.5) + \frac{1}{2} \right\rfloor, -N, N\right) \\
&= -g(x)
\end{aligned}$$

However, if we now assume that the term in the floor function *is* an integer, then we will see that $g(-x) \neq -g(x)$. To fix this, we change the floor function to a truncation in such a way that the value of g is not changed on non-integers. This is done by $[x] \approx \left\lfloor x + \frac{\operatorname{sign}(x)-1}{2} \right\rfloor$ (where $\approx$ denotes equality for all non-integers). No requirement is invalidated by this change.

$$g(x) = \operatorname{clamp}\left(\left[x(N+0.5) + \frac{\operatorname{sign}(x)}{2} \right], -N, N\right)$$

Now it follows easily that:

$$\begin{aligned}
g(-x) &= \operatorname{clamp}\left(\left[-x(N+0.5) + \frac{\operatorname{sign}(-x)}{2} \right], -N, N\right) \\
&= \operatorname{clamp}\left(\left[-x(N+0.5) - \frac{\operatorname{sign}(x)}{2} \right], -N, N\right) \\
&= \operatorname{clamp}\left(-\left[x(N+0.5) + \frac{\operatorname{sign}(x)}{2} \right], -N, N\right) \\
&= -\operatorname{clamp}\left(\left[x(N+0.5) + \frac{\operatorname{sign}(x)}{2} \right], -N, N\right) \\
&= -g(x)
\end{aligned}$$

The f simplifies as follows:

$$\begin{aligned}
f(i) &= -1 + \frac{i + 0.5 + N}{2N + 1} 2 \\
&= -1 + \frac{i + N + 0.5}{N + 0.5} \\
&= \frac{i}{N + 0.5} \\
f(-i) &= \frac{-i}{N + 0.5} \\
&= -f(i)
\end{aligned}$$

Last, we can extend to handle out-of-range values (in particular, $i = -(N + 1)$) gracefully by clamping and move the truncation out to suggest that the clamping should be done in floating point because the value might not fit into an integer:

$$\begin{aligned}
f(i) &= \text{clamp}\left(\frac{i}{N + 0.5}, -1, 1\right) \\
g(x) &= \left[\text{clamp}\left(x(N + 0.5) + \frac{\text{sign}(x)}{2}, -N, N\right)\right]
\end{aligned}$$

4 Acknowledgements

Discussions in the newsgroup comp.graphics.algorithms were helpful when forming this paper. Thomas Richter directed me to the field of quantization when I was doing 'conversions between real and integer ranges'. Niels Fröhling noticed that the function g in the generalized solution is the well-known mid-rise quantizer. Daniel Pitts suggested the error between $g(f(x))$ and x as an optimality criterion.