Analysis of Dyadic Approximation Error for Hybrid Video Codecs with Integer Transforms

Chau-Wai Wong, Student Member, IEEE, and Wan-Chi Siu, Senior Member, IEEE

(Accepted for Publication)

Abstract—In this paper, we present an analysis of the dyadic approximation error introduced by the integerization of transform coding in H.264/AVC-like codecs. We derive the analytical formulations for dyadic approximation error and nonorthogonality error. We further classify the dyadic approximation error into a “system error” and a “non-flat error,” and proposed two models for them. We found that the “non-flat error” has a substantial impact on video quality if the number of shifting bits at decoder side (DQ_BITS) is small. We also give a theoretical justification on why scaling factors at encoder side are better to be adapted to the rescaling factors at decoder side in H.264/AVC-like codecs.

Index Terms—Dyadic Approximation, Error Analysis, Integer Transform, Video Coding, Discrete Cosine Transform.

I. INTRODUCTION

The discrete cosine transform (DCT) [1-4] has been used as the preferred transform kernel in video codecs for more than twenty years. However, since the first video codec, the problem “drift” [5-6] persists, which lowers the coding efficiency due to the mismatch between encoder and decoder. The state-of-the-art video standard H.264/AVC [7-8] solved the drift problem by adopting one [9-10] of the integer variations of the DCT, which replaces the floating-point arithmetic with integer multiplications and bit-wise shifting operations [9]. Although the integer transform eliminates the drift problem, it introduces another problem inherited from the dyadic approximation of floating-point numbers for post- and pre-scaling stages. Let us call it the “dyadic approximation error.” Although the dyadic approximation
error has been noticed [11-14] during the standardization process of the H.264/AVC, no analytical investigation has been found in the literature. Hence, in this paper, we derive an analytical form of the dyadic approximation error in order to have a better understanding of its effects and characteristics. For example, the benefits of a technical suggestion can be theoretically proved, that the scaling factors for the post-scaling stage at the encoder side should be “adapted” to the rescaling factors for the pre-scaling stage at the decoder side [10]. Furthermore, for those cases that the number of shifting bits at the decoder side (DQ_BITS) is small, the sacrifice of this “adaptation” can also be observed.

In Section II, problems of the present study are formed, symbols are defined, an error analysis is carried out, and analytical models of different kinds of errors are derived. Section III verifies the proposed models with experimental results. Section IV concludes the paper.

II. DYADIC APPROXIMATION ERROR

A. Equivalence Model for Dyadic Approximation Error

In an ideal hybrid video codec model, the only error source is the quantization process. However, under the framework of integer transform, other error sources may exist. For example, if the normalized integer kernel is not strictly orthogonal, the nonorthogonality error exists [15-16]. For any H.264/AVC-like [7, 17-24] video codec, it suffers from the “dyadic approximation error” due to the integerization of transform coding with a finite number of shifting bits.

Let us use Fig. 1 to explain the source of the dyadic approximation error and the way we model it. Recall that the main idea of integerization of the transform coding is to split the “transform” in Fig. 1(a) into “integer transform” and “normalization” in Fig. 1(b), and then combine “normalization” and “quantization” in Fig. 1(b) together with a new name “post-scaling” in Fig. 1(c). The “pre-scaling” is constructed similarly at the decoder side. In spite of different block names of the three flows in Fig. 1(a), (b), and (c), the flows are totally equivalent, and do not suffer from any other error except quantization error and nonorthogonality error.
Fig. 1. Diagrams showing the source of dyadic approximation error: Each row contains stages that convert original signal (represented by “x”) into reconstructed signal (represented by “y_r”). Each row presents a variation: (a), (b), and (c) the ideal cases; (d) the practical case which includes the dyadic approximation error implicitly; (e) the practical case which models all error factors explicitly; and (f) the analytical expression of (e). Note that i) for each block in (a)–(e), the bottom part indicates whether this block causes error; and ii) “DA” stands for “dyadic approximation.”

Let us use the “post-scaling” as an example. The multiplication between an integer $m$ and a floating-point number $f$ (a factor takes normalization and quantization into consideration) in “post-scaling” of Fig. 1(c), depicted by Fig. 2(a), is implemented by the integer multiplication between $m$ and $k$ followed by an $n$-bit right-shifting operation (denoted with “$\gg$”), where the integer $k$ and natural number $n$ are defined in the dyadic fraction representation of $f$, $k/2^n$, in which $n$ is infinitively large for ideal cases. However, for real cases as shown in Fig. 2(b), $n$
can only be a finite number; hence the mapped multiplication is not identical to the original multiplication. This leads to the error which is actually introduced by the dyadic approximation of $f$ as shown below

$$f \approx \frac{k}{2^n}. \quad (1)$$

The “post-scaling” of Fig. 1(c) is the model for the ideal case that the number of shifting bits $n \to +\infty$, whereas “post-scaling” of Fig. 1(d) is the model for the real case that $n$ is finite. The error thus introduced by the dyadic approximation can be measured in terms of the ratio between the approximated value and the actual value, which is

$$S = \left(\frac{k}{2^n}\right) / f. \quad (2)$$

The dyadic approximation error is implicitly shown by the italic words “Finite Precision” in “post-scaling” and “pre-scaling” of Fig.1 (d), which is not traced by any block and thus inconvenient for mathematical modeling. Hence it is beneficial to explicitly trace the errors with the 2$^{nd}$ and 5$^{th}$ blocks in Fig. 1(e), in which the dyadic approximation error is explicitly modeled. The reason that “transform” and “quantization” are used instead of “integer transform” and “post-scaling” is due to their equivalence which has been explained before.

It is observed that the operations needed under the integer framework are so fundamental that the ALUs of different processors are able to produce exactly the same results if the number of shifting bits $n$ (more specifically, $n_{\text{forward}}$ (Q_BITS) for the post-scaling stage and $n_{\text{inverse}}$ (DQ_BITS) for the pre-scaling stage) is properly specified hence the potential mismatch between the encoder side and decoder side can be avoided.

![Fig. 2. Implementation of “post-/pre-scaling” without floating-point arithmetic in (a) ideal case, and (b) real case.](image-url)
It has been made clear that both the encoder side and decoder side suffer from the dyadic approximation errors. Hence, in order to reduce the defect due to the dyadic approximation, it is intuitive to adapt the approximation of the encoder side according to that of the decoder side (since video coding standard only defines decoder side). The VCEG recommends that the post-scaling factor $k_{\text{forward}}$ at the encoder side be adapted as pre-scaling factor $k_{\text{inverse}}$ at the decoder side [10]:

$$k_{\text{forward}}(i, j) = \text{round}[2^{Q\_B\_BITS+DQ\_B\_BITS} \times ||b_i||^2 \times ||b_j||^2 / k_{\text{inverse}}(i, j)],$$ (3)

where $i, j \in [0, N-1]$ are the indices, $N$ is the order of integer kernel used, $Q\_B\_BITS/DQ\_B\_BITS$ is the number of shifting bits at encoder/decoder side, $b_n$ is the $n^{th}$ row vector of the integer kernel, $\text{round}[]$ is the round-to-nearest-integer operation, and $||\cdot||_2$ is the norm-2 operator. Equation (3) reveals that an increase in pre-scaling factor $k_{\text{inverse}}$ means a decrease in post-scaling factor $k_{\text{forward}}$, or vice versa. It is equivalent to a recommendation that the dyadic approximation errors at encoder side ($S_1$) and decoder side ($S_2$) fulfill the following constraint:

$$S_1 \times S_2 = 1.$$ (4)

However, a theoretical justification for (3) is needed. Why the adaptation of post-scaling factor to pre-scaling factor can help lowering the overall error, when other factors, e.g. quantization and nonorthogonality, exist simultaneously?

Note that (4) can never be fulfilled exactly, since the number of shifting bits is finite. It becomes even worse for those scenarios that small $Q\_B\_BITS$ and $DQ\_B\_BITS$ are required for a small dynamic range. If the dyadic approximation error of the whole system is not unit, the perfect reconstruction property of the transform cannot be preserved even though the quantization is bypassed. This error causes a drift between the raw residual signal and the reconstructed residual signal, which lowers the coding efficiency and the visual quality. Let us name it the “system error.”

Even if the dyadic approximation error of the whole system is unit, another problem also exists. Recall in Fig. 1(e), the blocks “DA Error Equivalent Model” and “Quantization” are
next to each other. We can create equivalence by deliberately setting the dyadic approximation error to zero and moving the errors into the quantization stage. For example, setting quantizer step $q = 5$, $S_1 = 1.1$ and $S_2 = 0.9091$ is equivalent to setting $q = 5/1.1 = 4.545$ and $S_1 = S_2 = 1$. For the general case, even though the dyadic approximation error of the whole system is unit, the transformed coefficients of different positions are quantized with different quantizer steps, equivalently. This can be considered as being quantized with a non-flat quantization matrix. The corresponding error arisen can be named as the “non-flat error,” and a theoretical investigation is needed.

### C. Analysis of Dyadic Approximation Error in Terms of MSE

In order to measure the significance of dyadic approximation error, it should be compared to other errors, e.g. quantization error and nonorthogonality error. We thus derive the mean square error (MSE) between the reconstructed signal and the input signal taking the effects of dyadic approximation, quantization and nonorthogonality into the consideration. Let us first denote

1) an input signal as an $N \times 1$ vector $\mathbf{x}$,

2) the normalized nonorthogonal transform kernel as an $N \times N$ matrix $\mathbf{H}$,

3) the transformed vector as $\mathbf{z}$,

4) the matrix which models the error introduced by the dyadic approximation at the post-scaling stage as $S_1$ (which is a diagonal matrix with diagonal elements slightly larger or smaller than 1 (see Fig. 3(a) for an example)),

5) the distorted vector due to the dyadic approximation as $\mathbf{v}$,

6) the quantized vector as $\mathbf{u}$,

7) the vector modeling the quantization error as $\mathbf{n}_q$,

8) the matrix which models the error introduced by the dyadic approximation at the pre-scaling stage as $S_2$ (which is a diagonal matrix with diagonal elements slightly larger or smaller than 1 (see Fig. 3(b) for an example)),

9) the output vector as $\mathbf{y}_r$, and

10) the output obtained by omitting the quantization as $\mathbf{y}$. 
Fig. 3. Equivalence models for dyadic approximation error for H.264/AVC’s default 4×4 kernel $1(1,2,1)$ [29] with quantizer step = 5 at (a) encoder side with number of shifting bits = 7, and (b) decoder side with number of shifting bits = 3.

Note that throughout this paper, we use a letter in bold to represent a matrix or a vector (for example $H$ or $x$), whereas a letter in italic with index term(s) to represent one particular element in that matrix or vector (for example, $H(i,j)$ or $x(i)$). Let us also define the following symbols and equivalences for clarity of presentation

$$N_1 = N_1^T = S_1^{-1} - I,$$  \hspace{1cm} (5)

$$N_2 = N_2^T = S_2 - I,$$  \hspace{1cm} and

$$E_r = E_r^T = HH^T - I.$$  \hspace{1cm} (7)

Note that the transpose relationship in (5) and (6) is obvious since the off-diagonal elements of $S_1$ and $S_2$ are all zero (See Fig. 3). Let us represent the operation from the forward transform to inverse transform in Fig. 1(f) as shown by the following set of equations

$$z = Hx,$$  \hspace{1cm} (8)

$$v = S_1z,$$  \hspace{1cm} (9)

$$u = quant^{-1}[quant(v)] = v - n_q,$$  \hspace{1cm} and

$$y_r = (S_2H)^T u.$$  \hspace{1cm} (10)

The reconstruction error can be calculated by using the average MSE between the input $x$ and output $y_r$ as shown below

$$\sigma_r^2 = \frac{1}{N} E[(x - y_r)^T(x - y_r)].$$  \hspace{1cm} (12)

The difference between $x$ and $y_r$ can be expanded into

$$x - y_r = H^{-1}z - (S_2H)^T u = (H^T - H_{cr}(S_1^{-1}v - (S_2H)^T(v - n_q) = H^T S_2 n_q - H c_r S_1^{-1}v + H^T(S_1^{-1} - S_2)v.$$  \hspace{1cm} (13)
where \( H_{er} = H^T - H^{-1} \). Let us define
\[
\begin{align*}
A &= H^T S_2 n_q - H_{er} S_1 v, \text{ and} \\
B &= H^T (S_1^{-1} - S_2) v.
\end{align*}
\]
Hence (12) can be represented by
\[
\begin{align*}
\sigma_i^2 &= \frac{1}{2} E\left[ (A + B)^T (A + B) \right] = \frac{1}{2} \left( E\left[ A^T A \right] + 2E\left[ A^T B \right] + E\left[ B^T B \right] \right).
\end{align*}
\]
The first term of (16) can be written as
\[
\begin{align*}
\frac{1}{2} E\left[ A^T A \right] &= \frac{1}{2} E\left[ (H^T S_2 n_q - H_{er} S_1 v)^T (H^T S_2 n_q - H_{er} S_1 v) \right] \\
&= \frac{1}{2} E\left[ n_q^T S_2 H H^T S_2 n_q + v^T S_1^T H_{er} S_1 v - 2v^T S_1^T H_{er}^T H^T S_2 n_q \right] \\
&= \sigma_q^2 + \frac{1}{2} E\left[ n_q^T (N_1^2 E r_{N_2} + N_2^2 E r_{N_2} + E r_{N_2} + 2E r_{N_2} + 2N_2) v_q + (y - x)^T (y - x) \right] \\
&= \sigma_q^2 + \frac{1}{2} E\left[ n_q^T (N_2^2 E r_{N_2} + N_2^2 E r_{N_2} + E r_{N_2} + 2E r_{N_2} + 2N_2) v_q \right] + \sigma_o^2,
\end{align*}
\]
where \( \sigma_o^2 = \frac{1}{2} E\left[ (y - x)^T (y - x) \right] \). It can be observed that \( \sigma_o^2 \) is the average MSE between the input and output without quantization. It has been shown in [16] that
\[
\sigma_o^2 \leq \sigma_i^2 \left[ \frac{1}{2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} M(k, j) \right],
\]
where
\[
M = (H^T H - I)^T (H^T H - I).
\]
For the second term of (16), we have
\[
\begin{align*}
\frac{1}{2} E\left[ A^T B \right] &= \frac{1}{2} E\left[ (H^T S_2 n_q - H_{er} S_1 v)^T (H^T (S_1^{-1} - S_2) v) \right] \\
&= \frac{1}{2} E\left[ n_q^T S_2 H H^T (S_1^{-1} - S_2) v + v^T (-S_1^{-1} (H^T H^{-1}) H^T (S_1^{-1} - S_2)) v \right],
\end{align*}
\]
where \( E[n_q^T S_2 H H^T (S_1^{-1} - S_2) v] = 0 \) since
\[
E[n_q(i)v(j)] = 0 \text{ for all } i, j \in [0, N-1].
\]
This can be proved by considering the cases for \( i = j \) and \( i \neq j \). For \( i = j \), \( E[n_q(i)v(j)] \) measures the correlation between a quantizer input and its quantization noise. As it is shown in [25], these two signals are uncorrelated provided that the quantizer input is “band-limited in the CF domain” of which general cases approximately fulfills the condition. For \( i \neq j \), \( E[n_q(i)v(j)] \) measures the correlation between a quantizer input and the quantization noise of another quantizer. By assuming that the DCT coefficients are independent of each other (which means
that each DCT coefficient can be regarded as an independent quantizer input), \( n_q(i) \) and \( v(j) \) are obviously uncorrelated. Hence

\[
\frac{1}{n} E\left[ A^T B \right] = \frac{1}{n} E\left[ x^T (H^T Er(S_2 - S_1) H) x \right].
\]  

(22)

For the third term of (16)

\[
\frac{1}{n} E\left[ B^T B \right] = \frac{1}{n} E\left[ \left( H^T (S_1^T - S_2) \right)^T \left( H^T (S_1^T - S_2) \right) \right]
\]

\[
= \frac{1}{n} E\left[ x^T \left( H^T S_1^T (S_1^T - S_2) \right) \left( Er + I \right) \left( S_1^T - S_2 \right) S_1 H x \right]
\]

\[
= \frac{1}{n} E\left[ x^T \left( H^T (I - S_1 S_2) (Er + I) (I - S_1 S_2) H \right) x \right].
\]  

(23)

Hence we obtain

\[
\sigma_i^2 = \sigma_q^2 + \sigma_{i0}^2 + \frac{1}{n} E\left[ n_q^T \Omega n_q \right] + \frac{1}{n} E\left[ x^T \Psi x \right]
\]  

(24)

where \( \sigma_q^2 \) is the average MSE of the quantization error, \( \sigma_{i0}^2 \) is the average MSE of the nonorthogonality error,

\[
\Omega = \sum_i \sum_j \Omega(i,j) n_q(i) n_q(j) = \sum_i \sum_j \Omega(i,j) E[n_q(i) n_q(j)].
\]  

(25)

\[
\Psi = H^T S_1 \left[ \left( S_1^T - S_2 \right) - \left( S_1^T + S_2 \right) Er \right] \left( S_1^T - S_2 \right) S_1 H.
\]  

(26)

Let us continue to simplify the third term of (24) as shown below

\[
E\left[ n_q^T \Omega n_q \right] = \sum_i \sum_j \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \Omega(i,j) n_q(i) n_q(j) = \sum_i \sum_{j=0}^{N-1} \sum_{j=0}^{N-1} \Omega(i,j) E[n_q(i) n_q(j)].
\]  

(27)

For quantization noise \( n_q \), elements of its autocorrelation matrix are zero except the diagonal terms, hence

\[
E\left[ n_q^T \Omega n_q \right] = \sum_{i=0}^{N-1} \Omega(i,i) E[n_q^2(i)] = \sigma_i^2 \sum_{i=0}^{N-1} \Omega(i,i).
\]  

(28)

Similarly, the fourth term of (24) can be simplified as shown below

\[
E\left[ x^T \Psi x \right] = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \Psi(i,j) x(i) x(j) E[x(i) x(j)]
\]

\[
= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \Psi(i,j) \rho(i,j) \sigma_x^2 \rho(i,j).
\]  

(29)

The correlation coefficient \( \rho(i,j) \) ranges from –1 to 1, hence the maximum value of \( E[x^T \Psi x] \) can be reached if (a) \( |\rho(i,j)| \to 1 \), and (b) \( \rho(i,j) \) has the same sign with \( \Psi(i,j) \). Condition (a) is easily fulfilled by setting the absolute value of all correlation coefficients to 1. Condition (b) can only be fulfilled for some positions. It is because the number of sign patterns of \( \Psi(i,j) \) is
$2^{(N-1)/2}$, whereas the number of sign patterns of $\rho(i,j)$ is only $2^{N-1}$. Although condition (b) cannot be really achieved, our experimental works show that (b) can lead to a good approximation to the real upper bound. Hence the fourth term of (24) can be written as

$$E[|x^T \Psi x|] \leq \sigma_n^2 \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |\Psi(i,j)|.$$  \hspace{1cm} (30)

Let us also modify (18) by using the above idea,

$$\sigma_n^2 \leq \sigma_i^2 \left[ \frac{1}{N} \sum_{k=0}^{N-1} |M(k,j)| \right].$$  \hspace{1cm} (31)

Hence the upper bound of the average variance of reconstruction can be written as follow

$$\sigma_n^2 + \sigma_i^2 \left[ \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |M(i,j)| \right] + \sigma_i^2 \left[ \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |\Omega(i,i)| \right] + \sigma_i^2 \left[ \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |\Psi(i,j)| \right],$$  \hspace{1cm} (32)

where $M$, $\Omega$ and $\Psi$ are defined by (19), (25) and (26) respectively.

D. Extension into 2-D Form

We have derived the analytical representation of reconstruction error by considering quantization, nonorthogonality and dyadic approximation in one-dimensional (1-d) form. Let us extend the analysis into two-dimensional (2-d) form so that the integer kernels for H.264/AVC-like video codecs can be evaluated.

To begin with, let us define the input signal as an $N \times N$ matrix $X$, hence the transformed matrix is $Z = HXH^T$. In order to simplify the derivation, matrices $X$ and $Z$ can be lexicographically ordered into column vectors $x$ and $z$ of length $N^2$, where $x(i+jN) = X(i,j)$ and $z(i+jN) = Z(i,j)$ for $i, j \in [0, N-1]$. We convert the 2-d transform into the lexicographical form $z = \mathcal{H}x$ where $\mathcal{H}$ is the Kronecker product of $H$ with itself, i.e. $\mathcal{H} = H \otimes H$ [26-27]. By considering the dyadic approximation error at the forward transform stage, we multiply the transform kernel $H$ by the scalar matrix $S_1$. Hence, the distorted transformed signal $v$ in lexicographical form can be formulated as $v = [(S_1H) \otimes (S_1H)]x = (S_1 \otimes S_1)(H \otimes H)x = S_1\mathcal{H}x$ where $S_1 = S_1 \otimes S_1$. It can be seen that $v = S_1\mathcal{H}x$ shares exactly the same form with the combination of (8) and (9) except that the order has been raised from $N$ to $N^2$. Hence, we rewrite (5)–(11) for the 2-d case by simply changing the variables from the typewritten form into the script form, as shown below.
\[ z = \mathcal{H}x, \]  
(33)

\[ u = S_1z, \]  
(34)

\[ u = \text{quant}^{-1}[\text{quant}(u)] = u - q, \]  
(35)

\[ y_t = (S_2\mathcal{H})^T u, \]  
(36)

\[ N_1 = N_1^T = S_1^{-1} - I, \]  
(37)

\[ N_2 = N_2^T = S_2 - I, \]  
(38)

\[ \mathcal{E}_x = \mathcal{E}_x^T = \mathcal{H}\mathcal{H}^T - I, \]  
(39)

where the definitions of symbols are exactly the same w.r.t. their 1-d counterparts, hence we can omit the definitions for the sake of simplicity. By going through the derivations in last subsection once again, we obtain similar representation of the average variance of reconstructed signal, as shown below

\[
\sigma_q^2 + \sigma_s^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathcal{M}(i, j) \right] + \sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N-1} \mathcal{W}(i, i) \right] + \sigma_s^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathcal{Y}(i, j) \right],
\]

(40)

where

\[ \mathcal{M} = (\mathcal{H}^T\mathcal{H} - I)(\mathcal{H}^T\mathcal{H} - I), \]  
(41)

\[ \mathcal{W} = N_1^T\mathcal{E}_x N_2 + N_2^T N_2 + \mathcal{E} + 2\mathcal{E}_x N_2 + 2N_2, \]  
(42)

\[ \mathcal{Y} = \mathcal{H}^T S \left[ (S^{-1} - S) - (S^{-1} + S) \mathcal{E}_x \right] (S^{-1} - S) S \mathcal{H}. \]  
(43)

E. Theoretical Interpretations

1) Overall

The reconstruction error derived in terms of MSE with the consideration of quantization, nonorthogonality, and dyadic approximation effects has been obtained in (40). We need to match the terms with their error sources in order to compare significances of various errors. However, some non-separable terms may simultaneously contain more than one error sources.

For example, for \( N_1^T\mathcal{E}_x N_2 \) in (42), the term \( N_2 \) relates to the dyadic approximation at decoder side, and term \( \mathcal{E}_x \) relates to the nonorthogonality of the transform kernel. Hence, some criteria must be determined before matching the terms with error sources. One reasonable approach...
we suggest is to make use of the following criteria: (a) when one type of error vanishes, its
error terms should have a sum of 0, (b) for a known error type (e.g. the nonorthogonality error
[15-16]), when any new error term is considered belonging to that type, only error terms that
are exclusively dependent on that error type should be added. Let us match the terms with
their error sources:

1. $\sigma_q^2$ is the quantization error by its definition.

2. $\sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} \sum_{j=0}^{N^2-1} |M(i, j)| \right]$ is the major error term that measures the nonorthogonality
error. (Recall that in [15-16], the nonorthogonality error term is also in this form.) By a
careful examination on (42) and (43), it is found that the term $\mathcal{E}_x$ is exclusively dependent
on the nonorthogonality, hence $\sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} \mathcal{E}_x(i, i) \right]$ is also an error term of
nonorthogonality. Hence the nonorthogonality error is

$$
\sigma_{\text{nonortho}}^2 = \sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} \sum_{j=0}^{N^2-1} |M(i, j)| \right] + \sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} \mathcal{E}_x(i, i) \right], \quad (44)
$$

which equals zero when kernel $\mathcal{H}$ is orthogonal.

3. Hence, the remaining terms are the error terms for dyadic approximation. Both of them are
zero when no dyadic error exists. The first one is the “system error”

$$
\sigma_{\text{system}}^2 = \sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} \sum_{j=0}^{N^2-1} |\mathcal{U}(i, j)| \right], \quad (45)
$$

the second one is the “non-flat error”

$$
\sigma_{\text{non-flat}}^2 = \sigma_q^2 \left[ \frac{1}{N^2} \sum_{i=0}^{N^2-1} |\mathcal{N}(i, i)| \right], \quad (46)
$$

where

$$
\mathcal{U} = \mathcal{W} - \mathcal{E}_x = \mathcal{N}_1^T \mathcal{E}_x \mathcal{N}_2 + \mathcal{N}_2^T \mathcal{N}_2 + 2 \mathcal{E}_x \mathcal{N}_2 + 2 \mathcal{N}_2. \quad (47)
$$

They are explained in details as follows.

2) **System Error**

According to (45) and (43), the system error $\sigma_{\text{system}}^2$ vanishes when the adaptation
condition $\mathcal{S}_1^{-1} = \mathcal{S}_2$ is fulfilled. This is a theoretical justification for VCEG’s recommendation
that post-scaling factor be adapted to pre-scaling factor as shown in (3). It means that when
the dyadic approximation errors at the encoder and decoder sides are coupled, the
reconstruction error in terms of MSE can be reduced.

Since the system error is determined by both Q_BITS and DQ_BITS, even if the
DQ_BITS specified by a standard is quite small, a large Q_BITS can be configured in
encoder to reduce the overall system error. Hence, the system error can be regarded as an
encoder issue.

3) Non-Flat Error

However, the non-flat error is solely a decoder issues as shown in (46) and (47). If no
dyadic approximation were found at the decoder side (N_2 = 0), the non-flat error \( \sigma^2_{\text{non-flat}} \) vanishes.

III. EXPERIMENTAL WORKS

A. Results Calculated From the Model

Let us calculate errors using the analytical models (45)–(46), and use the order-4
H.264/AVC default kernel IK(1,2,1) as an example (actually other kernels such as IK(5,7,3)
[29] can also be used). Note that our model is derived without making an assumption of the
order of transform, hence it can also be applied to kernels of any order, e.g. order 8 or 16.

Recall that for H.264/AVC-like codecs, the scalars and rescalars require huge storage [10]
for all QP values, if no periodicity design is applied. In the H.264/AVC specification [7], a
periodicity structure is employed that for each 6 steps increase of QP, the quantizer step \( q \) doubles:

\[
g(QP) = g(QP+6) / 2. \tag{48}
\]

Hence for all QP values, there are 6 different groups of scalars/descalars. In addition, the
dyadic ratio (2) is a function of \( q \) and \( n \), its value also repeats itself with a period of 6.

TABLE I shows the significances of quantization error, non-flat error, and system error, using
the IK(1,2,1) as an example. The values of QP were selected within a range of common
configurations. The DQ_BITS were selected to allow the non-flat error varying from low
(using 7 bits) to high (using 3 bits), and the Q_BITS were selected to maintain high system error. The error ratios are the terms embraced by “[ ]” in (45) and (46) respectively. (The nonorthogonality does not exist since the H.264/AVC kernel is orthogonal.) The table reveals that beside the dominating quantization error, the non-flat error is significant but the system error is almost 0. The non-flat error is more serve for small DQ_BITS, which is very reasonable.

Note that in some cases, the non-flat errors are negative. This can be mathematically explained that its error representation (46) and (47) does not contain any absolute sign. It can also be reasonably explained by considering this error as a kind of interference to quantization process—a non-flat quantization operation.

<table>
<thead>
<tr>
<th>Quantization Parameter</th>
<th>Quantizer Step</th>
<th>Average Variance*</th>
<th>No of Shifting Bits</th>
<th>Error Ratios</th>
<th>Quantization Error</th>
<th>Dyadic Approximation (Non-Flat Error)</th>
<th>Dyadic Approximation (System) Error</th>
<th>Reconstruction Error (Total Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Value</td>
<td>Weight</td>
<td>Value</td>
</tr>
<tr>
<td>24</td>
<td>10</td>
<td>33.7</td>
<td>8.33</td>
<td>6</td>
<td>3</td>
<td>3.84E+00</td>
<td>2.20E-03</td>
<td>8.3E+00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>4</td>
<td>1.64E+00</td>
<td>2.20E-03</td>
<td>8.3E+00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>5</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>8.3E+00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>8.3E+00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>7</td>
<td>1.13E-01</td>
<td>2.20E-03</td>
<td>8.3E+00</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>41.2</td>
<td>33.33</td>
<td>6</td>
<td>3</td>
<td>3.84E+00</td>
<td>2.20E-03</td>
<td>3.3E+01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>4</td>
<td>1.64E+00</td>
<td>2.20E-03</td>
<td>3.3E+01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>5</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>3.3E+01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>3.3E+01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>7</td>
<td>1.13E-01</td>
<td>2.20E-03</td>
<td>3.3E+01</td>
</tr>
<tr>
<td>36</td>
<td>40</td>
<td>48.7</td>
<td>133.33</td>
<td>6</td>
<td>3</td>
<td>3.84E+00</td>
<td>2.20E-03</td>
<td>1.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>4</td>
<td>1.64E+00</td>
<td>2.20E-03</td>
<td>1.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>5</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>1.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>1.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>7</td>
<td>1.13E-01</td>
<td>2.20E-03</td>
<td>1.3E+02</td>
</tr>
<tr>
<td>42</td>
<td>80</td>
<td>55.0</td>
<td>533.33</td>
<td>6</td>
<td>3</td>
<td>3.84E+00</td>
<td>2.20E-03</td>
<td>5.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>4</td>
<td>1.64E+00</td>
<td>2.20E-03</td>
<td>5.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>5</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>5.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
<td>-2.00E-01</td>
<td>2.20E-03</td>
<td>5.3E+02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>1.33E-01</td>
<td>2.20E-03</td>
<td>5.3E+02</td>
<td>89.9%</td>
</tr>
</tbody>
</table>

* Note that σ_i’s are obtained empirically, and σ^2_i’s are approximated by q_i^2/12.

B. Model Verification

In order to verify the model, we need to compare the results calculated from the model with the rate-distortion performance obtained from experiments. Due to the limitation of space, we focus on the non-flat error. The verification for system error model has also been carried out—no significant performance difference was obtained with/without the presence of the system error, which complies with the theoretical model.
Let us calculate the non-flat error in terms of MSE using our derived model (46) for $DQ_{BITS} \in [3,7]$ and $QP_{mod\_6} \in [0,5]$ as shown in TABLE II. It is observed that the non-flat error can either increase (positive entries) or reduce (negative entries) the reconstruction error in terms of MSE.

Experiments were also done on using the H.264/AVC evaluation software JM12.2 [18]. Two schemes are compared: the system error compensated scheme, and the system&non-flat errors compensated scheme. For the first scheme, only quantization error and non-flat error exist, whereas for the second scheme, only quantization error exists. TABLE III shows the experimental results in terms of BD-PSNR [28]. A positive value means that the gain has been achieved due to the compensation of non-flat error. Each entry is obtained by testing the performances for 4 QP values with a step of 6, and calculates the average PSNR improvement with BD metric [28]. For example, the upper-left entry is the average PSNR improvement calculated from QP = 24, 30, 36, and 42.

With a careful comparison between the estimated values calculated from the theoretical model in TABLE II and the BD-PSNR improvement from experiments in TABLE III, we observe a good match between them. 23 out of 30 co-located entries have the same sign. (The remaining 7 elements are formatted into italic in TABLE III.) Especially for entries with small $DQ_{BITS}$, the estimated values can better reflect the relative magnitudes of experimental results. This can be explained that when the non-flat error is large, other factors may not easily override the effect of non-flat error. We can also conclude that the non-flat error effectively leads to the fluctuation of video quality. As $DQ_{BITS}$ becomes less, the effect of fluctuation becomes more severe.

<table>
<thead>
<tr>
<th>DQ_BITS</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.84</td>
<td>1.64</td>
<td>-0.20</td>
<td>-0.20</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>2.48</td>
<td>0.66</td>
<td>-0.50</td>
<td>-0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>1.78</td>
<td>-0.66</td>
<td>0.38</td>
<td>-0.18</td>
<td>-0.18</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
<td>-0.55</td>
<td>-0.22</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>-0.19</td>
<td>1.38</td>
<td>-0.17</td>
<td>-0.17</td>
<td>-0.02</td>
</tr>
<tr>
<td>5</td>
<td>-0.99</td>
<td>0.69</td>
<td>0.46</td>
<td>-0.19</td>
<td>-0.01</td>
</tr>
<tr>
<td>BD-PSNR (dB)</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>-------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0</td>
<td>0.384</td>
<td>0.327</td>
<td>0.007</td>
<td>−0.013</td>
<td>0.021</td>
</tr>
<tr>
<td>1</td>
<td>0.279</td>
<td>0.199</td>
<td>0.042</td>
<td>0.087</td>
<td>0.110</td>
</tr>
<tr>
<td>2</td>
<td>0.003</td>
<td>0.032</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.115</td>
<td>0.030</td>
<td>−0.008</td>
<td>0.024</td>
<td>0.030</td>
</tr>
<tr>
<td>4</td>
<td>−0.090</td>
<td>0.169</td>
<td>−0.117</td>
<td>−0.041</td>
<td>−0.001</td>
</tr>
<tr>
<td>5</td>
<td>−0.184</td>
<td>0.183</td>
<td>0.071</td>
<td>−0.052</td>
<td>0.099</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

In this paper, we derived the analytical formulation of reconstruction error in terms of MSE with the consideration of quantization, nonorthogonality, and dyadic approximation effects. We further classified the dyadic approximation error into a “system error” and a “non-flat error.” Our analytical models can be used directly for the analyses of order-4, 8, or 16 transforms. We found if the number of shifting bits at decoder side (DQ_BITS) is small, the non-flat error can be substantial. We have also theoretically justified why scaling factors at encoder side are better to be adapted to the rescaling factors at decoder side in H.264/AVC-like codecs.

REFERENCES


**Chau-Wai Wong** (S’05) was born in Hangzhou, Zhejiang, China, in 1986. He received the B.Eng. degree with first class honors in 2008, and M.Phil. degree in 2010, both in electronic and information engineering from the Hong Kong Polytechnic University.

He is currently a Research Associate with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, under the supervision of Prof. Wan-Chi Siu. His research interests include video coding, transform coding, directional transforms, and image/video interpolation.

Mr. Wong is members of the APSIPA, and the HKIE. His undergraduate study was supported by the Hong Kong Government and the Hong Kong Polytechnic University with a 4-year full scholarship. He has received numerous scholarships and awards for his academic achievements, such as being included in the Dean’s Honors List, the HSBC Scholarship, the Hitachi (Hong Kong) Ltd. Scholarship, the CMA & Donors Scholarship, and the Sonca Products Gold Medal Award. He was the general secretary of the IEEE PolyU Student Branch from 2006 to 2007.

**Wan-Chi Siu** (S’77-M’77-SM’90) received the Associateship from The Hong Kong Polytechnic University and the MPhil degree from The Chinese University of Hong Kong in 1975 and 1977, respectively, and the Ph.D. Degree from Imperial College of Science, Technology & Medicine, London, UK, in October 1984. He was with The Chinese University of Hong Kong as a Tutor and later as an Engineer between 1975 and 1980. He then joined The Hong Kong Polytechnic University as a Lecturer in 1980, and has become Chair Professor in the Department of Electronic and Information Engineering (EIE) since 1992. He was Head of
EIE subsequently Dean of Engineering Faculty between 1994 and 2002. Professor Siu is now Director of the Centre for Signal Processing of the same university.

He is an expert in Digital Signal Processing, specializing in fast algorithms and video coding, and has published 380 research papers, over 160 of which appeared in international journals, such as IEEE Transactions on Image Processing. His research interests include also transforms, image coding, transcoding, 3D videos, wavelets, and computational aspects of pattern recognition. His work on fast computational algorithms (such as DCT) and motion estimation algorithms have been well received by academic peers, with good citation records, and a number of which are now being used in hi-tech industrial applications, such as modern video surveillance and video codec design for HDTV systems of some million dollar contract consultancy works. Professor Siu is now Associate Editor of the IEEE Transactions on Image Processing, and member of editorial board of a number of other journals such as Journal of VLSI Signal Processing Systems for Signal, Image, Video Technology, etc. He was Guest Editor and Associate Editor of IEEE Transactions on Circuits and Systems, and editorial board member of Pattern Recognition and the EURASIP Journal on Applied Signal Processing. He is editor of the book, ‘Multimedia Information Retrieval and Management’, Springer Berlin Heidelberg 2003. He is also the recipient of a number of awards, including the Distinguished Presenter Award (1997, HK), IEEE Third Millennium Medal (2000, USA), the Best Teacher Award (2003, HK), the Outstanding Award in Research (2003, HK), Plaque for Exceptional Leadership (2003, IEEE SPCB, USA), and Honorable Mention Winner Award (Pattern Recognition, 2004, USA).

Professor Siu is a very popular lecturing staff member within the University, while outside the University he has been a keynote speaker of over 10 international/national conferences in the recent 10 years, and an invited speaker of numerous professional events, such as IEEE CPM’2002 (keynote speaker, Taipei, Taiwan), IEEE ISIMP’2004 (keynote speaker, Hong Kong), and IEEE ICICS’07 (invited speaker, Singapore) and IEEE ICNNSP’2008 (keynote speaker, Zhenjiang). He is the organizer of many international conferences, including the MMSP’08 (Australia) as General Co-Chair, and three IEEE Society sponsored flagship conferences: ISCAS’1997 as Technical Program Chair; ICASSP’2003 as the General Chair; and recently ICIP’2010 as the General Chair (2010 IEEE International Conference on Image Processing, which was held in Hong Kong, 26-29 September 2010). Professor Siu is also the President Elect (2011-13) of a new professional association, the “Asia-Pacific Signal and Information Processing Association”, APSIPA. He is a member (2010-2012) of the Engineering Panel and also was a member of the Physical Sciences and Engineering Panel (1991-1995) of the Research Grant Council (RGC), Hong Kong Government. In 1994, he chaired the first Engineering and Information Technology Panel of the Research Assessment Exercise (RAE) to assess the research quality of 19 departments from all universities in Hong Kong. (http://www.eie.polyu.edu.hk/~wcsiu/mypage.htm)