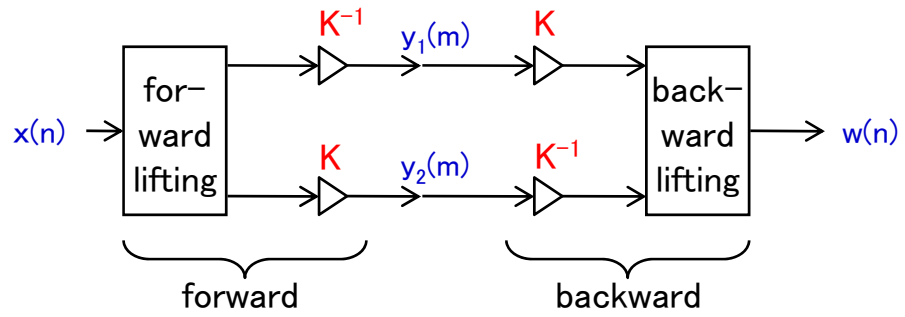
Finite Word Length Error Analysis based on Basic Formula of Rounding Operation

Masahiro IWAHASHI
Nagaoka University of Technology
Nagaoka-shi, Niigata, 940-2188 Japan

Hitoshi KIYA .
Tokyo Metropolitan University
Hino-shi, Tokyo, 191-0065 Japan

Wavelet Transform & Perfect Reconstruction (PR)

Wavelet Transform



Perfect Reconstruction

$$w(n) - x(n) = 0$$

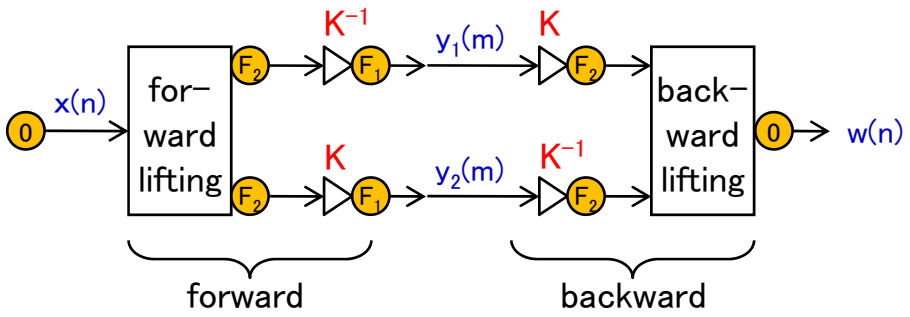
- Signals $\in$ Real Number
- Coefficients $\in$ Real Number

assumption (design)

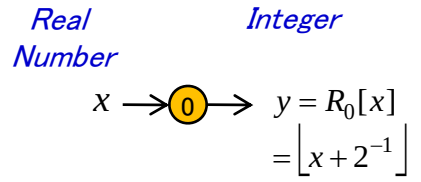
Effect of Rounding

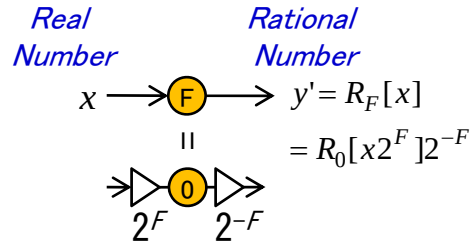
--- Finite Word Length ---

Wavelet Transform



Rounding





~~Perfect Reconstruction~~

$w(n) - x(n) \neq 0$

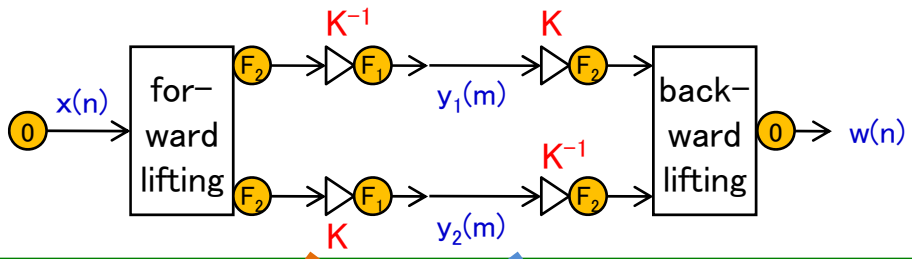
- Signals ∈ Integer or Rational Number
- Coefficients ∈ Rational Number

⇒ implementation (circuit)

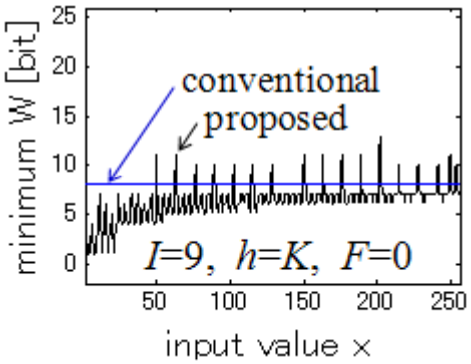
Find the "Minimum Word Length"

which satisfies the Perfect Reconstruction (PR)

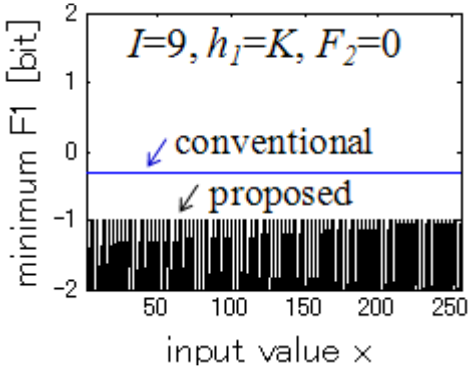
Wavelet Transform



Word Length of coefficients



Word Length of signals



Precisely Estimate the Minimum Value

Analysis

Conventional ...

L^∞ norm
⇒ precise,
⇒ too much, too strict

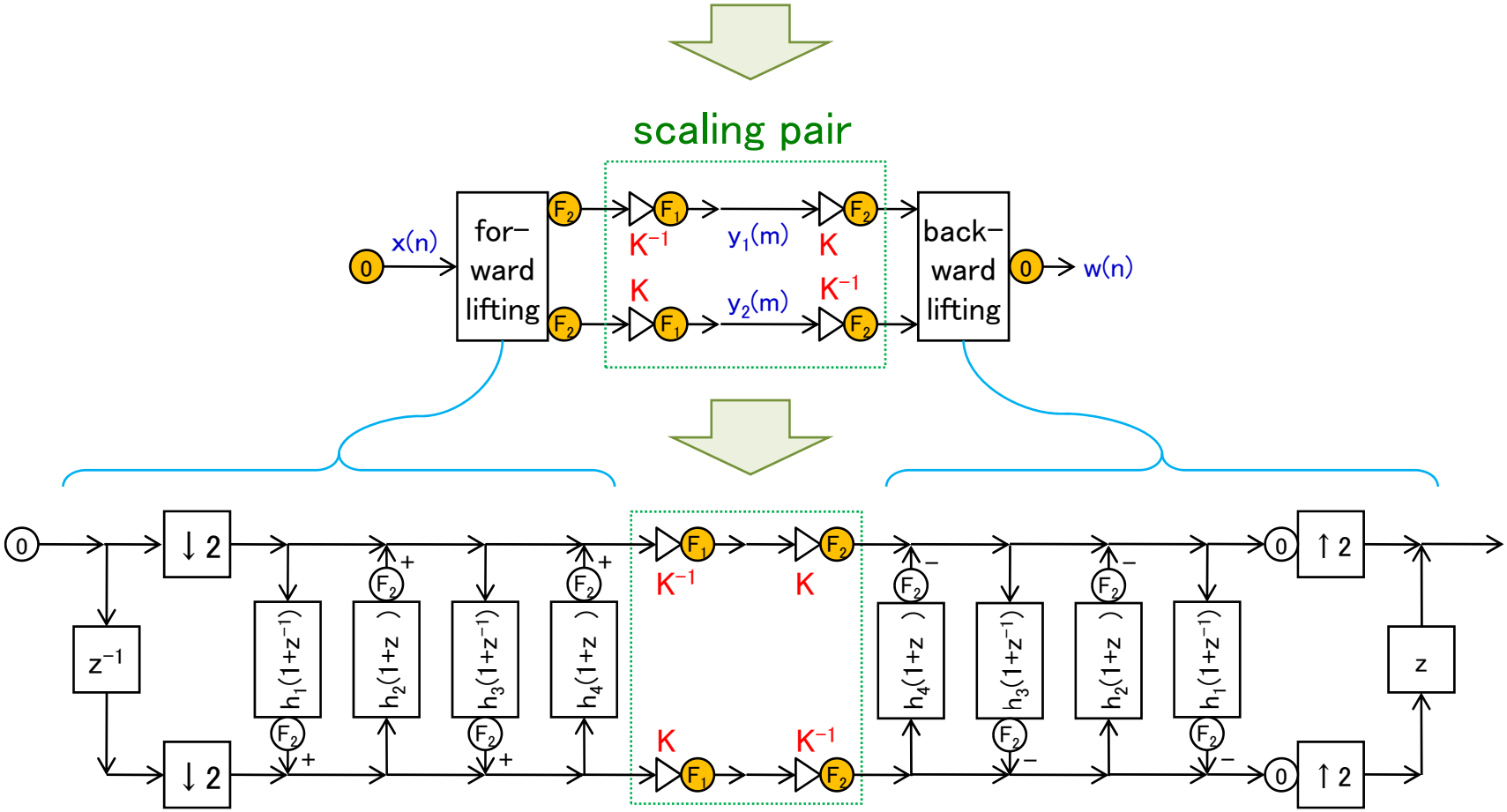
L^2 norm
⇒ Not Precise
⇒ not satisfactory

We propose ...

New analysis
⇒ Precise
⇒ exactly same as experimental results

Conditions on Signals & Coefficients

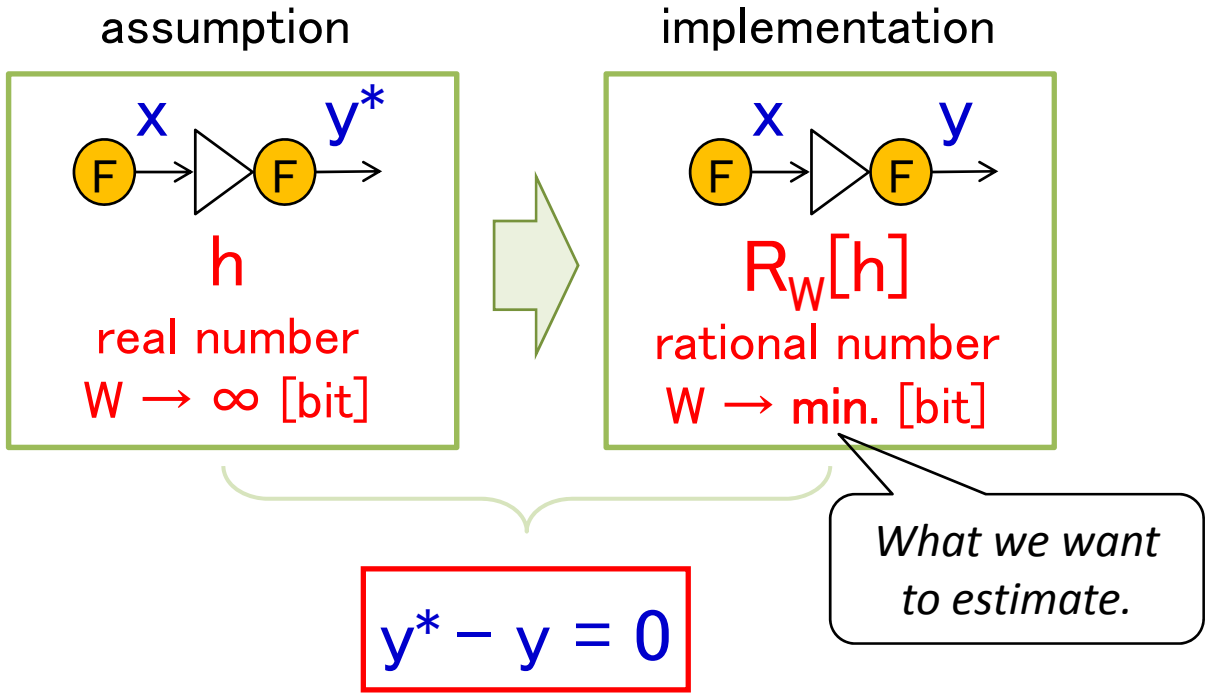
Focus on Scaling



cause of

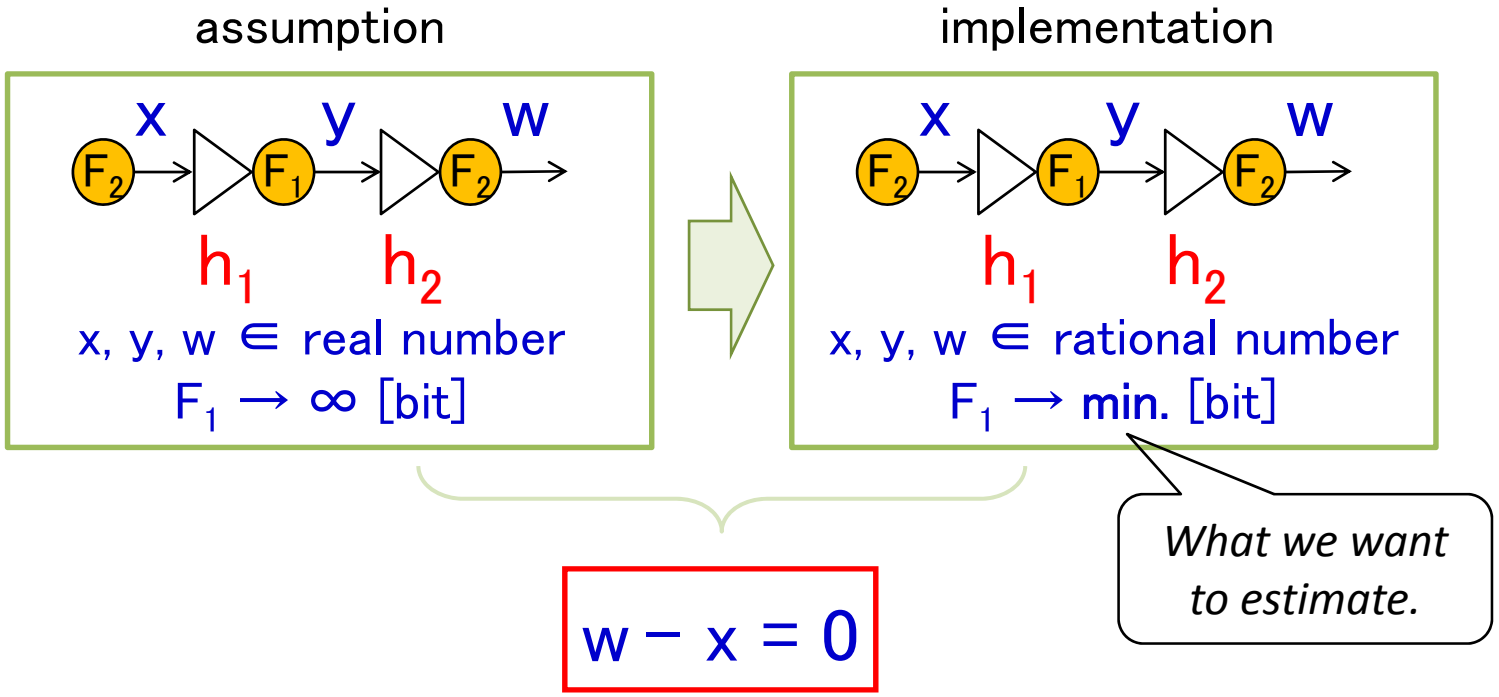
$w(n) - x(n) \neq 0$

[A] Condition on Coefficient



Mapping Invariant Condition

[B] Condition on Signals

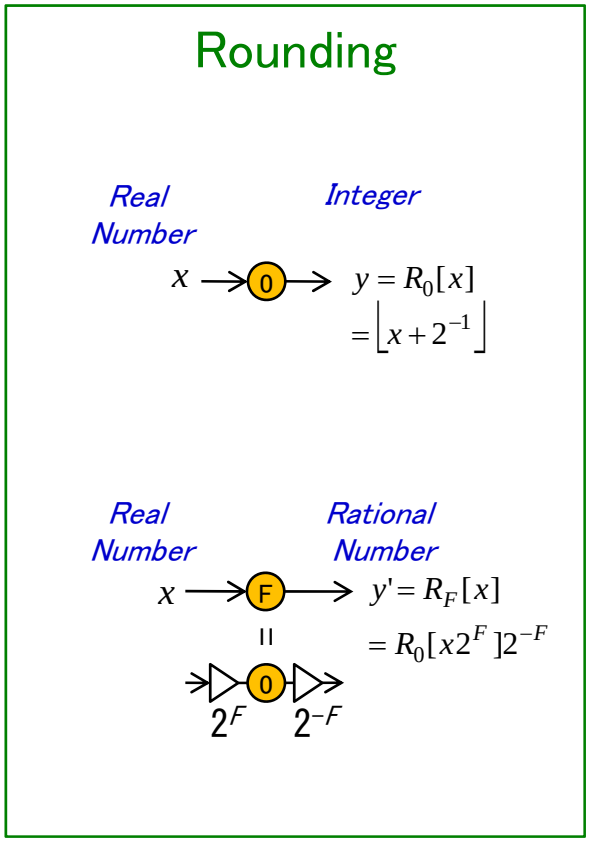


Lossless Scaling Pair Condition

"F_2" is given and "$h_1 h_2 = 1$"

Estimation

Expression of Rounding



Rounding to Integer

$$\begin{cases} R_0[x] = \lfloor x + 2^{-1} \rfloor = x' - (x' \mod 1) \\ x' = x + 2^{-1} \end{cases}$$

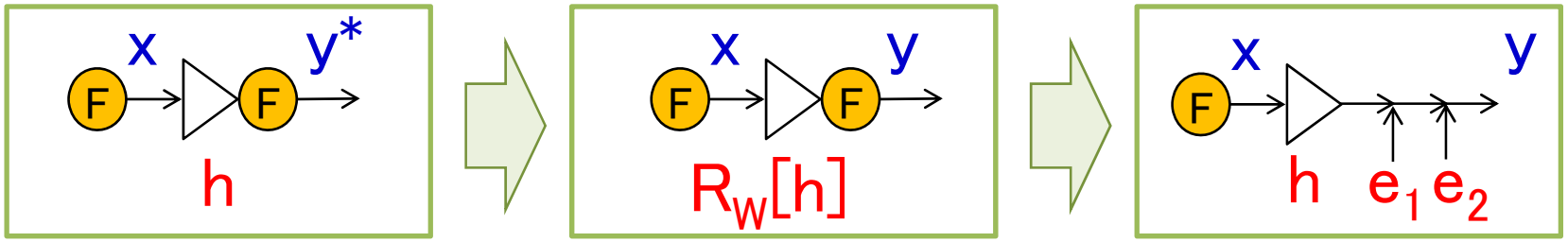
rounding error

$$\Delta_0[x] = x - R_0[x]$$
$$= \{(x + 2^{-1}) \mod 1\} - 2^{-1}$$

Rounding to Rational Number

$$\begin{cases} R_F[x] = R_0[x2^F]2^{-F} \\ \Delta_F[x] = x - F[x] = \Delta_0[x2^F]2^{-F} \end{cases}$$

[A] Condition on Coefficient



$$\begin{cases} y = hx + e_1 + e_2 \\ e_1 = -\Delta_W[h]x \\ e_2 = -\Delta_F[W[h]x] \end{cases}$$

Conventional Analysis

L^∞ norm
⇒ precise,
⇒ too much, too strict

L^2 norm
⇒ Not Precise
⇒ not appropriate

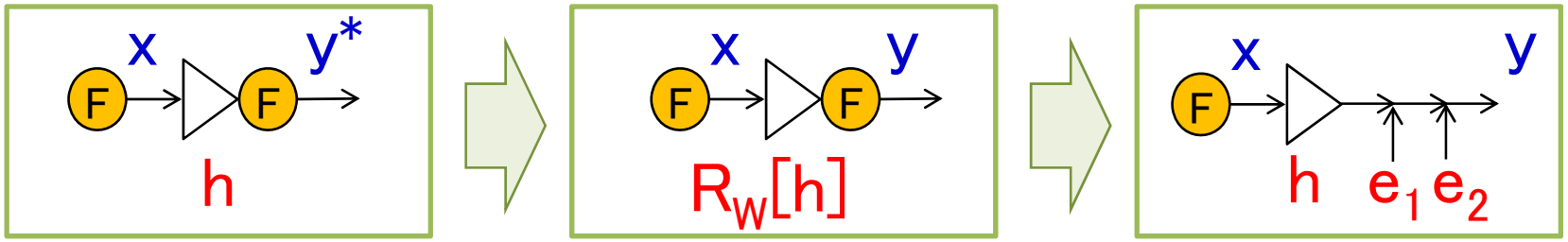
$$\max|e_1| = 2^{-1-W}2^{-1+I} < \max|e_2| = 2^{-1-F}$$

$$\sigma_{e_1}^2 = \frac{1}{12}\left(2^{-1-W}2^{-1+I}\cdot 2\right)^2 << \sigma_{e_2}^2 = \frac{1}{12}\left(2^{-1-F}\cdot 2\right)^2$$

$$W > F + I - 1$$

estimated.

[A] Condition on Coefficient



$$\begin{cases} y = hx + e_1 + e_2 \\ e_1 = -\Delta_W[h]x \\ e_2 = -\Delta_F[W[h]x] \end{cases}$$

Our Analysis

New analysis
⇒ Precise
⇒ exactly same as experimental results

$$E_m = 0 \quad for \quad \begin{cases} E_m = R_F[R_W[h]x] - R_F[hx] \\ R_W[h] = h - \Delta_W[h] \end{cases}$$

$$\left|\Delta_W[h]x - \Delta_F[hx]\right| < 2^{-1-F}$$ ***estimated.***

[A] *proof*

$$E_m = 0 \qquad for \qquad \begin{cases} E_m = R_F\left[R_W[h]x\right] - R_F[hx] \\ R_W[h] = h - \Delta_W[h] \end{cases}$$

proof

$$\begin{aligned} & R_F\left[R_W[h]x\right] - R_F[hx] \\ &= R_F\left[hx - \Delta_W[h]x\right] - R_F[hx] \\ &= R_F\left[\Delta_F[hx] - \Delta_W[h]x\right] + R_F[hx] - R_F[hx] \\ &= R_F\left[\Delta_F[hx] - \Delta_W[h]x\right] = 0 \\ \therefore \quad & \left|\Delta_F[hx] - \Delta_W[h]x\right| < 2^{-1-F} \end{aligned}$$

Q.E.D.

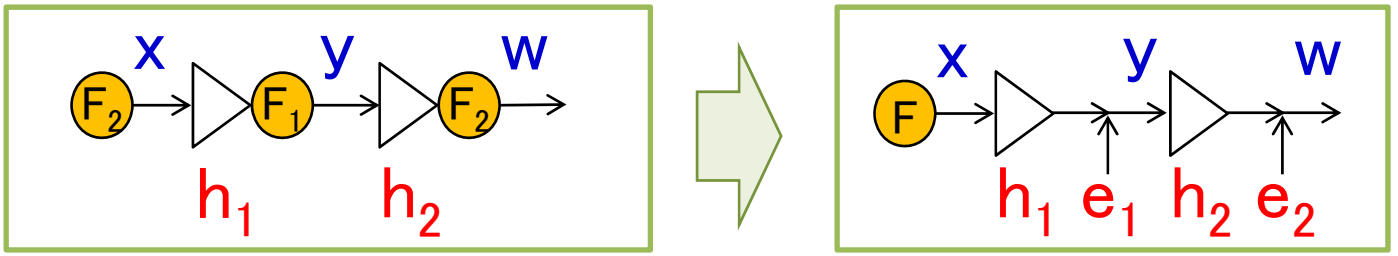
Addition Formula

$$R_F[x+y] = R_F[x] + R_F\left[y + \Delta_F[x]\right]$$
$$x \in \mathbf{R}, y \in \mathbf{R}$$

Properties

$$\begin{cases} R_0[x] = 0 \Leftrightarrow x \in [-2^{-1}, 2^{-1}) \\ R_F[x] = 0 \Leftrightarrow x \in [-2^{-1-F}, 2^{-1-F}) \end{cases}$$

[B] Condition on Signals



$$\begin{cases} w = h_2 y + e_2 \\ y = h_1 x + e_1 \end{cases}, \qquad \begin{cases} e_1 = -\Delta_{F_1}[h_1 x] \\ e_2 = -\Delta_{F_2}[h_2 y] \end{cases}$$

Conventional Analysis

L^∞ norm
⇒ precise,
⇒ too much, too strict

L^2 norm
⇒ Not Precise
⇒ not appropriate

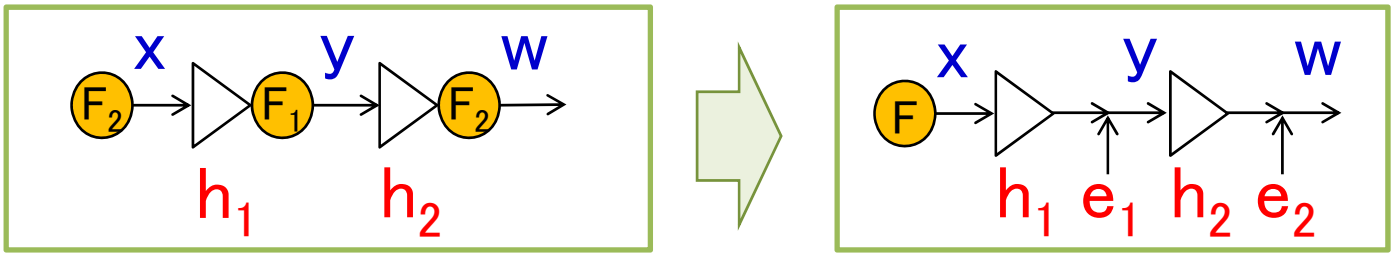
$$\max|h_2 e_1| = |h_2| 2^{-1+F_1} < \max|e_2| = 2^{-1-F_2}$$

$$\sigma^2_{h_2 e_1} = \frac{1}{12}\left(h_2 2^{-1+F_1} \cdot 2\right)^2 << \sigma^2_{e_2} = \frac{1}{12}\left(2^{-1-F_2} \cdot 2\right)^2$$

$$F_1 > F_2 + \log_2|h_2|$$

estimated.

[B] Condition on Signals



$$\begin{cases} w = h_2 y + e_2 \\ y = h_1 x + e_1 \end{cases}, \qquad \begin{cases} e_1 = -\Delta_{F_1}[h_1 x] \\ e_2 = -\Delta_{F_2}[h_2 y] \end{cases}$$

Our Analysis

New analysis
⇒ Precise
⇒ exactly same as experimental results

$$E_p = 0 \quad for \quad \begin{cases} E_p = R_{F_2}\left[h_2 R_{F_1}[h_1 x]\right] - x \\ h_1 h_2 = 1 \end{cases}$$

$$\left| h_2 \Delta_{F_1}[h_1 x] \right| < 2^{-1-F_2}$$ ***estimated.***

[B] *proof*

$$R_{F_2}\left[xR_{F_1}[y]\right]$$
$$= R_{F_2}[xy] + R_{F_2}\left[-x\Delta_{F_1}[y] + \Delta_{F_2}[xy]\right]$$

Multiplication Formula

$$R_F[xy]$$
$$= R_F\left[xR_F[y]\right] + R_F\left[x\Delta_F[y] + \Delta_F[xR_F[y]]\right]$$
$$x \in \mathbf{R}, y \in \mathbf{R}$$

proof

$$R_{F_2}\left[xR_{F_1}[y]\right]$$
$$= R_0\left[-x\Delta_{F_1}[y]2^{F_2} + xy2^{F_2}\right]2^{-F_2}$$
$$= R_0\left[-x\Delta_{F_1}[y]2^{F_2} + \Delta_0[xy2^{F_2}]\right]2^{-F_2} + R_{F_2}[xy]$$
$$= R_{F_2}\left[-x\Delta_{F_1}[y] + \Delta_{F_2}[xy]\right] + R_{F_2}[xy]$$

Q.E.D.

$$R_{F_2}\left[h_2\Delta_{F_1}[h_1x]\right] = 0$$
$$\left|h_2\Delta_{F_1}[h_1x]\right| < 2^{-1-F_2}$$

Properties

$$\begin{cases} R_0[x] = 0 \Leftrightarrow x \in [-2^{-1}, 2^{-1}) \\ R_F[x] = 0 \Leftrightarrow x \in [-2^{-1-F}, 2^{-1-F}) \end{cases}$$

Proof of the Formulas

Addition Formula

$$R_F[x+y] = R_F[x] + R_F\left[y + \Delta_F[x]\right]$$
$$x \in \mathbf{R}, y \in \mathbf{R}$$

proof

$$\begin{aligned} & R_F\left[x+y\right] \\ &= R_F\left[R_F[x] + \Delta_F[x] + y\right] \\ &= R_0\left[R_0[x2^F] + \Delta_F[x]2^F + y2^F\right]2^{-F} \\ &= R_F[x] + R_F\left[\Delta_F[x] + y\right] \end{aligned}$$

Q.E.D.

Multiplication Formula

$$\begin{aligned} & R_F[xy] \\ &= R_F\left[xR_F[y]\right] + R_F\left[x\Delta_F[y] + \Delta_F[xR_F[y]]\right] \end{aligned}$$
$$x \in \mathbf{R}, y \in \mathbf{R}$$

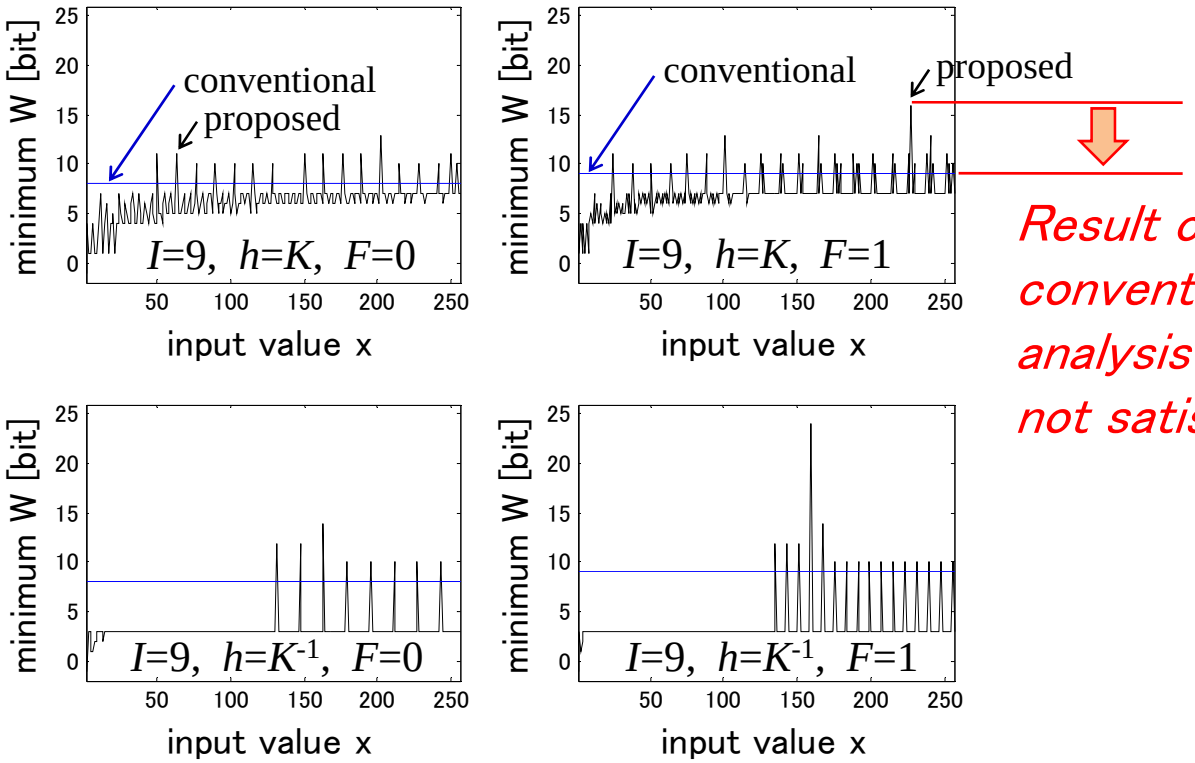
proof

$$\begin{aligned} & R_F[xy] \\ &= R_F\left[x\Delta_F[y] + xR_F[y]\right] \\ &= R_F\left[x\Delta_F[y] + \Delta_F\left[xR_F[y]\right] + R_F[xR_F[y]]\right] \\ &= R_F\left[x\Delta_F[y] + \Delta_F[xR_F[y]]\right] + R_F\left[xR_F[y]\right] \end{aligned}$$

Q.E.D.

Results

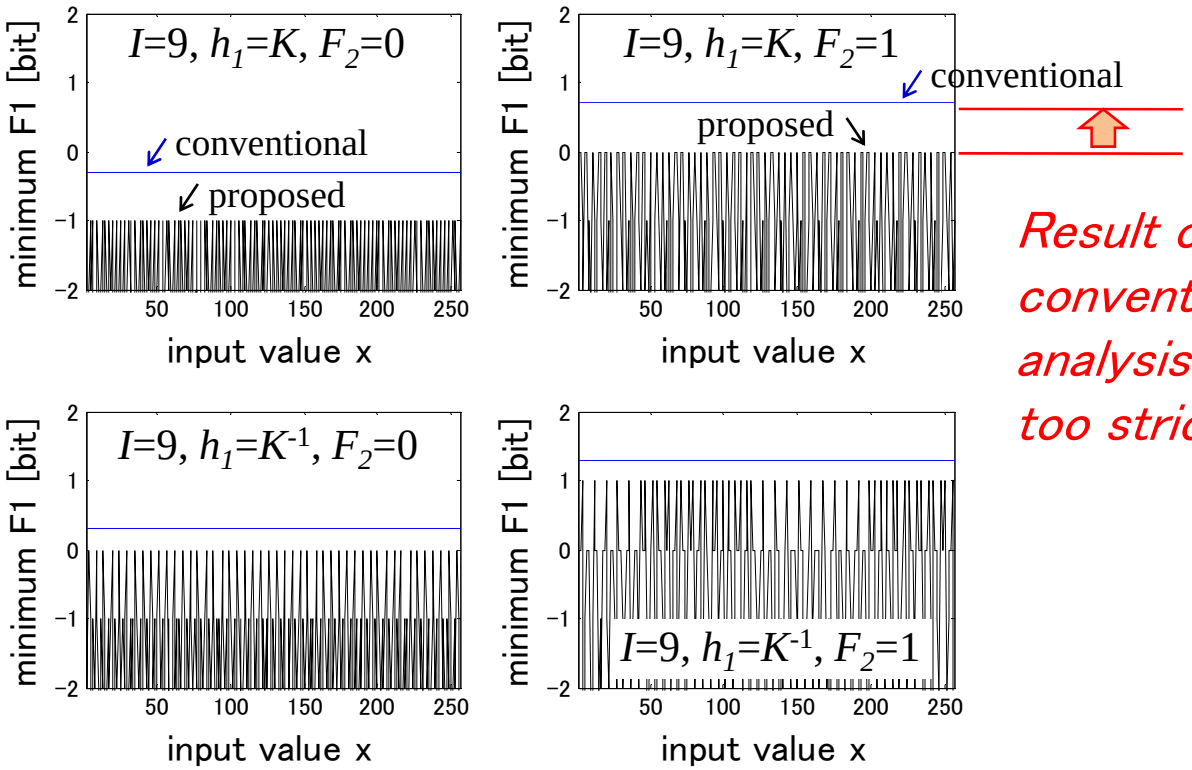
Min. Word Length of Coefficients



Result of the conventional analysis is not satisfactory.

$$W_{\min} = \{W \in \mathbf{Z} \mid E_m = 0, \ \forall W \geq W_{\min}\}$$

Min. Word Length of Signals



Result of the conventional analysis is too strict.

$$F_{1,\min} = \{F_1 \in \mathbf{Z} \mid \ E_p = 0, \ \ \forall F_1 \geq F_{1,\min}\}$$

Conclusion

Conclusion Remarks

In this report, we mathematically derived formulas of the rounding operation

to analyze the lossless condition on word length of signals and coefficients.

It was confirmed that our equations can determine the minimum word length for each input value more precisely than the conventional analysis.