DATA QUALITY METRIC (DQM) – HOW ACCURATE DOES IT NEED TO BE?

Mark Geoghegan
Quasonix, West Chester, OH
markg@quasonix.com

ABSTRACT

A Data Quality Encapsulation (DQE) protocol for improving telemetry link quality has recently been standardized and added to IRIG 106. It periodically provides a Data Quality Metric (DQM) required for optimal Maximum Likelihood Bit Detection (MLBD) when more than one receive source is available. The resulting diversity can provide tremendous improvements in link quality. In order to be effective, the estimated DQM value should respond quickly and accurately to reflect the actual Bit Error Probability (BEP).

This paper investigates the MLBD performance loss caused by DQM estimation error. The objective is to gain insight into the sensitivity of the overall bit recovery system and to use the results to help establish tolerance levels in DQM test procedures. This relationship provides the means to guarantee that the DQM accuracy is sufficient to meet or exceed a specified level of system performance which is the goal of DQM testing.

KEYWORDS

Data Quality Metric (DQM), Data Quality Encapsulation (DQE), MLBD

INTRODUCTION

A Data Quality Encapsulation (DQE) protocol for improving telemetry link quality has recently been standardized and added to IRIG 106 [1]. Current DQM test methods, such as those described in RCC IRIG 118-22 Chapter 11 – "Test Procedures for Assessing Telemetry Receiver Data Quality Metrics", describe detailed test procedures and identify specific data to be collected. However, it is unclear whether the observed DQM estimation performance is acceptable or not. In other words, how good is good enough?

To answer this question, one needs to understand how DQM accuracy impacts the performance of the overall bit recovery process which is ultimate measure of system performance. The approach to quantify how good is good enough begins with examining the basic characteristics of DQM and the optimal bit processing method described by Rice and Perrins [2]. Next, mathematical analysis and simulations will be used to evaluate the role that DQM plays in the bit recovery process. After this relationship is developed, a method of establishing thresholds on DQM accuracy that bounds the worst-case system BEP degradation will be presented.

DQE/DQM BACKGROUND

When multiple receive channels are available as shown in Figure 1, the performance of the communication link may be significantly improved by combining the individual sources into a single 'best' composite stream. There are many approaches to combining including selective and majority vote. Selective combining makes its decision solely based on the strongest individual channel, while majority vote chooses the bit with the most votes as its combined decision. For channels where one path is clearly dominant, selective combining works well while majority vote performs poorly. Conversely, when all paths are of similar quality, majority vote works well and selective combining is poor. The optimal combining approach makes its bit decision by comparing the weighted sums of channel outputs of 0 and 1 and outperforms all other methods. The weighting is based on the log-ratio of the error transition probability for each received channel.

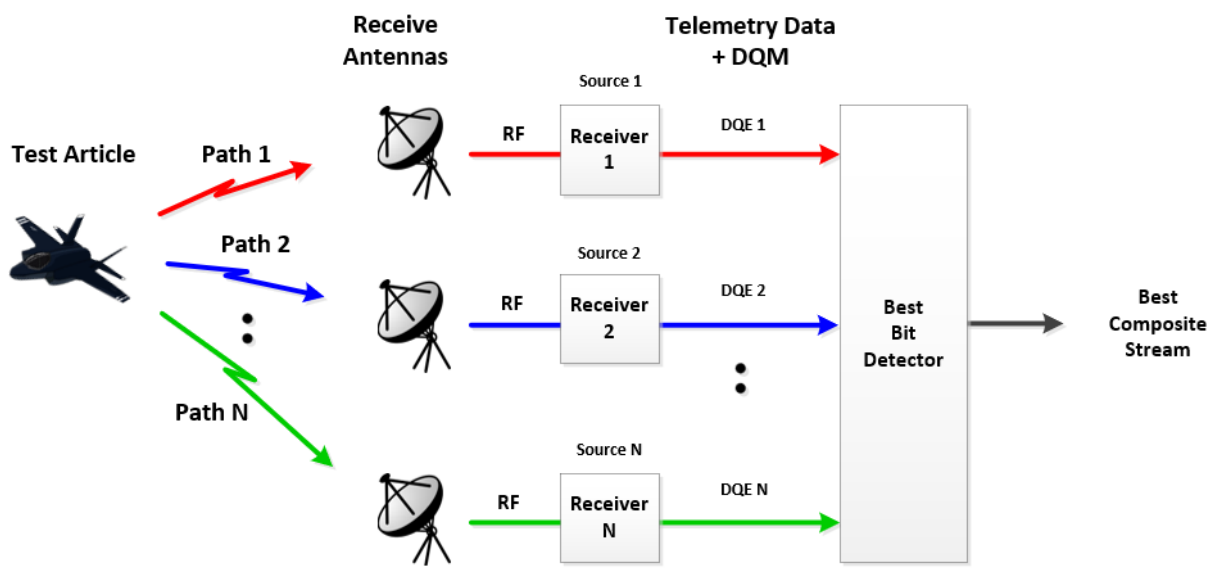
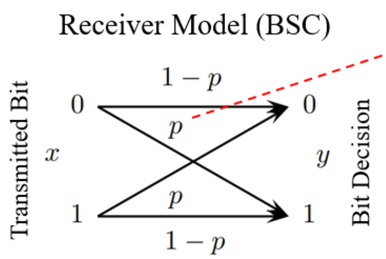


Figure 1: Telemetry System with Multiple Receive Channels and Best Source Selector.

Optimal combining requires both the individual channel decisions as well as the current channel error transition probabilities. A method for transporting telemetry data along with a data quality metric is described in the RCC 106-22 telemetry standards chapter 2 section 2.7 appendix 2-G and is a vital ingredient for supporting optimal Best Source Selection. It allows telemetry receivers to generate a serial stream that includes a standardized measurement of the real-time probability of error for a grouping of bits. The DQE process is illustrated in Figure 2 where the receiver translates the error transition probability p into a 16-bit DQM value that accompanies N bits of telemetry payload. The basis for the protocol was initially described in “Metrics and Test Procedures for Data Quality Estimation in the Aeronautical Telemetry Channel” by Hill [3]. The optimal processing technique was presented in “Maximum Likelihood Detection from Multiple Bit Sources” by Rice and Perrins [2]. Further work on how to test this capability was described in “Some Thoughts on Testing the Data Quality Metric” by Temple [4]. Finally, approved test methods for evaluating DQM performance were added to IRIG 118-22 Vol. 2 [5].



BEP Versus DQM

BEP	LR	DQM
0.5	1.00	0
1e−01	1.11111e−01	5211
1e−02	1.01010e−02	10898
1e−03	1.00100e−03	16381
1e−04	1.00010e−04	21845
1e−05	1.00001e−05	27306
1e−06	1.00000e−06	32767
1e−07	1.00000e−07	38229
1e−08	1.00000e−08	43690
1e−09	1.00000e−09	49151
1e−10	1.00000e−10	54613
1e−11	1.00000e−11	60074
1e−12	1.00000e−12	65535

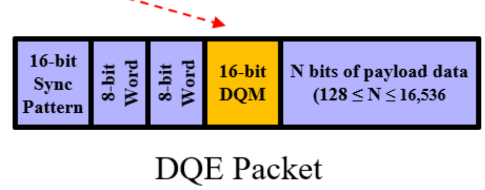


Figure 2: Data Quality Encapsulation (DQE) Packet Generation

EFFECT OF DQM INPUT ACCURACY ON MLBD ERROR PERFORMANCE

Although specific tests are defined for various channel conditions (AWGN, Flat Fading, Adjacent Channel Interference, and Multipath), there is no criteria or defined thresholds for what level of DQM accuracy is acceptable. The issue is that it is unclear how errors in DQM estimation degrade the system BEP performance. Understanding this relationship would allow a telemetry engineer to specify a maximum system degradation which would then bound the DQM estimation accuracy. For example, imagine an ideal bit recovery device (perfect DQM + MLBD) that can operate at an output BEP of 10^{-8}. If one is willing to accept 10^{-7} from a real implementation, the limits on DQM accuracy to guarantee that level of performance could be determined.

It is important to recognize that the optimal MLBD algorithm is a fixed set of deterministic calculations. The only quantities that are estimated are the input transition probabilities $p_1, p_2, ..., p_N$. It is analogous to a card game in which the channel outputs and their probabilities are the cards that are dealt and the MLBD algorithm is a procedure to perfectly play each hand for maximum possible return (lowest error probability). Accurately representing each hand and using the ‘perfect play’ recipe guarantees ideal play results. However, it will be shown that misestimating p can cause a serious loss in MLBD performance.

What happens when the estimate $\hat{p}$ differs from the actual value of p? Figure 3 shows the results of extreme over and under estimation using a channel example presented in [2]. With N=3, $\boldsymbol{p}$ = [10^{-4}, 10^{-2}, 10^{-3}] the MLBD was shown to have a BEP of 1.1098 x 10^{-5} which has approximately one order of error magnitude improvement over the best individual channel (p_1 = 10^{-4}). Now imagine that p_1 was optimistically misjudged and presented as $\hat{p}_1$ = 10^{-12}. The MLBD algorithm would now always choose y_1 resulting in an output error rate of 10^{-4} which degrades the ideal system error performance by one order of magnitude. Similarly, an overly pessimistic estimate of $\hat{p}_1$ = 0.5 would always cause channel 1 to be ignored and channel 3 to be selected resulting in an output error rate of 10^{-3}, a loss of two orders of error magnitude.

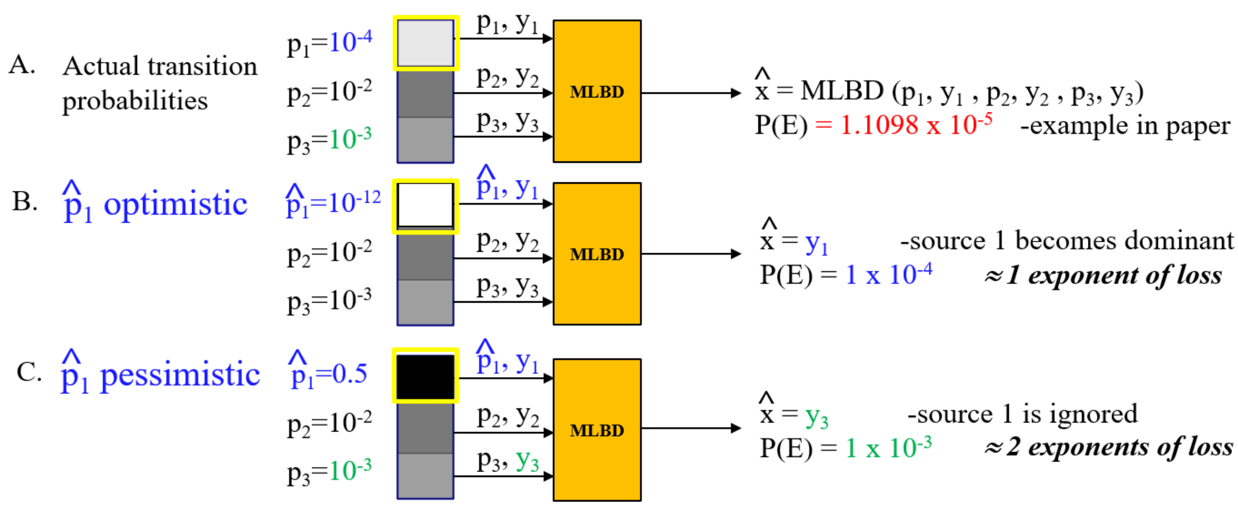


Figure 3: Example showing effect of extreme over and under estimation of p

A detailed performance analysis of the MLBD detector was presented in [2] that describes a procedure for calculating its output error probability. First, all possible channel output tuples $[y_1, y_2, ..., y_N]$ were formed and indexed by their binary error pattern [i]. They are separated into two sets ($\mathcal{N}_0$ or $\mathcal{N}_1$) based on their y_n value. Assuming x = 0 was transmitted, the probability of the i-th error pattern P[i] was computed by (1). For each error pattern, the MLBD metric M[i] is calculated by summing the log-ratio of the transition probabilities p_n for each set and taking their difference as shown in (2). The set of indexes $\mathcal{I}_+$ identify which M[i] > 0 indicating tuples where an error occurs. Finally, the sum of the probabilities over this set yield the final MLBD error probability (3).

$$P[i] = \prod_{n \epsilon \mathrm{N}_0[i]} (1 - p_n) \prod_{n \epsilon \mathrm{N}_1[i]} (p_n)\,. \tag{1}$$

$$M[i] = \sum_{n \epsilon \mathrm{N}_1[i]} \log\left(\frac{1 - p_n}{p_n}\right) - \sum_{n \epsilon \mathrm{N}_0[i]} \log\left(\frac{1 - p_n}{p_n}\right). \tag{2}$$

$$P_e = \sum_{i \epsilon \mathcal{I}_+} P[i] \tag{3}$$

If the transition probabilities are estimated they may be different that the actual p values. The probability of the channel outputs P[i] do not change and are computed with the actual p values as before. However, the MLBD metrics $\hat{M}$[i] use the estimated values $\hat{p}$ and can therefore change which cases $\hat{\mathcal{I}}_+(\hat{M}[\text{i}] > 0)$ are perceived as errors. Consequently, the MLBD output error rate is $\hat{P}_e$. The equations that incorporate estimation show the linkage between misestimation of p (DQM) and MLBD performance degradation. Also, note that if there is a common estimation bias across all channels, there is a tendency for it to cancel out since the MLBD metric $\hat{M}$[i] is the difference of the log-ratio sums that contain identical bias terms as seen below.

$$\hat{M}[i] = \sum_{n \epsilon \mathrm{N}_1[i]} \log\left(\frac{1-\hat{p}_n}{\hat{p}_n}\right) - \sum_{n \epsilon \mathrm{N}_0[i]} \log\left(\frac{1-\hat{p}_n}{\hat{p}_n}\right). \tag{4}$$

$$\hat{P}_e = \sum_{i \epsilon \hat{\mathcal{I}}_+} P[i] \tag{5}$$

$$\log\left(\frac{1-\hat{p}_n p_{bias}}{\hat{p}_n p_{bias}}\right) \xrightarrow{\hat{p}_n p_{bias} \ll 0} -\log\left(\hat{p}_n\right) - \log\left(p_{bias}\right) \tag{6}$$

A second example illustrates the details of these calculations and shows the impact of even a small estimation error in p. Table 1 compares the calculations for the ideal case $p_1 = 10^{-4}$ and the case when $\hat{p}_1$ is misestimated as 10^{-5}. Assuming x = 0 was transmitted, the P[i] values are calculated as in [2] as the product of the transition probabilities (either p or (1-p)) depending on the output y_n[i]. The ideal MLBD metrics are calculated using (2) and the error cases with M[i] $>$ 0 are identified. The corresponding P[i] values (i=3,5,6,7) are summed to obtain 1.1098 x 10^{-5} as the MLBD error output probability (red values). The blue values correspond to the calculations when $\hat{p}_1 = 10^{-5}$. This causes M[4] to increase and M[3] (the error complement term) to decrease such that P[4] is now perceived as an error resulting in 1.0000 x 10^{-4} as the MLBD error output probability. This shows how a small error in p estimation can seriously degrade the MLBD detector performance.

Although these two examples show how errors in p (or DQM) estimation can adversely impact the output error rate, it is still unclear on how to determine or bound the cost of misestimation over the entire range of channel inputs. In order to develop a more general understanding of this relationship, detailed mathematical and simulation work was performed over a wide range of channels with estimation errors. A surprising result was obtained and verified through both simulation and analysis. **The performance degradation of the MLBD in error exponents is bounded by the DQM estimation error exponent. In other words, the worst-case output degradation changes exponentially 1 to 1 with the input estimation error.** This was not expected since small improvements in error rate at the input can result in remarkable improvements at the output. This relationship provides the means to specify accuracy limits on DQM estimation and guarantee a minimum level of MLBD performance and is shown in (7) below.

$$\boxed{\Delta P_{e_{\exp}} \;\leq\; |\Delta p_{1_{\exp}}|} \tag{7}$$

Table 1: MLBD performance with imperfect p_1 estimate. Example of how a small estimation error in p_1 can cause a similar degradation in MLBD error probability.

N=3, $p_1 = 10^{-4}$, $p_2 = 10^{-3}$, $p_3 = 10^{-2}$					Ideal Metrics ($p_1 = 10^{-4}$)			Estimated Metrics ($\hat{p}_1 = 10^{-5}$)		
i	y1	y2	y3	Probability P[i]	$\sum_{N_0}$	$\sum_{N_1}$	$M[i]$	$\sum_{N_0}$	$\sum_{N_1}$	$\hat{M}[i]$
0	0	0	0	9.8891 x 10^{-1}	20.7	0	-20.7	23.0	0	-23.0
1	0	0	1	9.9890 x 10^{-3}	16.1	4.6	-11.5	18.4	4.6	-13.8
2	0	1	0	9.8990 x 10^{-4}	13.8	6.9	-6.9	16.1	6.9	-9.2
3	0	1	1	* 9.9991 x 10^{-6}	9.2	11.5	* **2.3**	11.51	11.50	-0.01
4	1	0	0	9.8901 x 10^{-5} *	11.5	9.2	-2.3	11.50	11.51	**0.01** *
5	1	0	1	* 9.9900 x 10^{-7} *	6.9	13.8	* **6.9**	6.9	16.1	**9.2** *
6	1	1	0	* 9.9000 x 10^{-8} *	4.6	16.1	* **11.5**	4.6	18.4	**13.8** *
7	1	1	1	* 1.0000 x 10^{-9} *	0	20.7	* **20.7**	0	23.0	**23.0** *
P[i]'s do not change due to estimation					P(E | 0)=1.1098 x 10^{-5}			$\hat{P}(E \mid 0)$=1.0000 x 10^{-4}		

SIMULATION RESULTS

This section presents a summary of the simulation results over a range of N, p's, and $\hat{p}$'s. For each case, the ideal MLBD performance was computed along with the results when p_1 was misestimated over the entire range of p values. The results for N=3, 5 and 7 are shown in Figure 4 below. The graph plots the degradation in the MLBD error exponent $\Delta P_{e_{\text{exp}}}$ versus the misestimation of transition probability exponent $\Delta p_{1_{\text{exp}}}$. There are many cases for which the MLBD output is unaffected by p_1 misestimation. In the case of dominant channels, even sizable estimation errors may not change the ideal MLBD result. Other channel combinations may be degraded up to a point and then become either dominant or ignored. The most important takeaway is that the change in MLBD error exponent does not exceed the change in the transition probability exponent. This supports the conclusion of the mathematical analysis in the Appendix and provides a method to bound DQM accuracy based on the maximum allowable degradation in MLBD error.

APPLICATION TO RCC IRIG 118

Although test methods have been developed to measure the consistency and accuracy of receiver output data quality metrics (DQM), there is not good understanding on what level of performance is acceptable. Regardless of the specific test, it is envisioned that a key piece of test equipment capable of capturing the DQE frames, extracting the DQM values, and measuring the estimated BEP with the actual BER will be needed. The primary output is a DQM correlation plot that shows

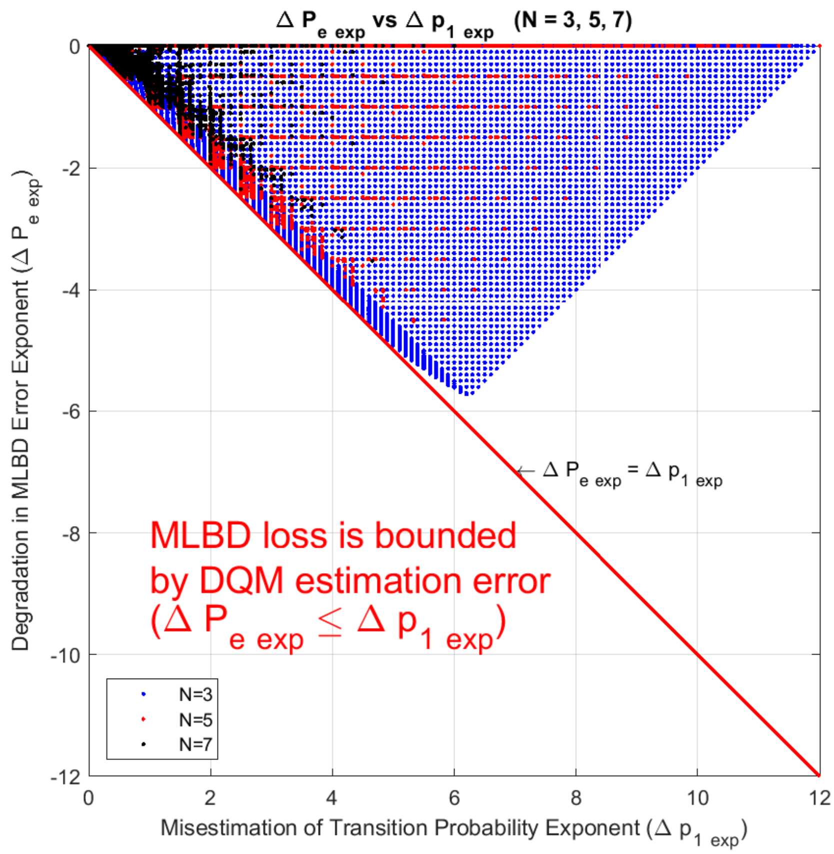


Figure 4: MLBD Degradation vs DQM Misestimation Plot

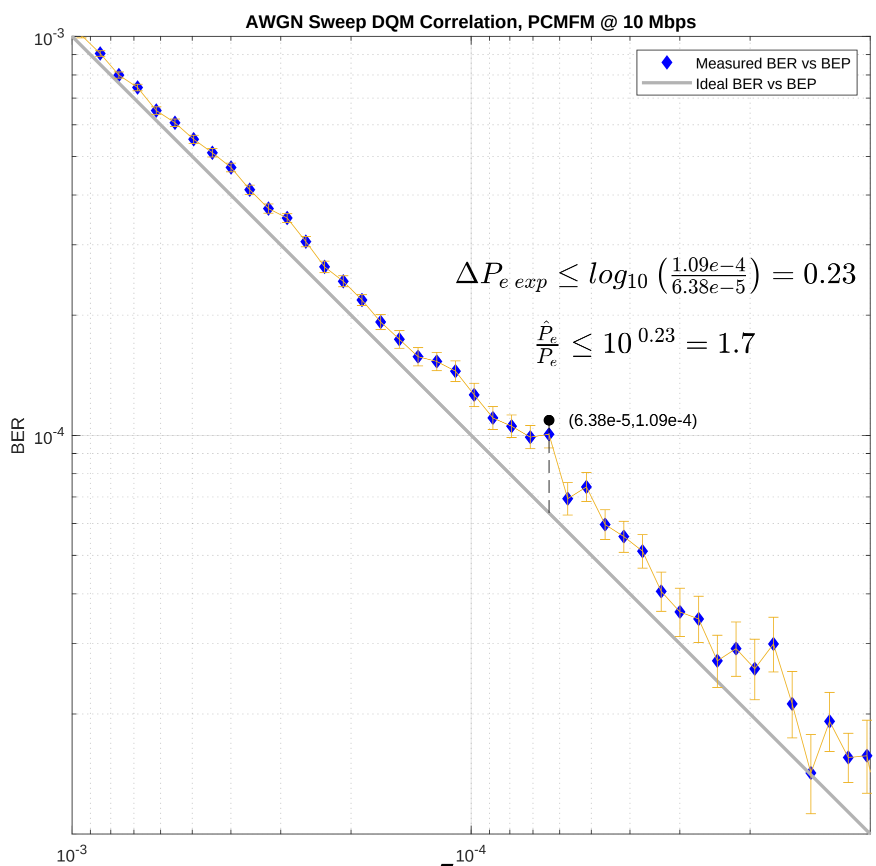


Figure 5: DQM Correlation Plot with Bound Calculation Example

how the estimated BEP value (DQM) compares to the actual BER for that frame. In practice, the 65,536 possible DQM values are grouped into a smaller number of bins to achieve a balance of resolution and efficiency and as described in [3].

A section of a DQM correlation plot is presented in Figure 5 showing the mean along with its corresponding confidence interval. Ideal DQM estimates will match the actual BER for that packet and lie on the Ideal vs BEP line. In practice, the DQM values are estimates and are represented by a mean and confidence interval. To show how the performance bound can be used, the DQM error at a single point on the graph is 0.23 error exponents which translates to a worst-case degradation factor of 1.7 times the ideal MLBD BEP. Figure 6 shows the entire DQM correlation plot with thresholds corresponding to various levels of system loss. Finally, Figure 7 shows how tapered thresholds can be applied which is well suited for typical test purposes.

CONCLUSIONS

This paper investigated the MLBD performance loss caused by DQM estimation error. The objective was to quantify the system cost as a function of DQM accuracy to assist in establishing tolerance levels in DQM test procedures. Starting with the ideal MLBD performance analysis from [2], the results were extended to include the case where the DQM transition probability values were non-ideal. Examples were presented that showed how errors in estimating DQM could cause significant system degradation. Using mathematical analysis and simulations, a method of calculating the misestimation cost in terms of output performance was presented.

It was shown that the performance degradation of the MLBD in error exponents is bounded by the estimation error exponent in DQM. This is a powerful result that directly describes their relationship. More significantly, it provides a means to quantify what level of DQM accuracy is sufficient to meet or exceed a specified level of system performance. This performance bound was applied to a measured DQM correlation plot to illustrate how DQM accuracy thresholds could be added and what they imply regarding the worst-case MLBD degradation. It was also shown how these results can be used to customize a tapered threshold based on the region of BER operation. In summary:

- *Using multiple receive channels can significantly improve the system BEP.*
- *The DQE/DQM RCC IRIG standard supports optimal processing for multi-channel telemetry reception.*
- *Accurate DQM values are a vital ingredient in achieving good performance.*
- *Common DQM estimation biases between receivers tend to cancel due to the differential nature of the metric calculations.*
- *Testing and verification methods are currently being developed to ensure vendor compatibility and consistent performance.*
- *Results from this paper can assist standards organizations in developing meaningful testing thresholds.*

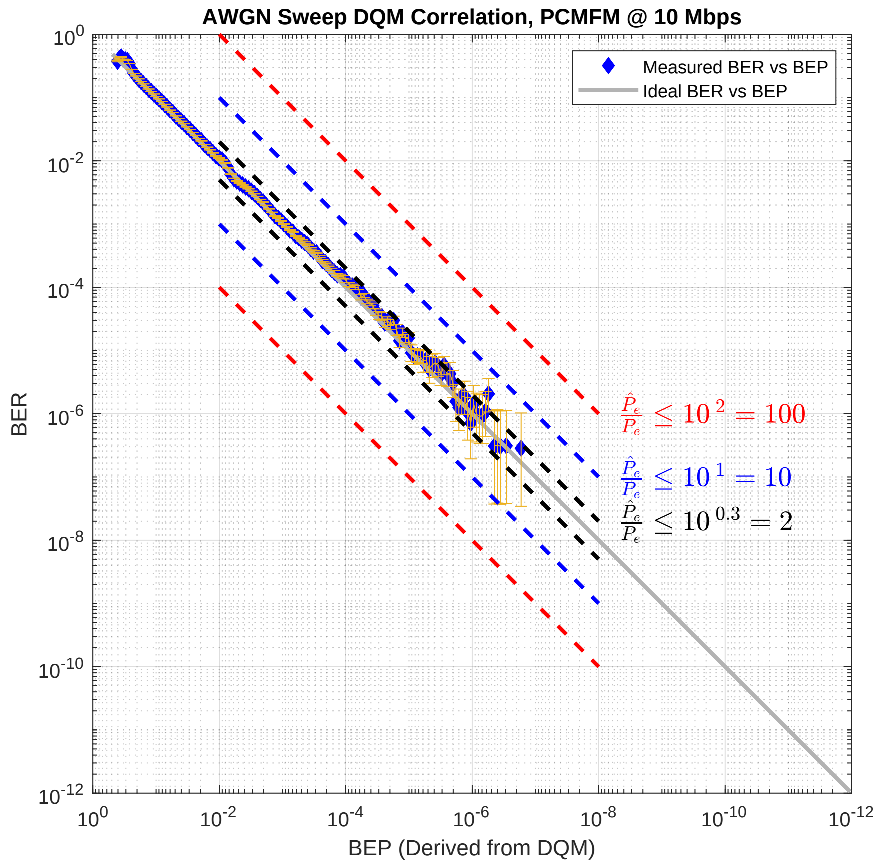


Figure 6: DQM Correlation Plot with various thresholds

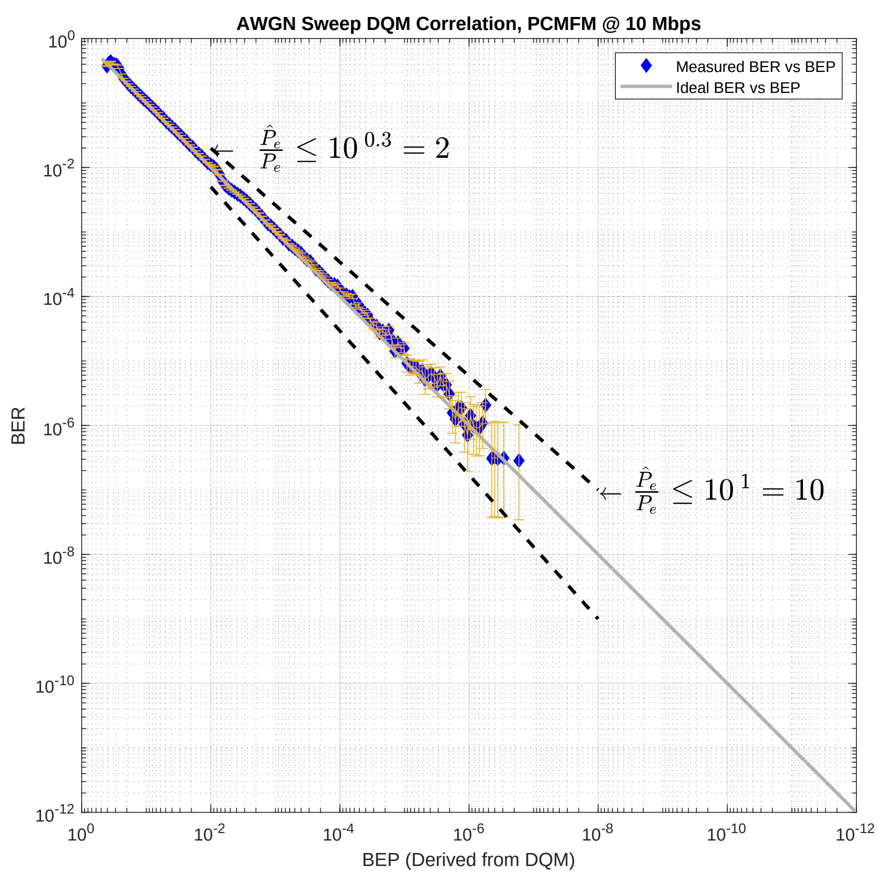


Figure 7: DQM Correlation Plot with tapered threshold

ACKNOWLEDGEMENTS

The author would like to thank Jim Uetrecht for providing his expertise and measured DQM data.

REFERENCES

[1] RCC, "Standards for data quality metrics and data quality encapsulation," in *Document IRIG 106-22 Chapter 2.*, Appendix 2-G, 2022.

[2] M. Rice and E. Perrins, "Maximum likelihood detection from multiple bit sources," in *Proc. Int. Telemetering Conf.*, (Las Vegas, NV), Oct. 2015.

[3] T. Hill, "Metrics and test procedures for data quality estimation in the aeronautical telemetry channel," in *Proc. Int. Telemetering Conf.*, (Las Vegas, NV), Oct. 2015.

[4] K. Temple, "Some thoughts on testing the data quality metric," in *Proc. Int. Telemetering Conf.*, (Las Vegas, NV), Oct. 2022.

[5] RCC, "Test procedures for assessing telemetry receiver data quality metrics," in *Document IRIG 118-22 Release 2 Volume 2.*, Chapter 11, 2022.

APPENDIX

In [2], the performance of the MLBD detector was analyzed and a procedure for calculating the probability of error was presented. This appendix extends the previous analysis to look at the effect of imperfect p estimation on the MLBD error probability. Starting with the principal equations from the paper for probability of error and MLBD metric

$$P(E|0) = \text{Pr}\left\{\sum_{n\epsilon \text{N}_1} \log\left(\frac{1-p_n}{p_n}\right) > \sum_{n\epsilon \text{N}_0} \log\left(\frac{1-p_n}{p_n}\right)\right\}. \tag{A.1}$$

$$M[i] = \sum_{n\epsilon \text{N}_1[i]} \log\left(\frac{1-p_n}{p_n}\right) - \sum_{n\epsilon \text{N}_0[i]} \log\left(\frac{1-p_n}{p_n}\right). \tag{A.2}$$

Let $\mathcal{I}_+$ be the set of indexes i for which $M[i] > 0$, and let $\mathcal{I}_-$ be the set of indexes i for which $M[i] < 0$. It was shown that the MLBD error probability P_e is the sum of the individual row probabilities P_i in the table listing all possible outputs where a positive error metric M_i occurs as shown below

$$P_e = \sum_{i\epsilon \text{I}_+} P[i] \quad \text{where} \quad P[i] = \prod_{n\epsilon \text{N}_0[i]} (1-p_n) \prod_{n\epsilon \text{N}_1[i]} (p_n). \tag{A.3}$$

It is important to note that misestimation of p only effects the MLBD metrics M[i] and does not change the probability products P[i] formed from the actual p values. Let $\hat{p}_1$ be an estimate of the actual channel transition probability p_1. If the estimation error is large enough, it will change the log-ratio sums such that the complementary error pair with the smallest difference will change signs. This causes the probability term associated with the smallest positive MLBD metric to be replaced with the probability term from the largest of the negative MLBD metrics resulting in an increase of the overall MLBD error output probability.

Starting with the ideal case, assume the smallest positive MLBD metric is $M[i_{\text{min+}}]$ which occurs at error pattern index $i = i_{\text{min+}}$ with complementary index $i_{\text{max-}}$. Misestimation of p_1 will not degrade the MLBD performance until the estimate $\hat{p}_1$ reaches a value that changes the sign of $M[i_{\text{min+}}]$ given by

$$\log\left(\frac{1-\hat{p}_1}{\hat{p}_1}\right) = \log\left(\frac{1-p_1}{p_1}\right) + M[i_{\text{min+}}] \quad \Longrightarrow \quad \hat{p}_1 = \frac{p_1}{p_1 + (1-p_1)e^{M[i_{\text{min+}}]}}. \tag{A.4}$$

substituting for $M[i_{\text{min+}}]$ yields

$$\hat{p}_1 = \frac{p_1}{p_1 + (1-p_1)e^{\left(\sum_{n\epsilon \mathrm{N}_1[i_{\text{min+}}]} \log\left(\frac{1-p_n}{p_n}\right) - \sum_{n\epsilon \mathrm{N}_0[i_{\text{min+}}]} \log\left(\frac{1-p_n}{p_n}\right)\right)}} \tag{A.5}$$

$$= \frac{p_1}{p_1 + (1-p_1) \prod_{n\epsilon \mathrm{N}_1[i_{\text{min+}}]} \left(\frac{1-p_n}{p_n}\right) \prod_{n\epsilon \mathrm{N}_0[i_{\text{min+}}]} \left(\frac{p_n}{1-p_n}\right)}. \tag{A.6}$$

Using the definition for the individual error probability $P[i_{\text{min+}}]$ and the fact that $P[i_{\text{max-}}]$ is its complementary probability pair yields

$$\hat{p}_1 = \frac{p_1}{p_1 + (1-p_1)\left(\frac{P[i_{\text{min+}}]}{P[i_{\text{max-}}]}\right)}. \tag{A.7}$$

Turning to the MLBD output error probability P_e, it consists of the sum of the probability rows with a positive MLBD metric. Using $\hat{p}_1$ instead of the actual p_1 will cause $P[i_{\text{min+}}]$ to be removed from the total and be replaced with $P[i_{\text{max-}}]$.

$$\hat{P}_e = \sum_{i\epsilon \mathrm{I}_+} P[i] - P[i_{\text{min+}}] + P[i_{\text{max-}}] \quad = \quad P[i_{\text{max-}}] + \sum_{i\epsilon \mathrm{I}_+ | i \neq i_{\text{min+}}} P[i]. \tag{A.8}$$

In order to understand the effect of misestimating p has on the MLBD error output probability, the proportional changes will be compared. First, the ratio of estimated to actual transition probability for p_1 is

$$\frac{\hat{p}_1}{p_1} = \frac{1}{p_1 + (1-p_1)\left(\dfrac{P[i_{\text{max-}}]}{P[i_{\text{min+}}]}\right)} \quad \xrightarrow{p_1 \ll 0} \quad \left(\frac{P[i_{\text{min+}}]}{P[i_{\text{max-}}]}\right). \tag{A.9}$$

Similarly, the ratio of MLBD probabilities can be written as

$$\frac{\hat{P}_e}{P_e} = \frac{P[i_{\text{max-}}] + \sum_{i\epsilon \text{I}_+ | i \neq i_{\text{min+}}} P[i]}{\sum_{i\epsilon \text{I}_+} P[i]} = \frac{P[i_{\text{max-}}] + \sum_{i\epsilon \text{I}_+ | i \neq i_{\text{min+}}} P[i]}{P[i_{\text{min+}}] + \sum_{i\epsilon \text{I}_+ | i \neq i_{\text{min+}}} P[i]} \tag{A.10}$$

Note that $P[i_{\text{max-}}]$ and $P[i_{\text{min+}}]$ are greater than or equal to any of the terms in the probability summation. Therefore, the ratio of $\hat{P}_e/P_e$ tends towards

$$\frac{\hat{P}_e}{P_e} \quad \xrightarrow{P[i_{\text{min+}}] \gg \sum_{i\epsilon \text{I}_+ | i \neq i_{\text{min+}}} P[i]} \quad \left(\frac{P[i_{\text{max-}}]}{P[i_{\text{min+}}]}\right). \tag{A.11}$$

These results show that the two probability ratios tend toward reciprocal values. This is a surprising result considering that small improvements in the transition probability can produce large performance improvements in MLBD error probability when using multiple channels. Define the ratio exponents as follows

$$\Delta p_{1_{\text{exp}}} = log_{10}\left(\frac{\hat{p}_1}{p_1}\right) \quad , \quad \Delta P_{e_{\text{exp}}} = log_{10}\left(\frac{\hat{P}_e}{P_e}\right). \tag{A.12}$$

Finally, it can be shown that since the probability terms are positive and the ratio $\hat{P}_e/P_e \geq 1$, the change in the transition probability exponent is greater than or equal to the change in the error probability exponent of the MLBD. Reordering the equation and applying an absolute value provides a convenient means to establish limits on DQM error accuracy.

$$\boxed{\Delta P_{e_{\text{exp}}} \;\leq\; |\Delta p_{1_{\text{exp}}}|} \tag{A.13}$$