Aloha collision detector

This one-page report[1] describes a collision detection classifier. It delivers satisfactory accuracy of collision detection, the expectation of AUC = 0.97 for one versus two transmitters. It is portable to an RFID reader. ***Run the code***.

Probability of a collision in a slot. The probability zero, one, two, and three or more transmitters sharing a given slot for N transmitters and 2^Q allocated time slots is

$$P_0 = \left(\frac{2^Q-1}{2^Q}\right)^N, P_1 = N \times \frac{1}{2^Q} \times \left(\frac{2^Q-1}{2^Q}\right)^{N-1}, P_2 = \frac{N(N-1)}{2} \times \frac{1}{2^{2Q}} \times \left(\frac{2^Q-1}{2^Q}\right)^{N-2},$$

and $P_{\geq 3} = 1 - P_0 - P_1 - P_2$.

The classifier model description. The proposed collision detection classifier model is the superposition of logistic regression, radial basis functions with Gaussian kernels, and self-modeling regression. The model acts in the complex space.

The dataset creation. Four classes were created: no-signal class with noise, a single transmission, two collided transmitters, and $3, \ldots, 5$ collided transmitters. The sample size and the signal mixture scaling keep characteristics of the orginal data.

The classifier accuracy analysis. The logistic regression returns the probability of collision. The quality criterion is Area Under ROC. Its expected value is AUC = 0.97 for the cross-validation of dataset, statistically isomorphic to original data. It means a reasonable probability of Aloha collision detection for one versus two or ≥ 3 transmitters.

Maximizing the accuracy on prior knowledge of N, Q. The collision classifier model was tested on datasets with various sample sizes and class balance ratios. The test shows that the accuracy is stable around borderline values of the sample size and class balance ratio. The AUC varies in $0.90, \ldots, 0.97$ (for balanced $0.94, \ldots, 0.98$) on the *test data*.

Model portability analysis. The preliminary number of operations: projection `complex`: $K \times T \times 2S$ multiplications, distance `complex`: $K \times T$ multiplications. The required memory `complex`: 128×39. For $K = 128$, $T = 39$, and $S = 1$ the model could be ported to the RFID reader after feasible simplificaton.

Future considerations. The problem of signal separation of two transmitters is important. Accorting to the obtained results it seems to be feasible even for a single reader.

Reproducibility of the experiment. The Google Colab Python notebook is accessible by ***the link in Appendix*** with detailed instructions how to run the experiment.

[1]Vadim Strizhov, `vadim.vct@gmail.com`, 7/2/2025

Appendix

1 Probability of a collision in a slot

There allocated 2^Q time slots for the set of N RFID tags, or transmitters. For a single transmitter the probability to occupy the empty slot is $\left(\frac{1}{D}\right)$. For simplicity denote $Q = \log_2 D$.

Probability of a slot having no transmitters. Each transmitter's time slot is independently and randomly chosen from D slots. The probability that a specific slot is not occupied by a *specific* transmitter is $\frac{D-1}{D}$. Since there are N transmitters, the probability that a specific slot is not occupied by anyone is

$$P_0 = \left(\frac{D-1}{D}\right)^N.$$

Probability of a slot having exactly two transmitters. To determine the probability that exactly two transmitters occupy a given slot, estimate the probability that a specific transmitter occupies this slot, $\frac{1}{D}$. The probability that exactly two specific transmitters occupy that slot is $\left(\frac{1}{D}\right)^2$, while the remaining $N-2$ transmitters *do not* occupy that slot, which occurs with probability $\left(\frac{D-1}{D}\right)^{N-2}$. Since any two of the N transmitters can be the ones sharing the slot, we choose two transmitters from N, which gives $\binom{N}{2} = \frac{N(N-1)}{2}$. Thus, the probability of exactly two transmissions in a given slot is:

$$P_2 = \binom{N}{2}\left(\frac{1}{D}\right)^2\left(\frac{D-1}{D}\right)^{N-2} = \frac{N(N-1)}{2} \times \frac{1}{D^2} \times \left(\frac{D-1}{D}\right)^{N-2}$$

Probability of a slot having three or more transmitters. The probability of exactly three transmitters occupying a given slot (while the others don't) is

$$P_3 = \binom{N}{3}\left(\frac{1}{D}\right)^3\left(\frac{D-1}{D}\right)^{N-3} = \frac{N(N-1)(N-2)}{6} \times \frac{1}{D^3} \times \left(\frac{D-1}{D}\right)^{N-3}.$$

The probability of four, five, or more transmitters occupying the same slot follows similarly. Thus, the probability of three or more transmitters sharing a given slot is $P_{\geq 3} = 1 - P_0 - P_1 - P_2$, where the probability of exactly one transmitter occurring is

$$P_1 = N \times \frac{1}{D} \times \left(\frac{D-1}{D}\right)^{N-1}$$

And since no transmission, exactly one transmission, or at least two transmission must occur on any given slot, we get:

$$P_{\geq 3} = 1 - P_0 - P_1 - P_2$$

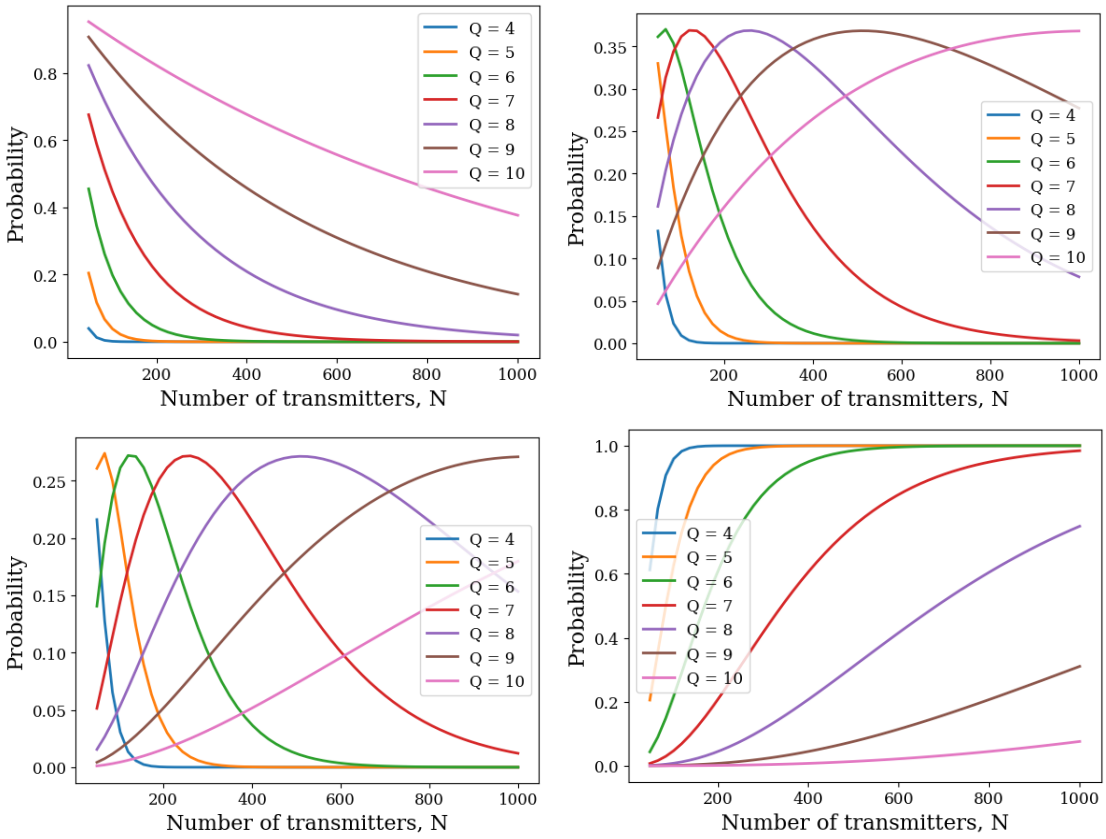

Figure 1: Probalility of no transmitter occupies a slot depends on the number of transmitters N and the number of slots D. The top left is P_0 (no transmission in a slot), right is P_1 (one transmission), the bottom left is P_2 (two transmitters collide), and right is P_3 (three and more transmitters collide).

This formula provides the probability of at least three transmitters sharing a specific slot.

The plots in Figure 1 show the fact that is if with an insufficiently small number of slots, there is no initial period where the probability of getting two transmitters in one slot increases. That is, if there are enough transmitters to overlap at all, they will *immediately start crowding into multiple transmissions per slot.* This fact brings importance to detect ≥ 3 collisions. We follow up this consideration in Section 7.

In terms of the variables N and Q the answer will be

$$P_0 = \left(\frac{2^Q - 1}{2^Q}\right)^N, P_1 = N \times \frac{1}{2^Q} \times \left(\frac{2^Q - 1}{2^Q}\right)^{N-1}, P_2 = \frac{N(N-1)}{2} \times \frac{1}{2^{2Q}} \times \left(\frac{2^Q - 1}{2^Q}\right)^{N-2},$$

and $P_{\geq 3}$ as above. This collision problem is known as the probabilistic birthday paradox [1, 2, 3, 4].

2 The classifier model description

The proposed collision detection classifier model is the superposition of logistic regression, radial basis functions with Gaussian kernels, and self-modeling regression. The first part returns the probability of collision as

$$p(y|\mathbf{w}) = \big(1 + \exp(-\mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}))\big)^{-1}.$$

The second part is the vector function $\boldsymbol{\varphi}(\mathbf{x})$ with Gaussian kernels

$$\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_K]^\mathsf{T}, \quad \text{where} \quad \varphi_k(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{d}(\mathbf{x}, \mathbf{c}_k) - \mathbf{c}_k\|^2}{2\sigma_k^2}\right).$$

And the last part is the self-modeling regression. It approximates the signal $\mathbf{x}$ with the centroid $\mathbf{c}$ as

$$\hat{\mathbf{c}} = v_1\big(\text{phase shift}(\mathbf{x}, v_2)\big),$$

with two parameters v_1, v_2. The first parameter v_1 is calculated as the dot product ratio of the projection

$$\text{proj}_\mathbf{c}\mathbf{x} = \frac{\mathbf{c}^\mathsf{T}\mathbf{x}}{\mathbf{c}^\mathsf{T}\mathbf{c}}\mathbf{c}, \quad \text{as long as} \quad \mathbf{x} \neq \mathbf{0}.$$

Note that this ratio could be negative, which is an admissible operation for the I/Q data signal. The second parameter calculated as an argument of minimum distance

$$v_2 = \underset{\text{admissible shift}}{\arg\min} \|\hat{\mathbf{c}} - \mathbf{c}\|^2.$$

For the simplicity of the computations and signal processing procedures, most part of the model acts in the complex space. The variables $\mathbf{x}, \mathbf{d}, \mathbf{c} \in \mathbb{C}^T$, where T is the length of

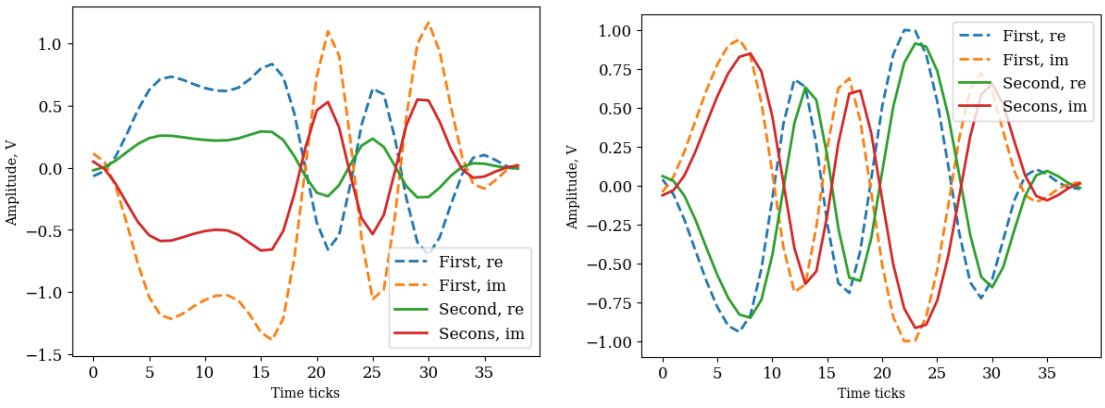

Figure 2: The self-modeling regression regresses a signal to centroid as a projection (left), while shifting the phase of the whole I/Q data signal to find the best fit (right). The legend shows real (re) and imaginary (im) part of the complex signal. The dashed line (First) shows the centroid, the solid line (Second) sows the approximated signal.

the transmitted I/Q data signal, $\boldsymbol{\varphi}, \mathbf{w} \in \mathbb{R}^K$, where K is the number of centroids, and the class $y \in \{0, 1\}$. The in-phase part of the complex vector $\mathbf{x}$ is real, while the quadrature part is imaginary.

Figure 2 illustrates the self-modeling regression. The first, dashed, signal is the centroid, scaled to 1 V, and the second, solid, signal is modified to approximate the centroid.

3 The dataset creation

The signal mixture procedure is organized as follows. The initial data seems to be scaled already and the scale does not always follow the amplitude in the $0.3, \ldots, 1.1$ V segment. However, the signal at the given scaling seems decent. The same with the noise. Its level is assumed to be properly set, so we leave it unchanged. The signal of small energy will express a bigger influence of the noise at the reader, so unchanging the given settings seems to be a natural decision.

Adding the signal increases their power for the same-phase case. To make the problem a little bit harder (signal attenuates with distance) we set the summing coefficients less than one: $0.5, \ldots, 0.8$ fixing them for the experiment. The scaling

$$A_{\text{result}} = \left| \sum_{i=1}^{N} A_i \exp(j\phi_i) \right|,$$

where A are the amplitudes of the signals and ϕ is the phase difference between them [5] (we find this with Gilbert transform, if the model complexity allows) requires additional discussion.

The generated data sample size. Decoupling indexes of data and noise datasets and assuming that the mixture of Gaussian noise is the Gaussian noise with zero expectation could provide us with a big sample size. We set the size of the original data for each of the four classes. In the case of ≥ 3 classes the mixture of 3 to five signals was generated.

4 The classifier accuracy analysis

Since the logistic regression model returns the probability of collision, the Area Under ROC is assigned as the quality criterion. For various cross-validations, its expected value is 9.7. For various normally-balanced datasets, it varies in the segment $0.94, \ldots, 0.98$. This means there is a reasonable probability of Aloha collision detection for one versus two or more transmitters.

Since the Aloha collision detection is the main goal of this experiment, and since the noise detection is the essential part of signal analysis, we suggest to analyze the classifier accuracy for the following combinations: 1) noise versus all, 2) single transmitter versus two transmitters, 3) single transmitter versus two or three or more transmitters, and 4) two transmitters versus three or more transmitters.

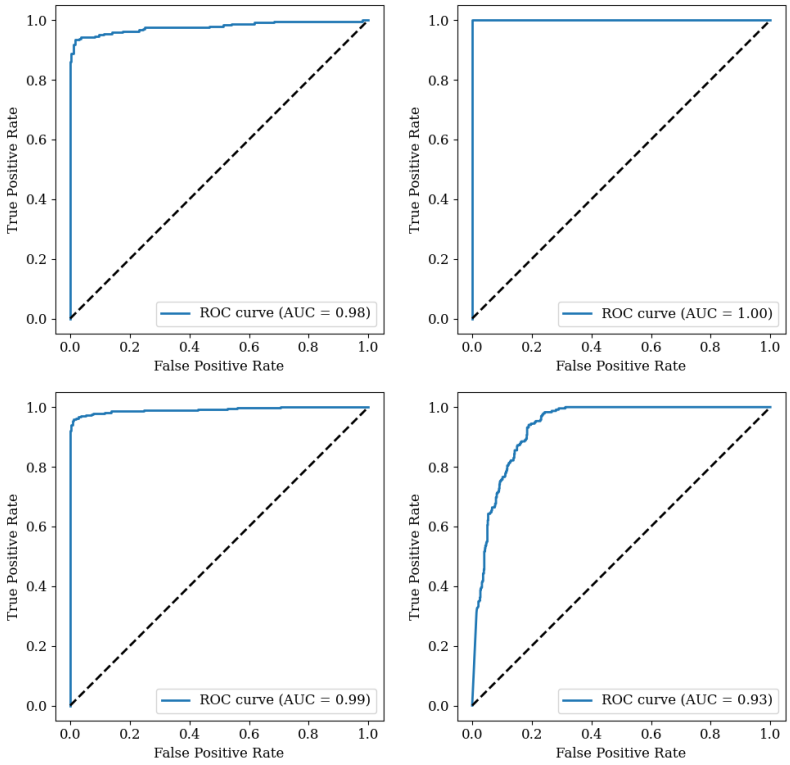

Figure 3: The model returns the probability of collision detection. The higher Operator Reciever Curve shows the better quality of detection (one transmitter versus two) on the *test part* of the dataset. The top row is 1 versus 2 (left), noise versus all (right). The bottom row is 1 versus 2 or ≥ 3 (left), and 0, 1 or 2 versus ≥ 3 transmitters (right).

The problem of 2 versus ≥ 3 transmitters *with complete separation and reconstruction* of both collided signals is feasible for the given data. Figure 3 shows the classification accuracy AUC $= 0.67$ without the reconstruction part; the latter is discussed in section 7.

Interesting case for the classes 1 versus $0, 2,$ or ≥ 3 when the RBF with logistic regression delivers AUC $= 0.66$ with two-hunch ROC curve, while kNN delivers AUC $= 0.96$ and accuracy 0.97. This case requires modification for the noise class, though the noise classification is straightforward with AUC $= 1.0$ on the test noise versus all classes.

Figure 3 shows the accuracy for the classification of a single transmitter versus the collision of two transmitters. The overall classification procedure is organized in the following way. The one-versus-all four classes classification problem is listed according to their probability: empty versus any transmission, one versus two transmitters, one versus two or more transmitters. The detection of three or more against two or more requires the blind signal separation procedure, which is introduced in Section 7.

5 Maximizing the accuracy on prior knowledge of N, Q

Assume as the prior knowledge 1) that the sufficient number of allocated for transmission time slots 2^Q brings a small number of collisions of two and a lesser number of collisions of ≥ 3 transmitters and, 2) the model parameters are subject to optimization during the reader's session. In this case, we suggest using predefined optimal parameters $\mathbf{w}_{\text{opt}}$, suitable for the current session or environment. For the model parameter optimization, apply a Bayesian penalty function for the current model. According to the Bayes' rule

$$p(\mathbf{w}|y, \mathbf{x}) \propto p(y|\mathbf{w}, \mathbf{x})p(\mathbf{w}_{\text{opt}}),$$

it changes the most likely parameters $p(y|\mathbf{w}, \mathbf{x})$ for the most probable parameters $p(\mathbf{w}|y, \mathbf{x}$. So the optimization criterion is

$$\mathcal{L} = \sum_{i=1}^{m} \Big(y_i \log p(\mathbf{x}_i, \mathbf{w}) + (1 - y_i) \log\big(1 - p(\mathbf{x}_i, \mathbf{w})\big) \Big) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{opt}})^{\mathsf{T}} \mathbf{A}^{-1} (\mathbf{w} - \mathbf{w}_{\text{opt}}).$$

The covariance matrix A is estimated in the multiple cross-validation procedure using prior knowledge datasets. The `LogisticRegression` model from `sklearn` has the regularization component (the right part of this formula) with $\mathbf{w}_{\text{opt}} = \mathbf{0}$ and $\mathbf{A} = \alpha\mathbf{I}$ by default.

Test the suggested collision classifier model on datasets with various sample sizes and class balance ratios. The approximate parameters for N and Q simulation were taken from the paper [6]. So let the number of single transmissions varies from 100 to 1000, and the ratio of collided transmission varies from 0.1 to 0.9.

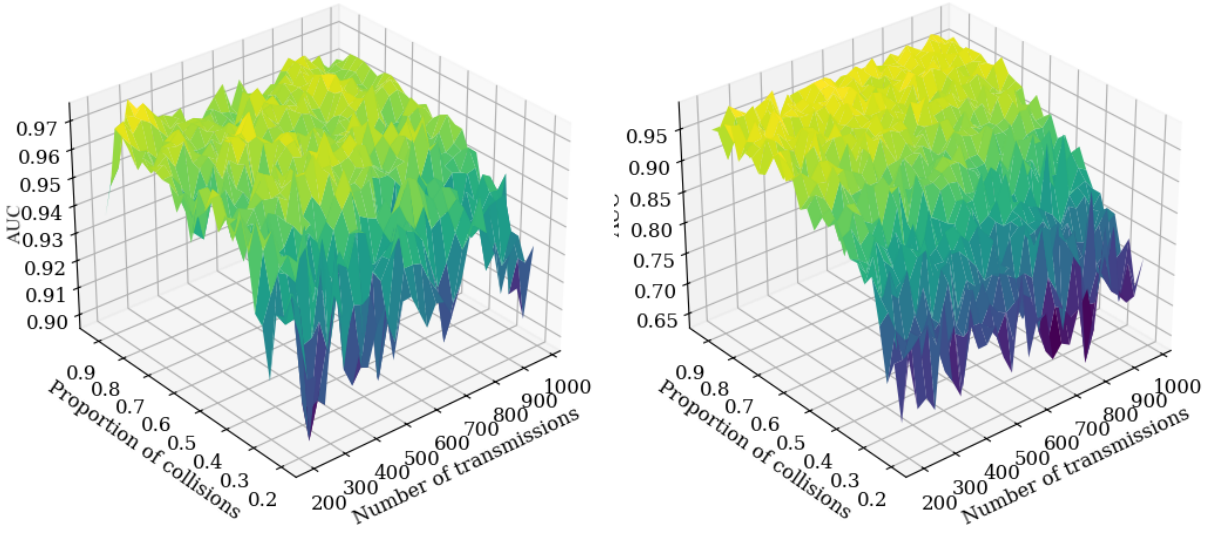

Figure 4: The accuracy of collision detection depends on the dataset sample size and class imbalance. The z-axis shows AUC. The left part is 1 versus two transmitters with logistic regression model, the right part is 1 versus 2 and ≥ 3 transmitters with kNN model.

Figure 4 illustrates *the class balancing* problem. It shows that the accuracy is stable around borderline values of the sample size and class balance ratio. It varies in the segment $0.90, \ldots, 0.97$ for training data. Each point of the surface is an independent train-test cycle for the regression model. The right axis shows the sample size for the single-transmission class. The left axis shows the proportion of this value in the collision class. So the summary of these two values makes the size of the dataset.

Since to optimize the model parameters we use the Newton-Raphson algorithm [7] that has small computational complexity and has good convergence to the optimal parameters, it could be ported to the reader. In this case, we suggest using the optimization criterion L to improve the imbalance of classes.

However, this suggested model has equal centroids according to the assumption that different transmitters has equal probability to act on the reader's request. So we suggest to fix the parameters of the model to their equal values. In this case, there is no need to optimize the model parameters in the reader.

6 Model portability analysis

To classify one transmitted signal the model uses the following operations:

1) projection `complex`: $K \times T \times 2S$ multiplications,

2) shifting `complex`: $K \times T \times 2S$ additions,

3) distance `complex`: $K \times T$ additions and multiplications,

4) classification `float`: T multiplications,

5) not counting single operations and padding,

6) there is `float`: $K + 1$ exponent operations.

Assuming the number of centroids $K = 128$, the length of the I/Q data signal $T = 39$, and the admissible phase shift parameter is $\pm$ time ticks, $S = 10$. With a single projection procedure $S = 1$, it takes $128 \times 39 \times 2$ operations. In this case the expected AUC = 9.3. However, without significant loss of accuracy, we could downsample I/Q data to the required number of operations. Providing the Gaussian noise expectation is zero, no denoising and additional operations are required. Also, since the logistic regression model parameters are assumed to be equal, we can classify the values of the kernel function by threshold. The model requires memory `complex`: 128×39. These estimations are preliminary.

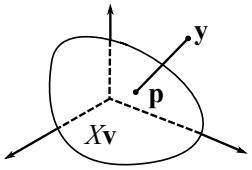

Figure 5: Two and more signals, mixture proportionally their attenuation, defines the vector span in the space of I/Q data signals. Here the vector $\mathbf{v}$ is the weights of the linear combinations of the signals. The vector $\mathbf{p}$ is the orthogonal projection to the span $\mathbf{Xv}$. The vector $\mathbf{y}$ is the mixture of signals and the added noise to be reconstructed. The basis of P independent (the transmitters can not send the same data) I/Q data signals $\mathbf{x}_1, \ldots, \mathbf{x}_P$ form the matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_P]$ as its columns.

7 Future considerations

As Figure 1 shows, the problem of detection of ≥ 3 transmitters becomes very important with increasing the number of transmitters. Once the probability of collision of two transmitters is significant, the further increasing number of transmissions greatly increases the number of collisions in a particular time slot.

For the majority of models, the pair of two transmitters versus ≥ 3 transmitters will not give accuracy that could be used in practical applications. However, models that could perform a blind signal separation could be useful not only to detect this type of collision but also to decode the signal at the receiver for two transmitters. Since the mixture of two or more signals from a single reader is inseparable, the challenge is to find a way to separate two signals with an accuracy high enough to tell that in case of the error the model declares the signal from ≥ 3 transmitters.

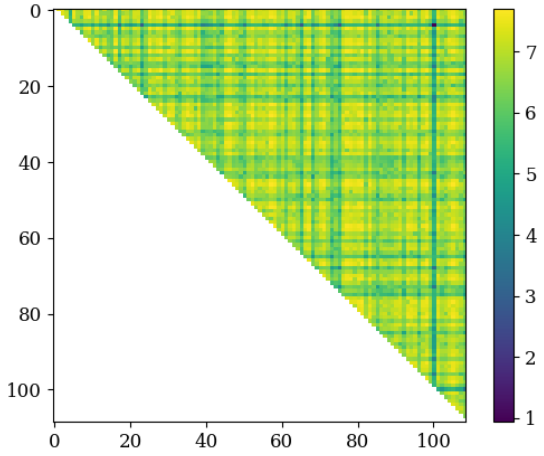

Figure 6: The rows and the columns denoted indexes of possible vectors. The dark blue dot in the upper right part of the matrix points to a couple of the I/Q data signals, whose linear combinations are the nearest, not counting the noise $\boldsymbol{\epsilon}$.

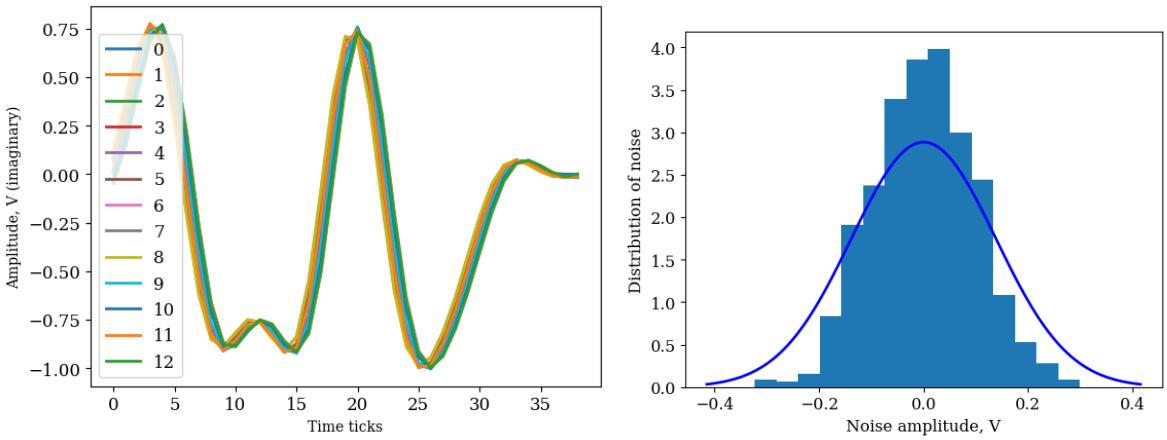

Figure 7: The left part of figure shows the cluster of transmitted signals with the same message without noise. They scaled to the same shift phase and amplitude. The ground truth noise (right part), samples from the dataset, has $\mu \leq 0.005$ (real or image) and standard deviation $\sigma = 0.1361$. The plot shows 3σ. This distribution function us used in the ≥ 3 signal reconstruction procedure.

Further, we plan to use the methods of Independent Component Analysis and Blind Signal Separation in the time domain and frequency domain [8, 9] of the I/Q data signal using a single reader (though this method requires at least two readers, and we do not count one reader has four antennas).

In the light of our recent advancement see the report on the signal separation and the papers [10, 11].

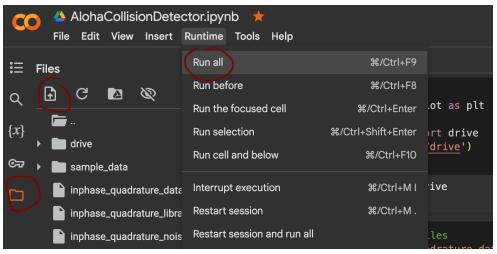

Figure 8: Upload the files to Google Colab Python notebook and run the computational experiment.

8 How to run the experiment

Open the link to the Google Colab file or copy the address to the browser (address is updated). Press the "Files" icon on the left panel of Colab (the fifth from the top, below the key). Then press the "Upload to session storage" icon right below the

word "Files". From your local disk upload the files inphase_quadrature_data.json, inphase_quadrature_noise.json, and inphase_quadrature_lib.npy. The last one is attached along with this text. In the Colab menu click "Runtime" and select the item "Run all". After the first cell runs, Colab asks the access to the uploaded files. Press the button "Continue" each time to let the Google Colab access your Google Disk, *there will be several consequent requests*. The experiment runs until the end. Figure 8 shows the orange "Files" icon and the "Runtime" menu open.

References

[1] Qiyang Sun. The probability principle of the birthday paradox and extended applications. *The Frontiers of Society, Science and Technology*, 3(3), 2021.

[2] Carla Santos and Cristina Dias. A probabilistic approach on coincidences: the birthday paradox. *Pensamiento Matematico*, 2015.

[3] Frederick Mosteller. Understanding the birthday problem. *The Mathematics Teacher*, 55(5):322–325, May 1962.

[4] Masoud Shakiba, Mandeep Jit Singh, Elankovan Sundararajan, Azam Zavvari, and Mohammad Tariqul Islam. Extending birthday paradox theory to estimate the number of tags in rfid systems. *PLoS ONE*, 9(4):e95425, April 2014.

[5] Constantine A. Balanis. *Antenna Theory*. Wiley-Interscience, 2005.

[6] Seungnam Kang and Zornitza Prodanoff. *RFID Model for Simulating Framed Slotted ALOHA Based Anti-Collision Protocol for Muti-Tag Identification*. InTech, July 2011.

[7] Anastasiya Motrenko, Vadim Strijov, and Gerhard-Wilhelm Weber. Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255:743–752, January 2014.

[8] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5):411–430, June 2000.

[9] Mark A. Elliott, Glenn A. Walter, Alex Swift, Krista Vandenborne, John C. Schotland, and John S. Leigh. Spectral quantitation by principal component analysis using complex singular value decomposition. *Magnetic Resonance in Medicine*, 41(3):450–455, March 1999.

[10] Alexandr Katrutsa and Vadim Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, 76:1–11, June 2017.

[11] A.M. Katrutsa and V.V. Strijov. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142:172–183, March 2015.