Form groups of 2–4 persons and work on a research-oriented project on information and coding theory. You need to culminate your work during the application period. In week 8 you need to make a presentation and provide us with final report as well as a source code.

In what follows we list possible project areas, initial project descriptions, bibliography in the area and our expectations. You need to express interest in up to 3 of these areas, or come up with your own suggestion, by March 4 and send them by e-mail to TAs. Groups will be formed after your suggestions are received.

1 Random Coding Bound for Massive Random-Access

Existing wireless networks are designed with the goal of increasing a spectral efficiency in order to serve human users. Next generation of wireless networks will face a new challenge in the form of machine-type communication. Analysts predict that the number of devices connected to the network will exceed 50 millions by 2020. The main challenges are as follows: (a) huge number (billions) of autonomous devices connected to one access point, (b) low energy consumption, (c) short data packets. This problem has attracted attention of 3GPP standardization committee under the name of mMTC (massive machine-type communication). In this project you will investigate possible approach to use codes with same codebook for each user in this scenario and derive a random coding bound in case of Additive White Gaussian Noise Multiple Access Channel (AWGN-MAC).

Our expectations:

1. Implement a tool to plot the random coding bound for AWGN-MAC (see reference)
2. Derive a bound for binary input AWGN-MAC (BI-AWGN-MAC). In this case the users utilize the alphabeth $\{-1, +1\}$.
3. Compare the bounds for the parameters: $K_a = 4$, $k = 100$ (inf. bits), $n = 1000$ (channel uses). Plot 2 curves of a user error rate P_e in dependence of E_b/N_0.

References and Resources:

- Y. Polyanskiy, A Perspective on Massive Random-Access, in Proc. IEEE International Symposium on Information Theory (ISIT), pp. 2523–2527, 2017.

2 Short codes and their applications to future 5G wireless networks and Internet of Things.

Most of the recent advances in the design of high throughput wireless systems are based on information-theoretic principles that demonstrate how to efficiently transmit long data packets. Until recently, the transmission of short data packets has not been one of the major tasks in the development of wireless networks. However, the upcoming wireless systems, notably the 5G system, will need to support novel traffic types that use short packets (i.e., codes with dimension k in the range of 50 to 1000 bits). In this project you will compare existing coding schemes in IoT scenarios.

Our expectations:

1. Compare the performance of binary LDPC, non-binary LDPC, Polar and Turbo codes for the following parameters: BI-AWGN channel, BPSK, QPSK and 8PSK modulations, $k = 64$, $R \in \{0.25, 0.5\}$.
2. Plot the results as curves of an error rate P_e in dependence of E_b/N_0.

References and Resources:

- G. Liva, L. Gaudio, T. Ninacs, T. Jerkovits, Code Design for Short Blocks: A Survey, arXiv:1610.00873.

3 Sum-Product Algorithm for LDPC Codes

Sum-Product algorithm (SPA) is a message passing algorithm, which works on the Tanner graph, corresponding to LDPC code. It is known, that SPA is equivalent to maximum a posteriori probability (MAP) decoder if a Tanner graph is a tree. Unfortunately, tree codes have minimum distance equal to 2 and bad error correcting capabilities. So SPA is usually applied for loopy Tanner graphs.

Our expectations:

1. Implement an SPA
2. TAs will provide you with LDPC parity-check matrix and encoding function (written in MATLAB)
3. Choose a digit from MNIST database
   http://yann.lecun.com/exdb/mnist/
4. Encode the digit, modulate ($0 \to +1$, $1 \to -1$), add Gaussian noise (choose an SNR value yourself, such that the decoding is possible)
5. Use the decoder to recover the original digit. Show the "evolution" of the decoding process for iterations $1 \dots 15$.

References and Resources:

- F. R. Kschischang, B. J. Frey and H. A. Loeliger, "Factor graphs and the sum-product algorithm," in IEEE Trans. Inf. Theory, vol. 47, no. 2, pp. 498-519, Feb 2001.

4 Polar Codes for Two-User BI-AWGN-MAC

Polar codes are a new and interesting class of codes proposed by E. Arikan in 2009. The usual way to decode polar codes is to apply successive cancellation algorithm. In this project you will implement a successive cancellation decoder of polar codes for the two-user BI-AWGN-MAC.

Our expectations:

1. Implement an encoding function
2. Implement a joint successive cancellation decoder
3. Choose two digits from MNIST database
   http://yann.lecun.com/exdb/mnist/
4. Encode the digits, modulate ($0 \to +1$, $1 \to -1$), sum the resulting vectors, add Gaussian noise (choose an SNR value yourself, such that the decoding is possible)

Cont.

5. Use the decoder to recover the original digits.

References and Resources:

- S. Onay, Successive cancellation decoding of polar codes for the two-user binary-input MAC, in Proc. IEEE International Symposium on Information Theory, Istanbul, 2013, pp. 1122–1126.

5 Private Information Retrieval

Private Information Retrieval (PIR) is a two-party cryptographic protocol that allows a user to retrieve an item from a database without revealing any information about the retrieved item to the database (however, with no guarantee for the confidentiality of the database). In 2005 Gentry and Ramzan propose an efficient single database Private Block Retrieval (PBR), an extension of PIR where a user can retrieve a block of data of size d bits, using smooth subgroups (i.e., subgroups that have many small primes dividing their order). In this project you will implement it.

Our expectations:

1. Implement PBR protocol
2. Test its' performance for different database and stored data size.

References and Resources:

- C. Gentry, Z. Ramzan, Single-database private information retrieval with constant communication rate, in Automata, Languages and Programming. Volume 3580. Springer Berlin Heidelberg, pp. 803-815, 2005.

6 A Minimax Approach to Supervised Learning

Given a task of predicting Y from X, a loss function L, and a set of probability distributions Γ on (X, Y), what is the optimal decision rule minimizing the worst-case expected loss over Γ? In this project you will investigate a novel approach proposed by David Tse (Shannon Award, 2017) which is in fact generalisation of the maximum entropy principle and apply it to binary classification problem.

Our expectations:

1. Implement minimax SVM
2. Compare its' performance with Support Vector Machines (SVM) , Tree Augmented Naive Bayes (TAN), and Discrete Renyi Classifiers (DRC) on datasets from the UCI repository

References and Resources:

- F. Farnia, D. Tse, A Minimax Approach to Supervised Learning, Neural Information Processing Systems, pp. 4240–4248, 2016

Cont.

7 Information Bottleneck Principle

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Tishby and Shwartz-Ziv suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer. They came to this conclusion after the analysis of a toy NN. In this project you will reproduce and verify these results for a "real-life" NN.

Our expectations:

1. Choose a problem, such that it can be solved with NN, e.g. recognition of digits from MNIST database http://yann.lecun.com/exdb/mnist/
2. Look at mutual information evolution during the training.
3. Present the results in the same way as in the reference. Do they look similar?

References and Resources:

- R. Shwartz-Ziv, N. Tishby, Opening the Black Box of Deep Neural Networks via Information, arXiv:1703.00810

8 Deep Learning for Decoding of Linear Codes – A Syndrome-Based Approach

The main goal is to suggest an application of deep learning to channel decoding. There were already attempts to construct neural network (NN) learning-based decoders in literature, here we face with so-called curse of dimensionality problem: even for a short code of length $N = 100$ bits and rate $R = 0.5$, 2^{50} different codewords exist, which are far too many to fully train any NN in practice. In this project you will reproduce a syndrome based approach to train NN. The main advantage is that we can perform training on zero codeword only (belongs to any linear code).

Our expectations:

1. Implement the algorithm from the reference
2. Perform training and validation for 5G LDPC code (the matrix will be provided by TA)
3. Compare the performance with Sum-Product decoder (the results will be provided by TAs).

References and Resources:

- A. Bennatan, Y. Choukroun, P. Kisilev, Deep Learning for Decoding of Linear Codes – A Syndrome-Based Approach, arXiv:1703.00810

9 Cooperative learning of encoding/decoding functions

The idea is to present encoder and decoder as NNs and perform joint training. As a result the system will learn the code and encoding/decoding functions.

Our expectations:

1. Implement the algorithm, train NNs
2. Simulate the performance for the following parameters: BI-AWGN channel, BPSK modulation, $k = 64$, $R = 0.25$

Cont.

3. Plot the results as curves of an error rate P_e in dependence of E_b/N_0
4. Compare the performance with known short-length codes (the reference will be provided by TAs).

10 Information Theory of DNA Shotgun Sequencing

DNA sequencing is the basic workhorse of modern day biology and medicine. Shotgun sequencing is the dominant technique used: many randomly located short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. A basic question is: given a sequencing technology and the statistics of the DNA sequence, what is the minimum number of reads required for reliable reconstruction? In this project you will investigate a novel approach proposed by David Tse (Shannon Award, 2017).

Our expectations:

1. Implement a tool to calculate a fundamental limit to the performance of any assembly algorithm
2. Implement an optimal assembly algorithm for shotgun sequencing under the criterion of complete reconstruction.

References and Resources:

- A. S. Motahari, G. Bresler and D. N. C. Tse, Information Theory of DNA Shotgun Sequencing, in IEEE Transactions on Information Theory, vol. 59, no. 10, pp. 6273-6289, Oct. 2013.
- G. Bresler, M. Bresler and D. N. C. Tse,Optimal Assembly for High Throughput Shotgun Sequencing, arXiv:1301.0068.

The End.