

Active Deep Decoding of Linear Codes

Ishay Be'ery, Nir Raviv, Tomer Raviv, and Yair Be'ery, *Senior Member, IEEE*

Abstract—High quality data is essential in deep learning to train a robust model. While in other fields data is sparse and costly to collect, in error decoding it is free to query and label thus allowing potential data exploitation. Utilizing this fact and inspired by active learning, two novel methods are introduced to improve Weighted Belief Propagation (WBP) decoding. These methods incorporate machine-learning concepts with error decoding measures. For BCH(63,36), (63,45) and (127,64) codes, with cycle-reduced parity-check matrices, improvement of up to 0.4dB at the waterfall region, and of up to 1.5dB at the error-floor region in FER, over the original WBP, is demonstrated by smartly sampling the data, without increasing inference (decoding) complexity. The proposed methods constitutes an example guidelines for model enhancement by incorporation of domain knowledge from error-correcting field into a deep learning model. These guidelines can be adapted to any other deep learning based communication block.

Index Terms—Deep Learning, Error Correcting Codes, Machine Learning, Active Learning, Belief Propagation

I. INTRODUCTION

Decoding of error-correcting codes has changed over the last few years. The rise of machine-learning methods, primarily of the deep learning subset, changed the field significantly and comprehensively.

Nachmani et al. [1], [2], proposed a model-based approach, placing learnable weights on Tanner graph's edges of the Belief Propagation (BP) algorithm [3] for linear codes. This approach is acknowledged as the original WBP, since had been the first in the field to learn a parameterized BP algorithm employing Stochastic Gradient Descent (SGD). The intuition offered was that the weights compensated for the short cycles in the Tanner graph. This addition improved the decoding performance. Lian et al. [4] validated these results and further explored spatial and temporal weights sharing. Xu et al. [5], [6] generalized the method for both Tanner and factor graphs of polar codes. Considering model-free approaches, Gruber et al. [7] proposed a fully connected (FC) neural networks (NN) approach, composed of linear and ReLU [8] layers. This model achieved Maximum a-posteriori (MAP) performance on very short polar codes. Bennatan et al. [9] presented a combination of model-based and model-free approaches in which a NN was trained by the syndrome of the received message. Utilizing concurrent NN designs in addition with learning the code properties, via the composed syndrome, achieved performance improvement. Further contributions to the field lie in [10] and [11], where neural decoders for

convolutional codes were proposed together with befitting training methodologies. However, these methodologies impose substantial increase of complexity, at both training and decoding (inference). Specifically, [10] explores a recurrent neural networks (RNN) architecture for decoding while [11] focuses on an unconstrained novel structure requiring no knowledge of the BCJR (Bahl-Cocke-Jelinek-Raviv) algorithm.

As evident from the above recap, many researches paid great attention to the decoder's architecture, revealing a typical trend in the field. Yet another aspect of the mentioned decoding problem is the training data. In [1], [2], training over varying SNR (signal to noise ratio) ranges was explored. This led to different decoding performances over the same validation set. Regarding choice of a single optimal training point, Kim et al. [10] provides guidelines for choosing the best training SNR value. Gruber et al. [7] showed that the choice of a training SNR value for generalization purposes is essential. A grid search is applied to locate the optimal single training SNR. This empirical result was followed by an analytical study in [12]. In the study, an entropy based analysis was performed, deriving a bound on the increase of the maximal error probability due to mismatched training and validation sets. The main conclusion is that no optimal training SNR for all validation sets *exists*, but rather *depends* on the specific validation data. One realization of this result is presented in [13], where the WBP parameters are assumed to be SNR dependent. Multiple NN are used to infer the value for each parameter in the WBP algorithm at the validation phase, conditioned on the SNR.

Data is a vital part of deep learning methods, yet we see that it is not fully comprehended. Many researchers focus on preliminary choice of training data, followed by passive generation of examples during training. We rather search for an adaptive scheme which actively samples the training data to feed the neural decoder. Regarding complexity, [13] emphasizes that distribution-specific data requires unique analysis, but the additional NN cause extra complexity. In this paper we narrowed the view for schemes with no additional modules.

Our main contributions are:

- 1) Active learning inspired approach is first applied, to our best knowledge, in the error-correcting codes field.
- 2) Performance improvement with no decoding complexity penalty.
- 3) Directing the effort of the machine-learning decoding community to data-tailored solutions.

We call our approach *active deep decoding*.

The paper is organized as follows. Section II covers notation and definitions. Section III explores different decoding parameters and sets the ground for the novel methods. Section IV introduces a detailed explanation of the methods. Section

This work was presented in part in the Future of Wireless Technology Workshop, Stockholm, Sweden, June 2019.

I. Be'ery, N. Raviv, T. Raviv and Y. Be'ery are with the School of Electrical Engineering, Tel-Aviv University, Tel-Aviv 6997801, Israel (e-mails: ishaye@gmail.com, nirravriv89@gmail.com, tomerraviv95@gmail.com, ybeery@eng.tau.ac.il).

V presents experiments and results and section VI concludes the paper.

II. PRELIMINARIES

A. Notation

We denote scalars in *italics* letters and vectors in **bold**. Capital and lowercase letters stand for a random vector and its realization, respectively. For example, \mathbf{C} and \mathbf{c} stand for the codeword random vector and its realization vector. \mathbf{X} and \mathbf{Y} are the transmitted and received channel words. $\hat{\mathbf{X}}$ denotes the decoded modulated-word, while $\hat{\mathbf{C}}$ denotes the decoded codeword. The i^{th} element of a vector \mathbf{v} will be denoted with a subscript v_i .

We will only deal with the AWGN channel in this work, denoting the SNR by ρ for convenience. We also denote the code by \mathbb{C} , with minimum Hamming distance d_{\min} and code length N . Let \mathbf{u} , \mathbf{x} and \mathbf{y} denote the message word, the transmitted word (after encoding and BPSK modulation) and the received word (with Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$) respectively, see Figure 1. Note that one always decodes the received LLR word \mathbf{z} , not \mathbf{y} . Let $\text{dist}(\mathbf{c}_1, \mathbf{c}_2)$ denote the Hamming distance between two codewords \mathbf{c}_1 and \mathbf{c}_2 . Specifically, we denote by d_H the Hamming distance between the encoded codeword \mathbf{c} and the decoded word $\hat{\mathbf{c}}$. The received word will always be decoded correctly by a hard-decision decoder if the Hamming distance between \mathbf{c} and \mathbf{y} demodulated by hard-decision is $t_H = \lfloor \frac{d_{\min}-1}{2} \rfloor$ at most. Let T be a latent binary variable [14], which denotes successful decoding of the NN decoder, with a value of 1 if $\mathbf{c} = \hat{\mathbf{c}}$ (similarly $d_H = 0$).

At last we denote $I(X; Y)$ as the mutual information between the two random variables, X and Y .

B. Training by Different Parameters

Let $\Gamma_\theta(S)$ be a distribution over received words \mathbf{Y} , parameterized by hyperparameters $\theta \in \Theta$ set with values S . For example, let θ be ρ and $S = 1\text{dB}$. Then, a training sample is drawn (assuming the all-zero codeword is transmitted) according to $P_{\mathbf{Y}}(\mathbf{y}; \rho = 1)$. For a batch of i.i.d. training samples, the entire sampling procedure is repeated n times, where n is the required batch size and both θ and S may vary in the same batch. We denote a batch sampled according to Γ by \mathbf{y}_γ .

C. Weighted BP Decoding

The Belief Propagation (BP) is an inference algorithm used to calculate the marginal probabilities of nodes in a graph efficiently. Pearl [3] also advocated the utilization of this algorithm for graphs with loops, along with a remark that it is an approximation only. This version is called the loopy belief propagation. A full derivation from the general case to linear codes can be found in [15]. We provide main details next.

The Tanner graph is an undirected graphical model, constructed of nodes and edges. The nodes are of two types - variables and checks nodes. A variable node corresponds to a single bit of the received codeword. Each check node corresponds to a row in the code's parity check matrix. An

edge exists between a variable v and a check node h iff variable v participates (has coefficient 1) in the condition defined by the h^{th} row in the parity check matrix.

The initialization of the variable nodes:

$$z_v = \log \frac{P(c_v = 0 | y_v)}{P(c_v = 1 | y_v)} = \frac{2y_v}{\sigma_n^2}$$

The subscript v indicates a variable node and z stands for a LLR (log-likelihood ratio) value. The last equality is true for AWGN channels with common BPSK mapping to $\{\pm 1\}$.

The message passing algorithm proceeds by iteratively passing messages over edges from variables nodes to check nodes and vice versa. The BP message from node a to node b at iteration i will be denoted by $m_{i,(a,b)}$ with the convention that $m_{0,(a,b)} = 0$ for all a, b combinations. Variable-to-check messages are updated in odd iterations according to the rule:

$$m_{i,(v,h)} = z_v + \sum_{(h',v), h' \neq h} m_{i-1,(h',v)}$$

While the check-to-variable messages are updated in even iterations by:

$$m_{i,(h,v)} = 2 \operatorname{arctanh} \left(\prod_{(v',h), v' \neq v} \tanh \left(\frac{m_{i-1,(v',h)}}{2} \right) \right)$$

Finally, the output variable node value is calculated by:

$$\hat{x}_v = z_v + \sum_{(h',v), h' \neq h} m_{2\tau,(h',v)}$$

Where τ is the number of BP iterations and all values considered are LLR values. In [1], [2], learnable weights are assigned to the variable-check message passing rule:

$$m_{i,(v,h)} = \tanh \left(\frac{1}{2} \left(w_{i,v} z_v + \sum_{\substack{(h',v) \\ h' \neq h}} w_{i,(h',v,h)} m_{i-1,(h',v)} \right) \right) \quad (1)$$

And to the output marginalization:

$$\hat{x}_v = \sigma \left(- \left[w_{2\tau+1,v} z_v + \sum_{\substack{(h',v) \\ h' \neq h}} w_{2\tau+1,(h',v)} m_{2\tau,(h',v)} \right] \right) \quad (2)$$

where σ is the sigmoid function. We denote by $\mathbf{w} = \{w_{i,v}, w_{i,(h',v,h)}, w_{i,(v,h')}\}$ the set of weights. Note that no weights are assigned to the check-variable rule, which now takes the form:

$$m_{i,(h,v)} = 2 \operatorname{arctanh} \left(\prod_{(v',h), v' \neq v} m_{i-1,(v',h)} \right) \quad (3)$$

This decision is explained by expected numerical instabilities due to the arctanh domain.

This formulation unfolds the loopy algorithm into a NN. One can see that the hyperbolic tangent function was moved from check-variable rule to scale the message to a reasonable output range. A sigmoid function is used to scale the LLR values into the range $[0, 1]$. An output value in the range

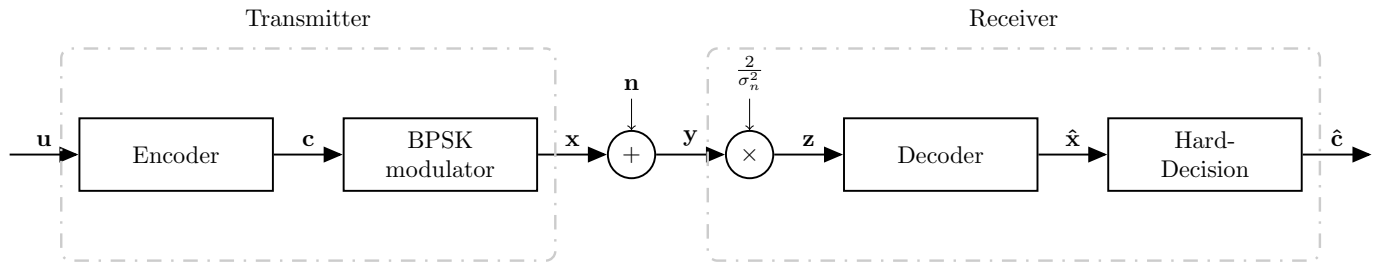


Fig. 1: System Diagram

(0.5, 1] is considered a '1' bit, otherwise a '0' bit (value of 0.5 was attributed to the '0' bit randomly). Training is done with the Binary Cross Entropy (BCE) multiloss:

$$L(\mathbf{c}, \hat{\mathbf{x}}) = -\frac{1}{N} \sum_{t=1}^{\tau} \sum_{v=1}^N [c_v \log \hat{x}_{v,t} + (1 - c_v) \log (1 - \hat{x}_{v,t})]$$

For a comprehensive explanation of the subject, please refer to [1], [2].

III. DATA EXPLORATION

We start exploring the data with a question in mind - *do all words contribute equally to the neural training?*

A. The SNR Parameter - A Motivation

We inspect how possessing the knowledge of ρ can affect the training data and model choices.

Regarding training data, Gruber et al. [7] trains multiple neural decoders, each decoder trained with data drawn from $\Gamma_{\rho}(i)$ where $-4 \leq i \leq 8, i \in \mathbb{Z}$. The $NVE(\rho_t, \rho_v)$ measure is suggested in [7] to compare between the trained models. One can notice that the model diverges when trained over only correct or noisy words, drawn from high or low SNR, respectively. In [10] guidelines for choosing ρ_t are provided. The value is chosen so that the neural decoder's training set is comprised from \mathbf{y} near the decision boundary.

Regarding model choices, a hidden assumption of [13] is that \mathbf{y}_{γ} which are drawn from $\Gamma_{\rho}(S_1)$ and $\Gamma_{\rho}(S_2)$ ($S_1 \neq S_2$) require different decoder weights, $\mathbf{w}_1, \mathbf{w}_2$. One may observe that knowledge possession of ρ_v is also mandatory for all LLR-based decoders (as an estimate is required to compute LLRs). It is quite straightforward to show that the next mutual information inequality holds:

$$I(\mathbf{Y}, \rho_v; T) \stackrel{(a)}{=} I(\mathbf{Y}; T) + I(\rho_v; T | \mathbf{Y}) \stackrel{(b)}{\geq} I(\mathbf{Y}; T)$$

where (a) follows from the mutual information chain rule, and (b) follows from the non-negativity of mutual information. Thus, the additional information of ρ_v can only aid decoding. This information of the channel and the decoder distributions, conditioned on the received word, may be non-zero for sub-optimal decoders. In [13], inference does not only require ρ_v knowledge, but is also ρ_v dependent. In other words - the model is data dependent.

B. Objective Formulation

Motivated by the above discussion, our main goal is to find parameters other than the SNR, which define a new Γ, Γ_{new} . We want that training the WBP over Γ_{new} will achieve as high decoding performance as possible.

Let κ denote the contribution of a word, in the training phase, to the validation decoding performance. We associate higher contribution words with higher κ value. Our goal is to find parameters $\theta \in \Theta$ and corresponding values S defining words distribution $\Gamma_{\theta}(S)$ such that the κ value integrated over the distribution is maximized:

$$\arg \max_{\theta, S} \int_{\mathbf{y} \in \Gamma_{\theta}(S)} \kappa(\mathbf{y})$$

The solution to this equation is intractable due to the infinite number of such parameters and values, thus we seek heuristic-based solution. We choose the parameters based on the vast decoding knowledge while using the above insights. In particular, \mathbf{y}_{γ} should be neither too noisy nor absolutely correct and should lie close to the decision boundary. Recall that throughout the paper we use the AWGN channel. Therefore, we search for parameters θ' which limit the feasible \mathbf{y}_{γ} of the channel distribution $\Gamma_{\rho}(S)$, associated with $K_{\rho}(S)$, to $\Gamma_{\rho, \theta'}(S, A)$, associated with higher $K_{\rho, \theta'}(S, A)$, where we denote $K_{\theta}(S) = \int_{\mathbf{y} \in \Gamma_{\theta}(S)} \kappa(\mathbf{y})$.

C. Distance Parameter

Some received words are undecodable due to the locality of the decoding algorithm, the Tanner graph structure induced by the parity-check matrix or a high Hamming distance. By sampling from specific $\Gamma_{\rho, d_H}(S, A)$ one can easily control the number of erroneous bits in \mathbf{y} . Choosing such words with a reasonable Hamming distance between them and the transmitted words decreases the amount of undecodable words in Γ .

To justify the above claims we trained a WBP decoder without any correct received words, $d_H=0$, and without high noise words, $d_H > t_H$. The training setup is similar to the one used in section V. The results show an improvement of up to 0.5dB by sampling according to this simple scheme, confirming our intuitions. By drawing data according to a distribution, and not according to the SNR, we have further control on training words properties. We elaborate more on this subject in IV-A.

With this short experiment we manage to answer the question we set to ask - do all words contribute equally to the training? A definitive answer is **no**.

D. Reliability Parameters

Soft in soft out (SISO) decoding compose the received signal to n LLR values, $\{z_1, \dots, z_n\}$. In general $z_v \in (-\infty, \infty)$ but in practice we limit their value by choosing appropriate threshold. The closer the z_v to 0, the less reliable it is. We consider mapping the LLR values to bits in two steps. First mapping LLR values to probabilities:

$$\Pi_{LLR \rightarrow Pr}(z_i) = \sigma(-z_i)$$

The next rule maps probability into corresponding bit:

$$\Pi_{Pr \rightarrow bit}(\tilde{z}_i) = \begin{cases} 1, & \text{if } \tilde{z}_i > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

The process of direct quantization from LLR to bits is called hard decision (HD) decoding:

$$\Pi_{HD}(z_i) = \Pi_{Pr \rightarrow bit}(\Pi_{LLR \rightarrow Pr}(z_i)) \quad (4)$$

Obviously there is information loss in the process:

$$\Pi_{HD}(z_1) = \Pi_{HD}(z_2) \not\Rightarrow z_1 = z_2$$

We seek numeric parameters which quantify reliability of a given \mathbf{z} . Two parameters that we inspected and found fitting to the task are defined below:

Average Bit Probability - the deviation of the channel output probabilities from the corresponding transmitted bits:

$$\eta_{ABP}(c_i, z_i) = \frac{1}{N} \sum_{i=1}^N |c_i - \Pi_{LLR \rightarrow Pr}(z_i)| \quad (5)$$

Mean Bit Cross Entropy - this parameter quantifies how close are the two probability distributions at the transmitter and at the receiver (before decoding):

$$\ell_{MBCE}(c_i, z_i) = \frac{1}{N} \sum_{i=1}^N |c_i \cdot \log(\Pi_{LLR \rightarrow Pr}(z_i)) + (1 - c_i) \cdot \log(1 - \Pi_{LLR \rightarrow Pr}(z_i))|. \quad (6)$$

By limiting the distribution to $\Gamma_{\rho, \eta_{ABP}, \ell_{MBCE}}(S, A_1, A_2)$, we have a better control of the distribution of \mathbf{y} , and consequently of \mathbf{z} , such that \mathbf{y}_γ has higher κ on average. The intuition guiding us, again, is that higher κ words lie close to the decision boundaries. Referring to III-B, we need to choose A_1, A_2 such that $K_{\rho, \eta_{ABP}, \ell_{MBCE}}(S, A_1, A_2)$ is maximized.

E. Correlation with SNR

Figures 2 and 3 show the correlation of the above parameters to ρ and T . In both figures 100,000 codewords were simulated per ρ on code with length of 63 bits. Regarding Figure 2, one can see that each ρ defines a different probability distribution of d_H values. This figure is unique for each code length and simulated ρ . The higher the SNR - the lower the d_H center of this probability distribution. High ρ includes high amount of no errors frames, while low ρ value induces lots of high noise

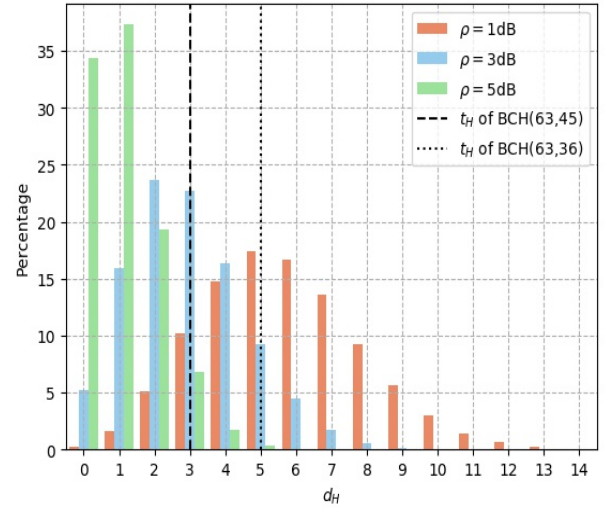


Fig. 2: Distribution of d_H by ρ

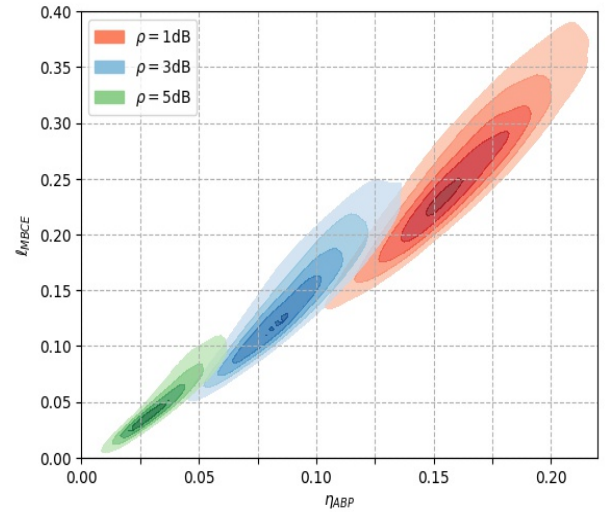


Fig. 3: Distribution of reliability by ρ

received words with d_H higher than t_H . Both t_H values for the two codes BCH(63,36) and BCH(63,45) are also plotted on this figure. Figure 3 represents similar notion in regard to reliability. Each ρ defines a probability distribution over the two parameters so that the higher the ρ is, the closer the distribution is to the origin. Here we do not have a defined threshold for correct and highly incorrect words, \mathbf{y} , as before, thus we must sample from this probability distribution much more carefully.

One thing we ignored so far is the evolution of the decoder during training. Obviously, as the decoder trains its' decision regions are altered - changing the optimal θ, S to sample by. In order to train the decoder with \mathbf{y} close to the decision boundaries at every stage, the distribution $\Gamma_\theta(S)$ we draw from must change actively during the training. A known method in machine-learning field for doing so is called - *active learning*.

IV. ACTIVE LEARNING

Active learning is a supervised learning method, which deals with an oracle that actively chooses the samples from a large pool of unlabelled data to feed the model. The oracle can be human annotator or a machine based one. Two important questions regarding this process are "why is active learning used" and "how is a batch queried". The solution to the former question is straightforward - the reason for using this method is the queried batch is assumed to benefit the training of the model more than using a random training batch, on average. The solution to the second question is achieved by introducing a metric which shows informativeness. At each training step, the batch with the highest metric value is considered the most informative thus it is queried. It is widely used in medical systems and in situations when annotating data is expensive, thus training data must be chosen punctiliously. For additional information on active learning see [16], [17].

In a stream based approach batches are generated one by one. A selective sampling approach is one in which the data to be queried is selected based on some metric. An underlying assumption in active learning stream-based selective sampling approach is that data is free to obtain. In our error-correcting codes domain data is unlimited when the channel model is known (as AWGN) or can be fairly easily collected when channel model is unknown, and we do not need to annotate it by hand. This is a huge advantage and a strong claim in favor of using this method in decoding. Traditionally, this method is used on unlabeled stream or pool of data. In this case one would want to choose which samples are worth labeling and training. In our case all samples are labeled. Therefore, our goal is to perform the training procedure with the highest κ for the received words.

We hereby present the main two active learning approaches taken.

A. Stream-Based Selective Sampling by Hamming Distance

The first approach is presented in Algorithm 1, where at each time step, the current neural model (line 6) determines the next queried batch (line 8) for the model update (line 10). This algorithm is based on intuitions from Subsection III-C, remove successfully decoded \mathbf{y} in addition to very noisy \mathbf{y} from training (lines 7-8). These received words are far from the decision boundary thus harm training. Why these \mathbf{y} can harm the training can also be explained from the learning signal perspective. On one hand, the real signal is nearly impossible to be recovered from a very noisy \mathbf{y} , thus the learning signal towards a minima is very low. On the other hand, for very reliable \mathbf{y} , the learning signal is low, since for every direction of decision the model takes these reliable words will be decoded successfully. Thus, they are not informative for learning.

B. Stream-Based Selective Sampling by Reliability Parameters

The second approach we present exploits the reliability of a given \mathbf{y} , see Algorithm 2. Inspired by the common uncertainty sampling query framework, we first calculate

Algorithm 1: Stream-based sampling by Hamming distance

Initialization: DEC as in [1]
Input : current decoder DEC
 $S = \{s_1, \dots, s_n\}$ set of SNR values
 $A = \{1, \dots, d_{max}\}$ set of d_H values
 \mathbf{c} encoded word
Output : improved model DEC

- 1 **SampleByDistance** (DEC, S, A, \mathbf{c})
- 2 **while** *error decreases* **do**
- 3 sample batch Q from $\Gamma_{\rho, d_H}(S, A)$;
- 4 **for** \mathbf{y} in Q **do**
- 5 $d_{in} \leftarrow \text{dist}(\Pi_{HD}(\mathbf{y}), \mathbf{c})$;
- 6 $d_{out} \leftarrow \text{dist}(\hat{\mathbf{c}}, \mathbf{c})$;
- 7 **if** $d_{out} = 0$ or $d_{out} \geq d_{in}$ **then**
- 8 $Q \leftarrow Q \setminus \mathbf{y}$;
- 9 **end**
- 10 $DEC \leftarrow$ update model based on Q ;
- 11 **end**
- 12 **return** DEC ;

Algorithm 2: Stream-based sampling by reliability parameters

Initialization: DEC as in [1]
Input : current decoder DEC
 $S = \{s_1, \dots, s_n\}$ set of SNR values
 b desired batch size
 \mathbf{c} encoded word
Output : improved model DEC

- 1 **SampleByReliability** (DEC, S, m, \mathbf{c})
- 2 $\boldsymbol{\mu}, \boldsymbol{\Sigma} \leftarrow$ ChoosePrior(S, \mathbf{c});
- 3 **while** *error decreases* **do**
- 4 sample batch Q from $\Gamma_{\rho}(S)$;
- 5 $\eta_{ABP} \leftarrow$ calculate by equation (5) per word;
- 6 $\ell_{MBCE} \leftarrow$ calculate by equation (6) per word;
- 7 $\boldsymbol{\theta} \leftarrow [\eta_{ABP}, \ell_{MBCE}]$;
- 8 $\mathbf{w} \leftarrow f(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$;
- 9 $\tilde{\mathbf{w}} \leftarrow \mathbf{w} / \|\mathbf{w}\|_1$;
- 10 $\tilde{Q} \leftarrow$ random sampling b words from Q w.p $\tilde{\mathbf{w}}$;
- 11 $DEC \leftarrow$ update the model based on \tilde{Q} ;
- 12 **end**
- 13 **return** DEC ;

$\Gamma_{\rho, \eta_{ABP}, \ell_{MBCE}}(S, A_1, A_2)$ for several untrained BP decoders with different number of iterations $\tau_{set} = \{\tau_1, \dots, \tau_r\}$ empirically. We chose to query each batch by setting a prior on η_{ABP}, ℓ_{MBCE} . We elaborate on the prior and batch selections. Firstly, the prior was chosen as a Normal distribution with expectation, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, over \mathbf{y} that are decodable by adding iterations to the standard BP decoder. The prior selection is summarised in Algorithm 3. These \mathbf{y} are assumed to be close to the decision boundaries, since BP decoders with additional iterations are able to decode them. We want the WBP to compensate for these additional iterations by training. Secondly, in Algorithm 2, the batch was queried by

Algorithm 3: Choose prior

Input : $S = \{s_1, \dots, s_n\}$ set of SNR values
 \mathbf{c} encoded word

Data : $\tau_{set} = \{\tau_1, \dots, \tau_r\}$ set of iterations
 $\epsilon_{set} = \{\epsilon_1, \dots, \epsilon_{r+1}\}$ set of colors

Output: prior distribution parameters μ, Σ

```

1 ChoosePrior ( $S, \mathbf{c}$ )
2   sample batch  $Q$  from  $\Gamma_\rho(S)$ ;
3    $\eta_{ABP} \leftarrow$  calculate by eq. (5);
4    $\ell_{MBCE} \leftarrow$  calculate by eq. (6);
5   for  $\tau_i$  in  $\tau_{set}$  do
6      $DEC \leftarrow$  construct BP with  $\tau_i$  iterations;
7     for  $\mathbf{y}$  in  $Q$  do
8        $d \leftarrow \text{dist}(\hat{\mathbf{c}}, \mathbf{c})$ ;
9       if  $d = 0$  then
10        | Plot point  $(\eta_{ABP}, \ell_{MBCE})$  in color  $\epsilon_i$ ;
11       else
12        | Plot point  $(\eta_{ABP}, \ell_{MBCE})$  in color  $\epsilon_{i+1}$ ;
13       end
14     end
15    $\mu, \Sigma \leftarrow$  set empirically on decodable words;
16   return  $\mu, \Sigma$ ;
```

performing a few trivial steps (lines 2-9). The last step (line 10) includes random sampling of a given size batch by the normalized weights as the probabilities, without replacement.

One important note is that the uncertainty sampling method is usually performed over the neural model output signal, while here we use it over the input signal. That is because the multiple BP decoders are the baseline for improvement, not the weighted decoder.

V. EXPERIMENTS AND RESULTS

We present the results of training and applying the approaches mentioned in IV for three different linear codes BCH(63,45), BCH(63,36), BCH(127,64) with $t_H = 3$, $t_H = 5$ and $t_H = 10$, respectively. We use the cycle-reduced (CR) parity-check matrices as appear in [18], thus evaluating our method when the number of short cycles is already small and improvement by altering weights is harder to achieve. The number of iterations is chosen as 5 as in [1], [2], [4]–[6], [9], who set a benchmark in the field. The zero codeword is used for training, due to symmetry, as in [1], [2]. It also serves as the codeword in Algorithms 1 and 2. All other training relevant hyperparameters are summarised in Table I. All WBP decoders are trained until convergence. We apply two methods - Hamming distance based and reliability based. Regarding the active learning hyperparameters, for the distance approach, and in order to stay consistent, we chose the same d_{max} for the two short codes. All hyperparameters are summarised in Table II. As a follow-up to section III-C, we also apply d_H filtering to the reliability method. This is referred as the reliability & d_H filtering in Table II.

We simulate the WBP over a validation set of 1dB to 10dB until at least 1000 errors are accumulated at each given point. In addition, we adopt the syndrome based early

TABLE I: Training Hyperparameters

Hyperparameters	Values
Architecture	Feed Forward
Initialization	as in [1] (*)
Loss Function	BCE with Multiloss
Optimizer	RMSPROP
ρ_t range	4dB to 7dB
Learning Rate	0.01
Batch Size	1250 / 300 words per SNR (**)
Messages Range	(-10, 10)

(*) $w_{i,v}$ in eqs. (1) and (2) set to constant 1 since no additional improvement was observed.

(**) for the 63 / 127 code length, respectively.

termination, as we saw that some correctly decoded codewords were misclassified again by the following layers. This can also benefit complexity since the average number of iterations is less than or equal to 5 when using this rule.

Results for the simulation are presented in Figure 4. One can see that both distance-based and reliability-based approaches outperform the original BP-FF model with hyperparameters as in [1], [2]. We separate the contribution of our methods to two different regions. At the waterfall region the improvement varies from 0.25dB to 0.4dB in FER and 0.2dB to 0.3dB in BER for the different codes. At the error-floor region, the gain is increased by 0.75dB to 1.5dB in FER and by 0.75 to 1dB in BER for all codes. The best decoding gains per code are summarized in Table III. The measured error value, where the gain is observed, is specified in parentheses. Comparing to [13] in the BER graphs, a gain of 0.25dB is achieved in the CR-BCH(63,36) code, while in CR-BCH(127,64) one can observe similar performance. Furthermore, the difference in gains between the reliability curve and the reliability & d_H filtering curve indicates that the two methods indeed train on different distributions of words.

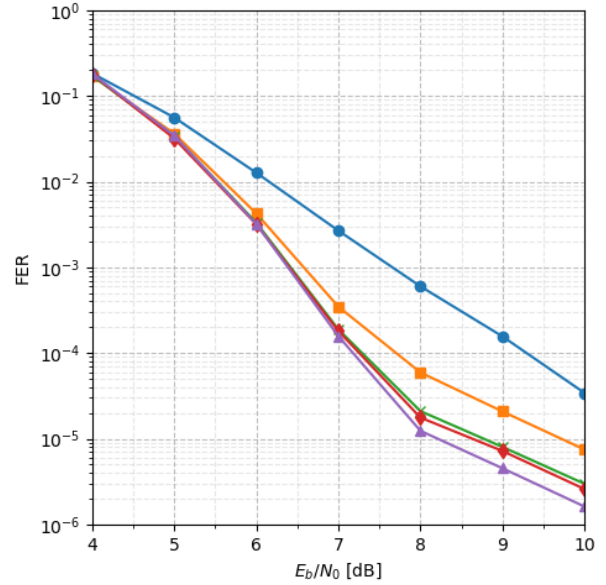
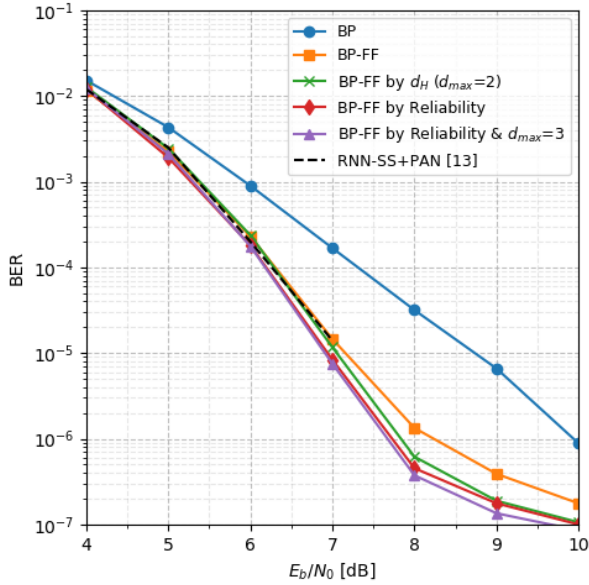
The FER metric is observed to gain the most from all approaches, with the reliability & d_H filtering approach having the best performance. One conjecture is that all these methods are optimized to improve FER directly. For the Hamming distance approach, lowering the number of errors in a single codeword reflects the FER directly. The reliability parameters are taken as a mean over the received words, thus adding more information on each \mathbf{y} rather than on each single bit, y_i . One can see that all methods achieve better performance while keeping the same decoding complexity as before in [1], [2]. This is achieved solely by smartly sampling the data to train the neural decoder.

VI. CONCLUSIONS

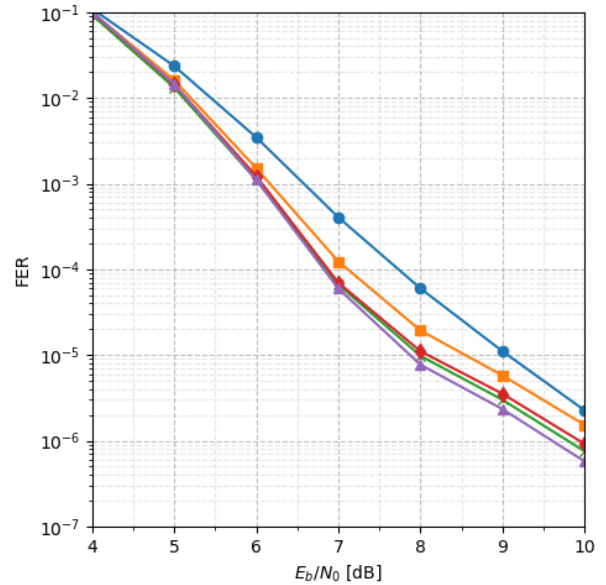
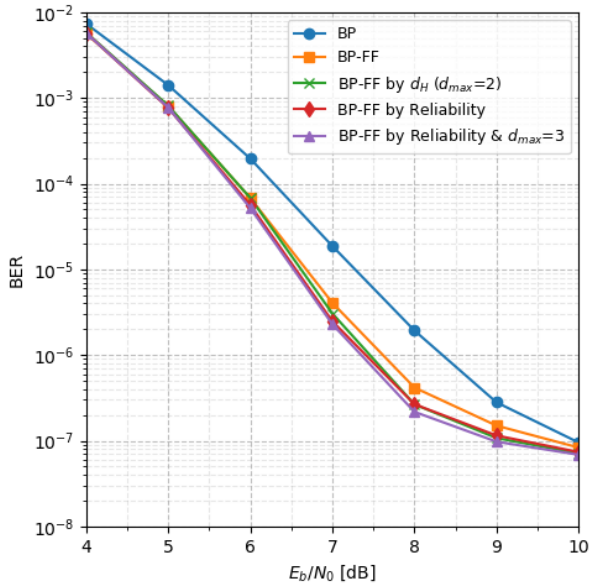
In this paper we proposed two novel sampling methods, incorporating error decoding measures with methodologies from the vast machine-learning field. Increases in performance of up to 0.4dB at the waterfall region, and of up to 1.5dB at the error-floor region, compared to the original WBP, are possible with no decoding complexity penalty, only by smartly sampling the training set. Furthermore, note that an aggregated

TABLE II: Active Learning Hyperparameters

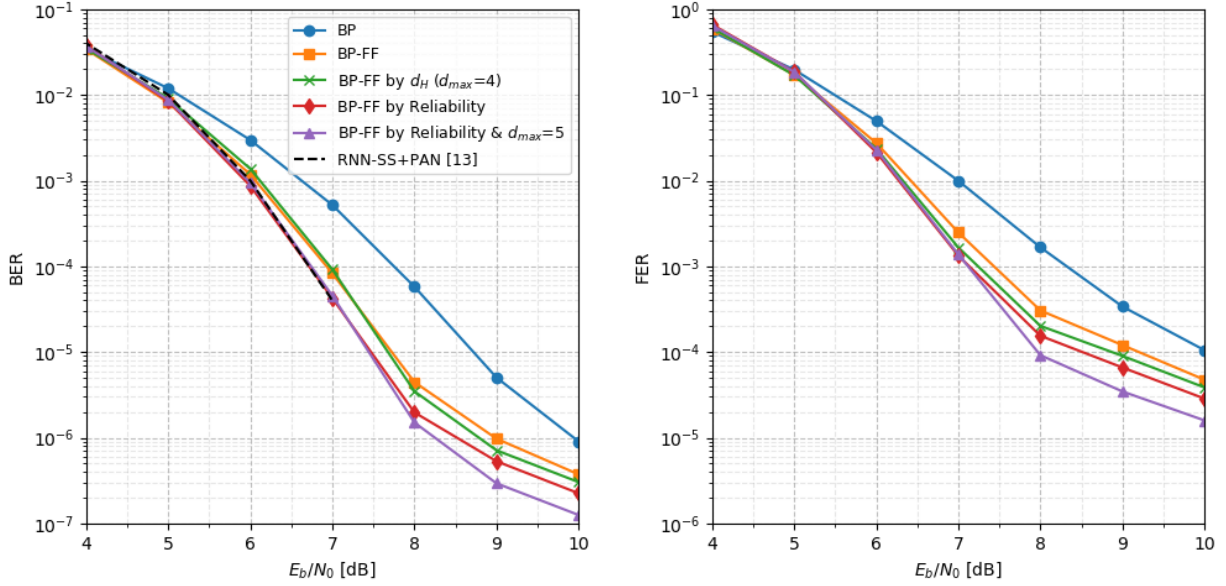
Method	Hyperparameters	CR-BCH N=63	CR-BCH N=127
Hamming distance	d_{max}	2	4
Reliability	τ_{set}	{5, 7, 10, 15}	
	μ	(0.025, 0.1)	(0.03, 0.1)
	Σ	$\begin{bmatrix} 6.25 \cdot 10^{-4} & 0 \\ 0 & 5.625 \cdot 10^{-3} \end{bmatrix}$	
Reliability & d_H filtering	d_{max}	3	5
	τ_{set}	{5, 7, 10, 15}	
	μ	(0.025, 0.1)	(0.03, 0.1)
	Σ	$\begin{bmatrix} 6.25 \cdot 10^{-4} & 0 \\ 0 & 5.625 \cdot 10^{-3} \end{bmatrix}$	



(a) CR-BCH(63,36)



(b) CR-BCH(63,45)



(c) CR-BCH(127,64)

Fig. 4: BER and FER comparison: (●) Regular BP, (■) BP-FF [1], (×) Hamming distance approach, (◆) Reliability approach, (▲) Reliability approach with filter by d_H

TABLE III: Best Decoding Gains (*)(**)

Code \ Region	Waterfall		Error-floor	
	BER[dB]	FER[dB]	BER[dB]	FER[dB]
CR-BCH(63,36)	0.2 (10^{-5})	0.25 (10^{-3})	1 ($4 \cdot 10^{-7}$)	1.5 (10^{-5})
CR-BCH(63,45)	0.2 (10^{-5})	0.25 (10^{-4})	0.75 ($2 \cdot 10^{-7}$)	0.75 ($3 \cdot 10^{-6}$)
CR-BCH(127,64)	0.3 (10^{-4})	0.4 (10^{-3})	0.75 (10^{-6})	1.25 (10^{-4})

(*) all gains are WRT the original WBP [1],[2].

(**) The measured error value, where the gain is observed, is specified in parentheses.

increase in gain of about 2dB in high SNR, compared to BP, is achieved. We provided general guidelines for choosing training data in communications, starting in data exploration, validating assumptions by experiments and finally developing active learning based algorithms. We highlighted that SNR does not reveal the whole story. By introducing other key parameters one can have more control over the training data. Our conjecture is that sampling close to the decision boundary is crucial. At last, we urge the readers to seek sampling schemes in their communication application.

As for the next step, one may aim to find new ways of incorporating important parameters in training and validation for improved results. Likewise, one may explore a reinforcement learning algorithm which finds the optimal parameters during training with no conjectures whatsoever. Another direction is applying the proposed methods into the mRRD decoder [19], [20] for approaching maximum-likelihood performance with further complexity reduction. Lastly, further analysis of dropped training samples could enhance explainability and provide insights about the proposed methods.

ACKNOWLEDGMENT

We would like to thank Eran Asa for the insightful discussions. We also thank the reviewers and the editor for their beneficial comments.

REFERENCES

- [1] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL., USA, 2016, pp. 341–346.
- [2] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 119–131, Feb. 2018.
- [3] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.
- [4] M. Lian, C. Häger, and H. D. Pfister, "What can machine learning teach us about communications?" in *IEEE Information Theory Workshop (ITW)*, Guangzhou, China, 2018, pp. 410–414.
- [5] W. Xu, Z. Wu, Y.-L. Ueng, X. You, and C. Zhang, "Improved polar decoder based on deep learning," in *IEEE International Workshop on Signal Processing Systems (SiPS)*, Lorient, France, Oct. 2017, pp. 1–6.
- [6] W. Xu, X. You, C. Zhang, and Y. Be'ery, "Polar decoding on sparse graphs with deep learning," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2018, pp. 599–603.

- [7] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *IEEE Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, 2017, pp. 1–6.
- [8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *27th international conference on machine learning (ICML)*, Haifa, Israel, 2010, pp. 807–814.
- [9] A. Bennatan, Y. Choukroun, and P. Kisilev, "Deep learning for decoding of linear codes—a syndrome-based approach," in *IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, 2018, pp. 1595–1599.
- [10] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," *arXiv preprint arXiv:1805.09317*, 2018.
- [11] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Deepturbo: Deep turbo decoder," *arXiv preprint arXiv:1903.02295*, 2019.
- [12] M. Benammar and P. Piantanida, "Optimal training channel statistics for neural-based decoders," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2018, pp. 2157–2161.
- [13] M. Lian, F. Carpi, C. Häger, and H. D. Pfister, "Learned belief-propagation decoding with simple scaling and snr adaptation," *arXiv preprint arXiv:1901.08621*, 2019.
- [14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [15] T. Richardson and R. Urbanke, *Modern coding theory*. Cambridge university press, 2008.
- [16] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.
- [17] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowledge and information systems*, vol. 35, no. 2, pp. 249–283, 2013.
- [18] M. Helmling, S. Scholl, F. Gensheimer, T. Dietz, K. Kraft, S. Ruzika, and N. Wehn, "Database of channel codes and ml simulation results," www.uni-kl.de/channel-codes, 2019.
- [19] E. Nachmani, Y. Bachar, E. Marciano, D. Burshtein, and Y. Be'ery, "Near maximum likelihood decoding with deep learning," in *International Zurich Seminar on Information and Communication (IZS)*, Zurich, Switzerland, 2018, pp. 40–44.
- [20] I. Dimnik and Y. Be'ery, "Improved random redundant iterative hdpc decoding," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1982–1985, 2009.