Optimal Training Channel Statistics for Neural-based Decoders

Meryem Benammar
Department of Electronics, Optronics, and Signal processing
ISAE-Supaéro
Toulouse, 31000, France
Email: meryem.benammar@isae-supaero.fr

Pablo Piantanida
L2S, CentraleSupélec-CNRS-Université Paris-Sud
MILA, Université de Montréal
Montréal, QC H3T 1N8, Canada
E-mail: pablo.piantanida@centralesupelec.fr

***Abstract*—This work investigates the design of End-to-End channel coding based on deep learning. The focus is on the design of neural networks based channel decoders. We demonstrate the existence of an optimal training statistic for the cross-entropy loss which allows the network to generalize to channel statistics unseen during training while performing close to their optimal decision rule. Numerical results illustrate an application to Polar coding on binary input memoryless channels.**

I. Introduction

The channel coding problem, shown in Fig. 1, is that of transmitting reliably a message M through a noisy memoryless stationary communication channel defined by a pmf (resp. pdf) $P_{Y|X}$ (resp. $f_{Y|X}$). To cope with the noise of this channel, the message $M \in \{1, \dots, 2^{nR}\}$ if rate R is mapped through a Forward Error Correction (FEC) code into the channel input X^n blocklength n, which allows to introduce redundancy in the transmitted symbols. At the receiver, the channel decoding problem consists thus in recovering the message M with arbitrarily low probability of error from the noisy channel output Y^n. The optimal decoder, i.e., *maximum a posteriori (MAP) decoder*, identifies the most likely message M given an observation Y^n, relying on an exhaustive enumeration of all possible codewords. As such, its complexity is exponential in the dimension of the blocklength n. Thus, approaching the MAP probability of error with reasonable complexities has been the leitmotiv for algorithmic designs of error correction codes. Many structured non-exhaustive decoding algorithms (Viterbi, BCJR, Belief Propagation,...) have been developed to perform close to MAP for a family of error correction codes such as Convolutional codes, LDPC codes, and Polar Codes. Their performances and their respective complexities have been widely assessed and are, to date, well understood. Yet, the MAP performance can be achieved with non-structured solutions based, for instance, on supervised learning and more specifically, on neural networks as shown in Fig. 2.

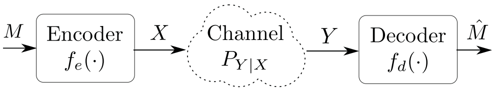

Fig. 1. Cannel coding problem.

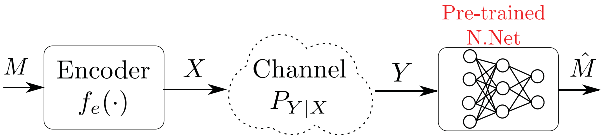

Fig. 2. Neural Network based channel decoder.

Early works [1]–[3] and more recent ones [4]–[6] all prove that quasi-optimal performances can be obtained with NN-based channel decoders for a variety of communication settings. NN-based channel decoders offer the advantage of trading the on-line complexity of the decoding for an off-line complexity at the training, and are, thus, rather appealing for low complexity/low latency communications. The design of these NN-based solutions is, however, rather heuristic (e.g. on the number of layers, number of neurones, activation function and structure of the network). In particular, authors in [4] showed that a pre-trained NN could achieve arbitrarily close performances to those of the MAP for a family of polar codes provided that it is trained at the adequate channel noise. In this work, we infer the existence and design of this training channel statistic through a surrogate loss, cross-entropy loss, which describes differently the performance of the neural network, rather than the probability of misdetection. Though derived through a surrogate loss, and not through the effective loss, the theoretic channel statistic we obtain coincides with the observed optimal training statistic for a family of codes and channel statistics, namely Polar codes [7] over Binary Symmetric Channels (BSC) and Additive White Gaussian Noise (AWGN) channels.

This work is organized as follows. In Section II, we introduce the channel coding problem and draw an analogy with a detection problem. In Section III, we describe neural-network based channel decoders and state the optimization problem on the training statistic. In Section IV, we characterize the optimal training statistic for two families of channels, namely, the BSC and the AWGN channels. Conclusions and further analysis are drawn in Section V.

Notations

Throughout this paper, pmf (resp. pdf) stands for probability mass (resp. density) function, while iid stands for independent

978-1-5386-9218-9/18/$31.00 ©2018 IEEE

and identically distributed. Sets and alphabets are denoted in caligraphic letters, e.g., $\mathcal{X}$. Random variables (resp. their realizations) are denoted by capital case letters (resp. lower case latters), e.g. X (resp. x). The pmf (resp. pdf) of a random variable X is denoted as $P_X(\cdot)$ (resp. $f_X(\cdot)$), while the conditional pmf (resp. pdf) of a random variable X knowing Y, is denoted as $P_{X|Y}(\cdot|\cdot)$ (resp. $f_{X|Y}(\cdot|\cdot)$). The operators $\mathbb{E}(\cdot)$ and $\mathbb{1}[\cdot]$ stand respectively for the expectation operation and the indicator function. The entropy of a random variable X is denoted as $H(X)$, while the conditional entropy of X knowing Y is denoted as $H(X|Y)$. We denote the *Kullback Leibler* (KL) divergence between two pmfs P and Q as $D_{\mathrm{KL}}(P\|Q)$, and the conditional KL divergence between two conditional pmfs $P_{M|Y}$ and $Q_{M|Y}$ as:

$$D_{\mathrm{KL}}(P_{M|Y}\|Q_{M|Y}|P_Y) \triangleq \mathbb{E}_{P_Y} D_{\mathrm{KL}}(P_{M|Y=y}\|Q_{M|Y=y}).$$

An n-dimensional sequence is denoted as $x^n = (x_1, \ldots, x_n)$ where x_i is the i-th component of the sequence. Scalars and vectors are denoted similarly unless the dimension is explicitely given in a superscript. $\otimes$ denotes the Kronecker product while $\oplus$ stands for the binary XOR operation.

II. Preliminary Definitions

A. The channel decoding problem

The channel decoding problem can be regarded as a task of inferring an unobserved random variable M through a noisy observation Y^n, as follows:

- The random variable of interest M has an associated prior P_M defined over its support set $\mathcal{M}$;
- The variable M is mapped through a deterministic mapping into X^n, the mapping being one to one and represents the FEC encoder;
- The noisy observation Y^n is related to X^n through the memoryless channel statistic $P_{Y|X}$;
- The estimator, or decision rule, which associates to each observable y^n an estimate $\hat{m}$ is assumed to be stochastic following the conditional p.m.f $Q_{\hat{M}|Y^n}$[1].

The Markov chain which describes then this detection problem is given by

$$M \xrightarrow{P_{X^n|M}} X^n \xrightarrow{P^n_{Y|X}} Y^n \xrightarrow{Q_{\hat{M}|Y^n}} \hat{M}. \tag{1}$$

B. Probability of misdetection:

The probability of misdetection associated with a decision rule $Q_{\hat{M}|Y^n}$ is given by

$$P_e\big(Q_{\hat{M}|Y^n}\big) \triangleq \mathbb{P}\big(\hat{M} \neq M\big) = 1 - \mathbb{E}_{P_{MY^n}}\big(Q_{\hat{M}|Y^n}\big). \tag{2}$$

Let P_{MY^n} be the joint pdf (pmf) associated with the prior pmf P_M and the observation pdf (pmf) $P_{Y^n|M}$. In the following, we introduce two decision rules which will be crucial to the analysis of the present detection problem:

1) The A Posteriori Probability (APP) decision rule

$$Q^{\mathrm{APP}}_{\hat{M}|Y^n}(\hat{m}|y^n) = \frac{P_M(\hat{m})P_{Y^n|M}(y^n|\hat{m})}{P_{Y^n}(y^n)},$$

for all $(m, y^n) \in \mathcal{M} \times \mathcal{Y}^n$;

2) the Maximum A Posteriori (MAP) decision rule

$$Q^{\mathrm{MAP}}_{\hat{M}|Y^n}(\hat{m}|y^n) \triangleq \mathbb{1}\big[\hat{m} = f_{\mathrm{MAP}}(y^n)\big], \tag{3}$$

where f_{MAP} is defined by

$$\begin{aligned} f_{\mathrm{MAP}} &: \mathcal{Y} \to \mathcal{M} \\ & y^n \to \hat{m} = \operatorname*{argmax}_{m\in\mathcal{M}} P_{M|Y^n}(m|y^n). \end{aligned} \tag{4}$$

Hereafter, we state some well-known results, with sketch of proofs when necessary.

Lemma 1 (Optimal decision rule)**.** *The optimal decision rule for probability of misdetection is given by the MAP rule, i.e.,*

$$\operatorname*{argmin}_{Q_{\hat{M}|Y^n}} P_e\big(Q_{\hat{M}|Y^n}\big) = Q^{\mathit{MAP}}_{\hat{M}|Y^n}. \tag{5}$$

Proof. A simple proof can be found in [8]. □

The optimal decision rule from a probability of misdetection point of view is purely deterministic, and concentrates the APP around its maximum value. In the following, we analyze the cross-entropy loss and derive its optimal decision rule.

C. Cross-entropy loss:

Evaluating the probability of misdetection might be intractable in practice due to the large data dimensions. Thus, we will write a bound on the probability of error which relies on a surrogate loss, which then, will allow for richer interpretations and bounding techniques. The cross-entropy loss is defined as

$$\mathcal{L}\big(P_{Y^n|M}, Q_{\hat{M}|Y^n}\big) \triangleq \mathbb{E}_{P_{MY^n}}\Big[-\log\big(Q_{\hat{M}|Y^n}\big)\Big] \tag{6}$$

This surrogate loss is, comparatively to the probability of error, less common in error correction coding analysis, yet, its analysis will prove to yield intuitions and easier bounds on the channel decoding problem beforehand.

Lemma 2 (Cross-entropy loss)**.** *The optimal decision rule for the cross-entropy loss is the APP, i.e.,*

$$\operatorname*{argmin}_{Q_{\hat{M}|Y^n}} \mathcal{L}\big(P_{Y^n|M}, Q_{\hat{M}|Y^n}\big) = Q^{\mathit{APP}}_{\hat{M}|Y^n} = P_{M|Y^n}, \tag{7}$$

and the corresponding loss is the conditional entropy

$$\mathcal{L}\big(P_{Y^n|M}, P_{M|Y^n}\big) = H(M|Y^n). \tag{8}$$

Proof. The proof of this lemma follows by noticing that

$$\begin{aligned} \mathcal{L}\big(P_{Y^n|M}, Q_{\hat{M}|Y^n}\big) &= H(M|Y^n) \\ + D\left(P_{M|Y^n}\|Q_{\hat{M}|Y^n}|P_{Y^n}\right) &\overset{(a)}{\geq} H(M|Y^n), \end{aligned} \tag{9}$$

where (a) follows from the positivity of the Kullback-Leibler divergence. As such, the cross-entropy loss is minimized when the decision rule $Q_{\hat{M}|Y^n}$ is equal to the APP $P_{M|Y^n}$, and its minimum value amounts to $H(M|Y^n)$. □

[1] We assumed that the decoder can be stochastic which will be shown in Lemma 1 to be equivalent to restricting to deterministic decision rules.

Unlike the misdetection probability, the optimal cross-entropy decision rule does not concentrate the APP distribution around its maximum, and is thus, non-deterministic. The cross-entropy loss is related to the probability of error.

Lemma 3 (Cross-entropy and error probability)**.** *The cross-entropy loss is related to the misdetection probability by*

$$\begin{aligned} P_e\big(Q_{\hat{M}|Y^n}\big) &\leq 1 - \exp\Big(-\mathcal{L}\big(P_{Y^n|M}, Q_{\hat{M}|Y^n}\big)\Big) && (10)\\ &\leq \mathcal{L}\big(P_{Y^n|M}, Q_{\hat{M}|Y^n}\big), && (11) \end{aligned}$$

for any decision rule $Q_{\hat{M}|Y^n}$.

Proof. The proof of this lemma resides in a simple application of Jensen's inequality and is ommitted here. □

Thus, decreasing the cross-entropy loss amounts to decreasing the probability of misdetection. In the remainder of this work, we investigate the channel coding problem from a cross entropy point of view rather than a probability of error point of view in order to write meaningful bounds and evaluate them.

III. Neural-based channel decoding

Recent works on deep learning based channel decoding [4] have shown that promising performances, close to the MAP decoder, can be obtained by neural networks decoders with reasonable complexity. However, while most works concentrate on deriving new *deep* structures which mimic the optimal decoding performances, their design is still rather heuristic, and the roles of each of the design parameters are yet to investigate. Besides the conventional meta-parameters of neural networks (number of layers and neurones, activation functions,...), the training channel statistic plays as well a crucial role since it defines the distribution of the training dataset and conditions, thus, the learning capability as well as the generalization capability of the decoder.

Previous works [4] have shown the existence of a particular optimal design, in terms of the training channel statistics, which conditions the capacity of the network to generalize to channel statistics unseen during the training.

A. Training/validation statistics mismatch

Assume that the training is performed with a statistic $P^t_{Y^n|M}$ and that, at the end of the training process, the learned decision rule $Q^t_{\hat{M}|Y^n}$ is the one optimizing the cross-entropy risk $\mathcal{L}\big(Q_{\hat{M}|Y^n}\big)$, i.e. : $Q^t_{\hat{M}|Y^n} \equiv P^t_{M|Y^n}$.

Assume now that the validation of the obtained decision rule is performed in an environment with a mismatched (channel) statistic $P^v_{Y^n|M}$. The purpose of this section is to characterize the increase that is induced to the cross-entropy loss by this mismatched distribution, compared to the loss obtained with the optimal validation decision rule, i.e., when $Q^v_{\hat{M}|Y^n} \equiv P^t_{M|Y^n}$.

Lemma 4 (Training/validation mismatch)**.** *The increase in cross-entropy loss induced by mismatched training/validation statistics is tantamount to*

$$\begin{aligned} &\mathcal{L}\big(P^v_{Y^n|M}, Q^t_{\hat{M}|Y^n}\big) - \mathcal{L}\big(P^v_{Y^n|M}, Q^v_{\hat{M}|Y^n}\big)\\ &= D_{\mathrm{KL}}\big(P^v_{M|Y^n} \| P^t_{M|Y^n} | P^v_{Y^n}\big) && (12)\\ &= D_{\mathrm{KL}}\big(P^v_{Y^n|M} \| P^t_{Y^n|M} | P_M\big) - D_{\mathrm{KL}}\big(P^v_{Y^n} \| P^t_{Y^n}\big). && (13) \end{aligned}$$

Proof. The proof follows by writing that

$$\begin{aligned} &\mathcal{L}\big(P^v_{Y^n|M}, Q^t_{\hat{M}|Y^n}\big) - \mathcal{L}\big(P^v_{Y^n|M}, Q^v_{\hat{M}|Y^n}\big)\\ &= H^v(M|Y^n) + D_{\mathrm{KL}}\big(P^v_{M|Y^n} \| P^t_{M|Y^n} | P^v_{Y^n}\big) - H^v(M|Y^n) && (14)\\ &= D_{\mathrm{KL}}\big(P^v_{M|Y^pn} \| P^t_{M|Y^n} | P^v_{Y^n}\big) && (15) \end{aligned}$$

and by expanding

$$\begin{aligned} &D_{\mathrm{KL}}\big(P^v_{M|Y^n} \| P^t_{M|Y^n} | P^v_{Y^n}\big)\\ &= D_{\mathrm{KL}}\big(P^v_{Y^n|M} \| P^t_{Y^n|M} | P_M\big) - D_{\mathrm{KL}}\big(P^v_{Y^n} \| P^t_{Y^n}\big), && (16) \end{aligned}$$

which follows from standard analytic manipulations. □

The cross-entropy loss increases under mismatched training/validation statistics, and an optimal training statistic would be one which yields a decision rule close to the optimal for as many validation statistics as necessary.

B. Optimal training statistic

In this section, we formalize an optimization problem to determine the optimal training statistic for the channel decoding problem. Let us fix a training noise statistic $P^t_{Y^n|M}$, and assume that we wish to validate the obtained decision rule over a set $\mathcal{V}$ of V different validation statistics $P^v_{Y^n|M}$.

Authors in [4] derived, assuming an AWGN channel, a measure for the quality of a training noise variance σ_t^2, which is referred to as the NVE. The NVE is defined as the point-wise ratio between the Bit Error Rate (BER), i.e., probability of misdetection, obtained by a training noise variance σ_t^2, as compared to the BER of the MAP, i.e., the minimum possible probability of misdetection. The NVE writes, in the Gaussian case, as

$$\mathrm{NVE}(\sigma_t^2) = \sum_{\sigma_v^2} \frac{P_e\big(Q^{t,NN}_{\hat{M}|Y^n}\big)}{P_e\big(Q^{v,\mathrm{MAP}}_{\hat{M}|Y^n}\big)}. \quad (17)$$

According to [4], it was observed that the NVE exhibits a minimum around a training E_b/N_0 of 1dB.

In the sequel, we show that the optimal training statistic follows as a solution to the optimization problem:

$$\min_{P^t_{Y|M}} \sum_{v\in\mathcal{V}} \Big[\mathcal{L}\big(P^v_{Y^n|M}, P^t_{M|Y^n}\big) - \mathcal{L}\big(P^v_{Y^n|M}, P^v_{M|Y^n}\big)\Big]^2 \quad (18)$$

that is, the training statistic which minimizes the quadratic distance between the cross-entropy curve obtained by the trained neural network and the best possible curve, obtained with the APP. This target function behaves equivalent to minimizing the NVE as defined in [4].

Solving in closed form this optimization problem in (18) is rather challenging for generic channel models and channel codes. However, this can be solved in specific cases. We consider the case of a Polar code, as a FEC code, and two channel models: the BSC and the AWGN channel.

IV. Example: Polar coding over noisy channels

In the following, we assume that the FEC code is used as an (n, k)-Polar code.

A. The (n, k)-Polar code

Let us assume that the message of interest consists in k bits $M = (M_1, \dots, M_k)$ i.i.d. that are Bern(1/2) distributed, i.e.,

$$\forall (m_1, \dots, m_k) \in \{0:1\}^k \ , \ P_M(m_1, \dots, m_k) = 2^{-k}. \tag{19}$$

The input message M is mapped through a Polar code into a codeword C of n-bits trough a binary linear block mapping:

$$(c_1, \dots, c_n) = T_n \ (u_1, \dots, u_n), \tag{20}$$

where T_n indicates the n-th fold Kronecker power of the kernel $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, i.e., $T_n = T_2^{\otimes n}$ and the sequence $u^n = (u_1, ..., u_n)$ consists in the k input bits $(b_1, \dots, b_k)$ to which are appended $n-k$ bits set to 0, representing the $(n-k)$ frozen bits of the code. The indices of these frozen bits depend on the channel statistic considered at the design, and we consider in this work Arikan's construction [7]. The rate of this code is thus $R = k/n$.

In the following, we solve the optimization problem in (18) under the Polar code construction for different channel models, namely, the BSC and AWGN channels. To this end, we base our analysis on the work of [4] where the FEC code is a $(16, 8)$ Polar code, and where the Neural Network used for MAP approximation is a feedforward network with three hidden layers of corresponding sizes $[128, 64, 32]$ and with activation functions [ReLu, ReLu, ReLu, Sigmoid].

B. The Additive White Gaussian Noise (AWGN) channel

The channel input X^n is obtained by mapping the n coded bits C into n BPSK symbols, i.e., $X^n \equiv 2C - 1$.

The noisy observation Y^n consists in n consecutive iid realizations of an AWGN of variance σ^2, i.e.,

$$p^n_{Y|X}(y^n|x^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - x_i)^2}{2\sigma^2} \right). \tag{21}$$

Assume that the training and validation are performed under respective noise variances σ_t^2 and σ_v^2.

The optimization problem (18) writes then as follows:

$$\min_{\sigma_t^2} \sum_{\sigma_v^2 \in \mathcal{S}_v} D^2_{\mathrm{KL}}\left(P^v_{M|Y^n} \| P^t_{M|Y^n} | P^v_{Y^n}\right) =$$

$$\min_{p_t} \sum_{\sigma_v^2 \in \mathcal{S}_v} \left[D_{\mathrm{KL}}\left(P^v_{Y^n|M} \| P^t_{Y^n|M} | M\right) - D_{\mathrm{KL}}\left(P^v_{Y^n} \| P^t_{Y^n}\right) \right]^2.$$

Corollary 1. *The first KL divergence term*

$$D_{\mathrm{KL}}\left(P^v_{Y^n|M} \| P^t_{Y^n|M} | M\right) = \frac{n}{2} \left[\frac{\sigma_v^2}{\sigma_t^2} - 1 - \log\left(\frac{\sigma_v^2}{\sigma_t^2} \right) \right]. \tag{22}$$

Proof. The proof follows from standard analytic manipulations and is omitted here. □

The first KL divergence term is independent of the code construction and can be computed in a single-letter closed form. However, the second KL divergence term $D_{\mathrm{KL}}\left(P^v_{Y^n} \| P^t_{Y^n}\right)$ is difficult to obtain analytically, since it is the KL divergence between two Gaussian mixtures. We will thus resort to numerical computation for this term[2].

Fig. 3 shows the existence of an optimal training point in terms of

$$\frac{E_b}{N_0} = \frac{1}{2\sigma^2},$$

which coincides exactly to the one observed from numerical results in [4, Fig.2].

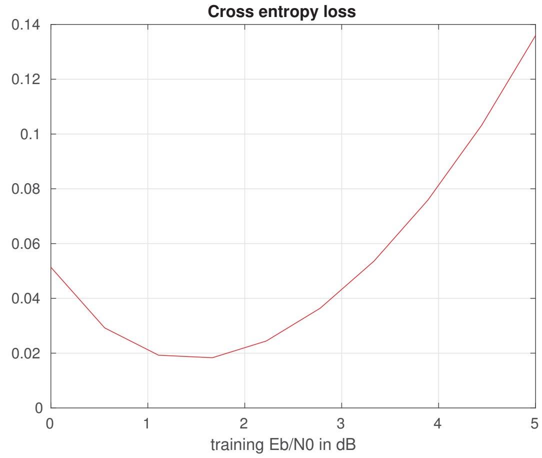

Fig. 3. Optimal training E_b/N_0 for an AWGN channel.

C. The Binary Symmetric Channel (BSC)

The noisy observations Y^n consist in n consecutive i.i.d. realizations of a Binary Symmetric Channel (BSC) with crossover probability $p \in [0, 0.5]$, i.e.,

$$P^n_{Y|X}(y^n|x^n) = \prod_{i=1}^{n} p^{y_i \oplus x_i} (1 - p)^{1 - y_i \oplus x_i}. \tag{23}$$

It is assumed that both training and validation channels follow the BSC model but with different crossover probabilities, denoted to as p_t and p_v.

Our purpose is to show the existence of an optimal p_t solution to the minimization problem:

$$\min_{p_t} \sum_{p_v \in \mathcal{P}_v} D^2_{\mathrm{KL}}\left(P^v_{M|Y^n} \| P^t_{M|Y^n} | P^v_{Y^n}\right) =$$

$$\min_{p_t} \sum_{p_v \in \mathcal{P}_v} \left[D_{\mathrm{KL}}\left(P^v_{Y^n|M} \| P^t_{Y^n|M} | M\right) - D_{\mathrm{KL}}\left(P^v_{Y^n} \| P^t_{Y^n}\right) \right]^2.$$

[2] There exist bounds on the KL divergence of two Gaussian mixtures, but we choose to compute them using Monte Carlo simulation.

Corollary 2. *For the above defined case, we have that:*

$$D_{\mathrm{KL}}\left(P^{v}_{Y^n|M} \| P^{t}_{Y^n|M} | M\right) = n\left[L_2(p_v, p_t) - h_2(p_v)\right], \quad (24)$$

where $L_2(p_v, p_t)$ is the binary cross-entropy of p_v and p_t, and is given by

$$L_2(p_v, p_t) \triangleq -p_v \log(p_t) - (1 - p_v)\log(1 - p_t). \quad (25)$$

and $H_2(\cdot)$ is defined as the binary entropy:

$$H_2(p_v) \triangleq -p_v \log(p_v) - (1 - p_v)\log(1 - p_v). \quad (26)$$

Proof. The proof follows from analytic manipulations and is omitted here. □

The first KL divergence term is computed in Corollary 2 and is independent of the code construction, and depends only on the channel statistics. However, the second KL divergence term, i.e., $D_{\mathrm{KL}}\left(P^{v}_{Y^n} \| P^{t}_{Y^n}\right)$, is more challenging to evaluate in closed form since the observations Y^n, due to the correlation introduced by the code, are no longer i.i.d. when not conditioned on the inputs X^n. When the coding rate $R = 1$, the avobe KL divergence term can be shown to be zero, however, for arbitrary positive rate, we will resort only to numerical calculations.

Fig. 4 shows the variation of the cross-entropy loss difference across various training crossover probability p_t, and proves the existence of an optimal training $p_t = 0.07$. Fig. 5 shows the BER of the MAP compared to three training statistics: $p_t = 0$, $p_t = 0.5$ and $p_t = 0.07$. The training crossover probability predicted by the optimization problem in (18) turns out to be representative of the optimal training crossover probability according to the misdetection probability.

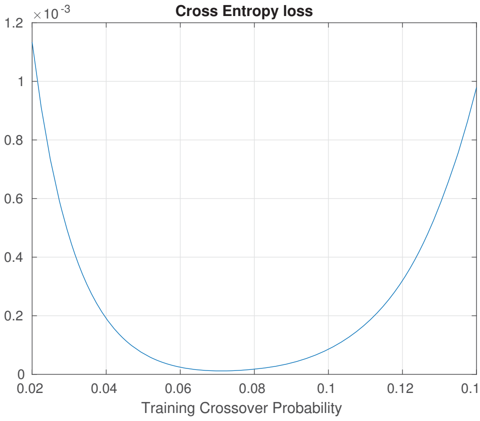

Fig. 4. Optimal training crossover probability for a BSC channel

V. Summary and Concluding Remarks

In this work, we investigated and formalized empirical evidence that neural-based channel decoders can achieve MAP performances provided that the training is performed at an optimal training statistic. This implies that –at least for the channel coding problem– training the neural networks over a noiseless setting or training it in a too noisy setting yield both poor performances. There exists an optimal noise level which allows these networks to generalize to statistics unseen during training, and though the analysis was carried out with a surrogate loss and not with the probability of error, numerical evidence matches with the theoretical predictions of our result.

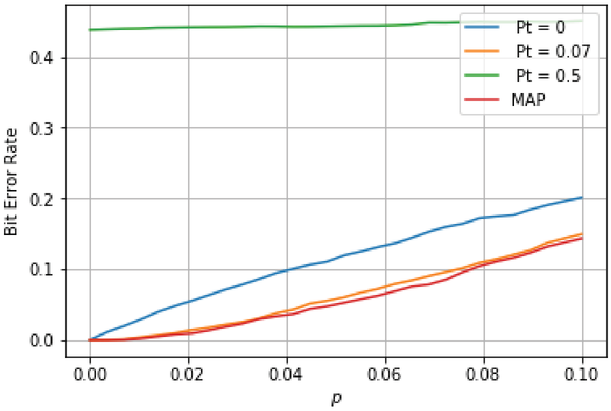

Fig. 5. BER for neural-based channel decoding under various training crossover probabilities p_t.

Acknowledgment

The work of Prof. Pablo Piantanida was supported by the European Commission's Marie Sklodowska-Curie Actions (MSCA), through the Marie Sklodowska-Curie IF (H2020-MSCAIF-2017-EF-797805-STRUDEL).

References

[1] J. Bruck and M. Blaum, "Neural networks, error-correcting codes, and polynomials over the binary n-cube," *IEEE Transactions on information theory*, vol. 35, no. 5, pp. 976–987, 1989.

[2] L. Tallini and P. Cull, "Neural nets for decoding error-correcting codes," in *Northcon 95. IEEE Technical Applications Conference and Workshops Northcon95*. IEEE, 1995, pp. 89–.

[3] X.-A. Wang and S. B. Wicker, "An artificial neural net viterbi decoder," *IEEE Transactions on communications*, vol. 44, no. 2, pp. 165–171, 1996.

[4] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *Information Sciences and Systems (CISS), 2017 51st Annual Conference on*. IEEE, 2017, pp. 1–6.

[5] E. Nachmani, Y. Beery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*. IEEE, 2016, pp. 341–346.

[6] A. Bennatan, Y. Choukroun, and P. Kisilev, "Deep learning for decoding of linear codes-a syndrome-based approach," *arXiv preprint arXiv:1802.04741*, pp. –, 2018.

[7] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[8] J. Muramatsu and S. Miyake, "On the error probability of stochastic decision and stochastic decoding," *arXiv preprint arXiv:1701.04950*, pp. – , 2017.