On Equivalence of Neural Network Receivers

Maqsood Careem, Aveek Dutta
Department of Electrical and Computer Engineering
University at Albany SUNY, Albany, NY 12222 USA
{mabdulcareem, adutta}@albany.edu

Ngwe Thawdar
US Air Force Research Laboratory
Rome, NY, USA
ngwe.thawdar@us.af.mil

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500703

***Abstract*—Neural Network (NN) based receivers have seen limited adoption in practical systems due to a lack of explainability and performance guarantees, despite their efficacy as a data-driven tool for physical layer signal processing. In order to bridge this gap in explainability, we present an equivalent NN-based receiver that performs the same optimizations used by classical receivers for symbol detection. Achieving equivalence is crucial to explaining how a NN-based receiver classifies symbols in high-dimensional channels and determining its structure that is robust to the underlying channel with minimum training. We realize this by deriving the risk function that guarantees equivalence, which also provides a measure of the disparity between NN-based and classical receivers. Consequently, this information allows us to derive mathematically tight data-dependent bounds on the bit error rate of NN-based receivers, and empirically determine its structure that achieves minimum error rate. Extensive simulation results show the efficacy of the derived bounds and structure of NN-based receivers for single and multi-antenna systems over a variety of channels.**

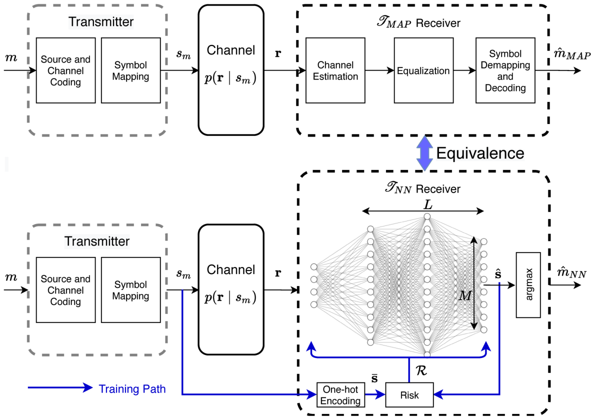


Fig. 1: Equivalence between $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ receivers

I. Introduction

Deep Learning (DL) has been shown to be a versatile tool for physical layer communication due to its proficiency in systems with model and algorithm deficit [1], and has been used to replace individual blocks in communication systems for optimized performance for a given dataset. However, much of the effort have focused on training neural networks as black-box systems that lack guarantees on their performance and model complexity, and have at best approaching the error rate of classical (non-DL) methods [2], [3]. These often require prohibitive amount of training and computational resources [4], [5] that limit their practical adoption. The stringent requirements on reliability and throughput in Next Generation (xG) propagation environments demand wireless transceivers to achieve provable performance guarantees over a variety of channels. Therefore, *explaining* how an NN-based receiver detects symbols (generally a multidimensional classification problem) is paramount to achieve such guarantees [6] while providing insights on amount of required training and robustness to different signal configurations and channels. In order to achieve this goal, two things should be mathematically and semantically understood: 1) A measure of the knowledge acquired by the NN-based receiver and 2) Bounds on learnability (or error) of the neural network model used for signal processing. Due to the difficulty of explaining the behaviour of neural networks in high-dimensional stochastic systems like wireless channels [1], [7], we have to define an equivalent NN-based receiver that eventually performs the same optimization as a classical symbol detector, which is optimal for stationary channel distributions with a known mathematical model [8]. The key to establishing this *equivalence*, shown in figure 1, is deriving the risk function and other structural criteria of the NN-based receiver $\mathscr{T}_{NN}$ that eventually leads to the error performance of classical symbol detector $\mathscr{T}_{MAP}$. This serves as the much-needed theoretical evidence on the similarity in the performance of NN-based and classical receivers that can be the stepping stone towards corroborating the efficacy and limitations of NN-based transceivers in different channel conditions. Such an equivalent $\mathscr{T}_{NN}$ receiver is not only robust to data that is not seen during training but also reduces the amount of training and leads to performance guarantees even when trained on channels with unknown distribution.

The risk function of the $\mathscr{T}_{NN}$ derived for equivalence also serves as a measure of the disparity between NN-based and classical receivers, that is central to deriving mathematically tight data-dependent bounds on its error rate. Unlike previous literature, this provides a means to evaluate the best and worst case performance of such receivers before field testing or practical implementation. Moreover, this information allows us to empirically determine the structure of $\mathscr{T}_{NN}$ that approach the error rate of $\mathscr{T}_{MAP}$, even under incomplete training. This allows for the implementation of practical receivers using data-driven, low-complexity NN structures that is both computationally efficient and robust to the underlying channel. Therefore, this work enables implementation of tractable NN-based receivers with guaranteed error rate and model complexity, which is critical to meet the reliability and throughput requirements in existing and xG networks.

978-1-7281-7122-7/21/$31.00 ©2021 IEEE

II. Related Work

Neural Network based Transceivers: Despite the increasing interest of DL to replace classical communication systems, these systems have come at the cost of increased training time and model complexity and have at best approached the error rate of classical methods [2], [3]. So far, much of the effort has been focused on replacing individual blocks [9], [10] or the end-to-end chain [11], [12] in classical transceivers using "black-box" neural networks to achieve a target error rate for a specific dataset. These fixed function implementations that fits a particular dataset cannot be generalized to other scenarios. This has limited the wide-spread practical adoption of neural network based communication systems, due to their lack of explainability [6] and performance guarantees on the error rate or the complexity (training or hardware complexity). In order to make wireless communications explainable, we take the first step towards understanding what happens inside a NN-based receiver that eventually performs equivalent to a classical receiver. We use this knowledge to derive tight bounds on the BER of NN-based receivers that achieve strict communication guarantees essential for practical adaptation.

Explainable Neural Networks: Explainability of NN-based transceivers is rare in the literature, attributed primarily due to their black-box implementation. Of particular relevance here, is the literature that instead rely on model-driven techniques to avoid the black box approach, and incorporate expert knowledge into autoencoder-based communication systems with the radio transformer network [13] and the OFDM-autoencoder [14]. The latter combines an autoencoder with a classical OFDM system to inherit its advantages including robustness to synchronization errors and multipath equalization. The model-driven OFDM receiver in [15] combines DL with expert knowledge and implements two modules for channel estimation and symbol detection similar to classical systems. However these approaches are limited by the flexibility and accuracy of the underlying model and the trained weights, which limits their applicability to specific channel conditions. In contrast, we achieve guarantees on the performance of data-driven NN-based receivers by explaining their internal mechanisms and deriving their structure that achieves equivalence to MAP receiver. This allows for low-complexity NN-based receivers that are robust to various channels.

III. Models and Preliminaries

The basic communications system consists of a transmitter, a channel, and a receiver. At the transmitter, the uncoded messages, $m{\in}[1,M]$ are modulated and the symbols, s_m are transmitted over the channel, where M is the order of the modulation scheme. The stochastic behavior of the wireless channel is represented by the conditional likelihood of receiving a random vector $\boldsymbol{r}$ when symbol s_m is transmitted, i.e., $p(\boldsymbol{r}|s_m)$ [8]. In general, $\boldsymbol{r}{\in}\mathbb{R}^{2K}$ consists of real and imaginary streams for each antenna $k{\in}[1,K]$, where K is the number of antennas at the receiver. Upon reception of $\boldsymbol{r}$, $\mathscr{T}_{MAP}$ and $\mathscr{T}_{NN}$ apply the transformations f_{MAP} and f_{NN} to produce the estimates $\hat{m}_{MAP}$ and $\hat{m}_{NN}$ of the transmitted message. The benefits of coding are complementary to this work, and yield an added gain in the error rate using existing coding schemes.

Classical Receiver: An optimal $\mathscr{T}_{MAP}$ receiver using conventional classical theory strives to achieve minimum probability of error, also referred to as the *Bayesian Risk* given by $P_e{=}\mathrm{P}[\hat{m}{\neq}m]$, by maximizing the posterior probability (MAP) to estimate the most likely transmitted symbol or the likelihood under the assumption of equal prior. The estimated message $\hat{m}_{MAP}$ of $\mathscr{T}_{MAP}$ is given by,

$$\hat{m}_{MAP}{=}\arg\max_{m} f_{MAP}(\boldsymbol{r});\ \ f_{MAP}(\boldsymbol{r}){:=}\mathrm{P}[s_m|\boldsymbol{r}], \qquad (1)$$

where $f_{MAP}(\boldsymbol{r})$ denotes the mapping from the input $\boldsymbol{r}$ to the output of $\mathscr{T}_{MAP}$ and is equal to the posterior probability of each symbol, s_m. While the low-complexity nearest neighbour symbol detection algorithm is provably optimal for Additive white Gaussian noise (AWGN) channels, there is a lack of closed-form expressions for $f_{MAP}(\boldsymbol{r})$ for most practical channels (*e.g.*, with unknown or non-Gaussian distributions) [8]. Moreover, the complexity of $\mathscr{T}_{MAP}$ grows exponentially with K and constellation size [16], and consequently $\mathscr{T}_{MAP}$ is often computationally prohibitive for practical realization and requires mathematically accurate channel models for optimal realization. Therefore, practical implementations of $\mathscr{T}_{MAP}$ relies on a modular structure with additional signal processing blocks, (*e.g.*, Channel estimation and Equalization) as shown in figure 1, to whiten the noise by estimating and mitigating distortions due to fading and interference, to employ the nearest neighbour decoder. However, we show that there exists a $\mathscr{T}_{NN}$ that is equivalent to $\mathscr{T}_{MAP}$, which is determined by learning a deterministic mapping function $f_{NN}(\boldsymbol{r})$ and hence is computationally tractable.

Neural Network based Receiver: The $\mathscr{T}_{NN}$ is derived by training a NN model to determine its parameters, δ_n^* that minimizes an *Empirical risk*, $\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))$ for a given training dataset $\mathcal{D}$ by using (2),

$$\delta_n^*{=}\arg\min_{\delta_n}\left[\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))\right]\ \text{ where,} \qquad (2)$$

$$\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n)){=}\frac{1}{n}\sum_{n}[\mathcal{L}(f_{NN}(\boldsymbol{r};\delta_n))]$$

where $\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))$ is the average value of a loss function, $\mathcal{L}(f_{NN}(\boldsymbol{r};\delta_n))$, f_{NN} is the mapping from the inputs $\boldsymbol{r}$ to the outputs of the $\mathscr{T}_{NN}$, and $n{=}|\mathcal{D}|$ is the number of training samples. $\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))$ is an empirical measure of the error in the estimated and true symbols of the $\mathscr{T}_{NN}$ such as the mean absolute error, mean-squared error (MSE) or the cross-entropy loss [17]. Gradient descent and the backpropagation algorithm [17] is used to determine the parameters, δ_n^* that minimizes (2) among all possible parameters with n training samples, δ_n.

For a fully-connected feed-forward deep neural network of depth L, f_{NN} is a composition of L functions, $f_1,\ldots,f_L$, that describe the transitions between neurons via $L{-}1$ intermediate hidden layers. Each function f_l is determined by the parameters $\delta_{n,l}{=}\{\boldsymbol{W}_l,\boldsymbol{b}_l\}$ and modeled as

$f_l\left(\boldsymbol{u}_{l-1};\delta_l\right)=\phi_l\left(\boldsymbol{W}_l\boldsymbol{u}_{l-1}+\boldsymbol{b}_l\right)$, where $\boldsymbol{W}_l$ is the weight matrix, $\boldsymbol{b}_l$ is the bias vector, and ϕ_l is a non-linear function referred to as an activation function. $\boldsymbol{u}_{l-1}$ is the input to the l^{th} layer and $\boldsymbol{u}_0=\boldsymbol{r}$. Therefore, f_{NN} is given by (3),

$$f_{NN}(\boldsymbol{r};\delta_n)=f_L\left(\ldots\left(f_2\left(f_1\left(\boldsymbol{r};\delta_{n,1}\right);\delta_{n,2}\right)\ldots\right);\delta_{n,L}\right) \quad (3)$$

where $\delta_n=\{\delta_{n,l}\}$, for $l\in[1,L]$. To understand the internal mechanism of $\mathscr{T}_{NN}$ and compare its performance to $\mathscr{T}_{MAP}$ over stochastic wireless channels, we define the concept of *probabilistic equivalence* between $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$.

Definition 1. *(Equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$). A $\mathscr{T}_{NN}$ with associated function f_{NN} and $\mathscr{T}_{MAP}$ with associated function f_{MAP}, are equivalent if and only if,*

$$\begin{aligned}&f_{NN}(\boldsymbol{r};\delta_n)\xrightarrow{P}f_{MAP}(\boldsymbol{r})\ \textit{or,}\\&\lim_{n\to\infty}\mathrm{P}\left[\left|f_{NN}(\boldsymbol{r};\delta_n)-f_{MAP}(\boldsymbol{r})\right|\right]\to 0\end{aligned} \quad (4)$$

where $\xrightarrow{P}$ refers to pointwise convergence in probability [17]. Therefore, a $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ are defined as equivalent if their outputs converge in probability. *Due to space constraint, the complete proofs of Theorems and Lemmas are provided in an external document in [18].*

A. Impact of Training on Equivalence

Establishing equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ is not a straightforward task because the output of conventional neural networks are deterministic estimates and not Bayesian probabilities, and the difficulty lies in bridging the gap between empirical approaches (*e.g.*, data-driven $\mathscr{T}_{NN}$) and Bayesian approaches (*e.g.*, model-driven $\mathscr{T}_{MAP}$). The disparity of these techniques are investigated under complete and incomplete training of the $\mathscr{T}_{NN}$.

1) Complete Training: This represents the scenario where the $\mathscr{T}_{NN}$ is trained on a sufficiently rich dataset that is representative of the entire channel distribution. Due to the difficulty in establishing effective measures of the quality of training data that is robust to various channel models, complete training is guaranteed asymptotically as $n\to\infty$. Therefore, under complete training criteria (2) is represented by (5),

$$\begin{aligned}\delta^*&=\lim_{n\to\infty}\arg\min_{\delta_n}[\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))]\text{ where,}\\\lim_{n\to\infty}\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))&=\mathbb{E}[\mathcal{L}(f_{NN}(\boldsymbol{r};\delta_n))]\end{aligned} \quad (5)$$

(5) is supported by the law of large numbers, where the average over the data in (2) is replaced by the statistical expectation, $\mathbb{E}[.]$ over the underlying channel distribution. Furthermore, from (5) due to the capacity of a feed-forward NN model with sufficient number of neurons to approximate any nonlinear function to a desired level of accuracy under sufficiently large n [19], [17], we can conclude (6),

$$\lim_{n\to\infty}\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n))\xrightarrow{P}\mathcal{R}(f_{NN}(\boldsymbol{r}))\implies\lim_{n\to\infty}f_{NN}(\boldsymbol{r};\delta_n)\xrightarrow{P}f_{NN}(\boldsymbol{r}) \quad (6)$$

Therefore to guarantee equivalence as in (4), it suffices to show that f_{NN} and f_{MAP} converge over the same input $\boldsymbol{r}$[1], i.e., $f_{NN}(\boldsymbol{r})\xrightarrow{P}f_{MAP}(\boldsymbol{r})$. In practice, n is finite, therefore complete training is approximately guaranteed by *empirically* determining a value n_{emp}, such that if $n\geqslant n_{emp}$ the *Structural Risk* of the $\mathscr{T}_{NN}$ (defined in Section V) is approximately zero, as explained further in Section VI.

2) Incomplete Training: This represents the scenario, where the $\mathscr{T}_{NN}$ is trained on a finite dataset that is not representative of the entire channel distribution. Under incomplete training, where n is finite (in practice, $n<n_{emp}$), it is difficult to bridge the gap between Empirical and Bayesian approaches [17]. Therefore, we derive the expected risk of the $\mathscr{T}_{NN}$, and use this to quantify the disparity between $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$.

IV. Equivalence

The $\mathscr{T}_{NN}$ unlike $\mathscr{T}_{MAP}$, *learns* to classify the symbols from the training data, without making any assumptions on the type of channel or noise. Compared to the exhaustive search employed by $\mathscr{T}_{MAP}$, the $\mathscr{T}_{NN}$ estimates the message, $\hat{m}_{NN}$ by deterministic processing of the received signal (via L layers), and thereby enables low-latency implementations of the receiver. Therefore, a single feed-forward neural network typically suffices to implement the communication receiver, regardless of the underlying channel distribution. Since $\mathscr{T}_{MAP}$ strives to maximize the posterior probability, i.e., $f_{MAP}(\boldsymbol{r}):=\mathrm{P}\left[s_m|\boldsymbol{r}\right]$ to estimate $\hat{m}_{MAP}$ as in (1), to achieve equivalence as per (4), we seek to derive the risk function of $\mathscr{T}_{NN}$, $\mathcal{R}(f_{NN})$ and other design criteria that guarantee that after training, the mapping function is precisely the posterior probabilities, i.e., $f_{NN}(\boldsymbol{r})\xrightarrow{P}\mathrm{P}\left[s_m|\boldsymbol{r}\right]$.

Lemma 1. *A $\mathscr{T}_{NN}$ with empirical risk, $\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n),\bar{\boldsymbol{s}})$ trained on labels, $\bar{\boldsymbol{s}}$ is equivalent to $\mathscr{T}_{MAP}$ under complete training and the three design criteria in* (7b),

$$\mathcal{R}(f_{NN}(\boldsymbol{r};\delta_n),\bar{\boldsymbol{s}})=\frac{1}{n}\sum_n\left\{\|f_{NN}(\boldsymbol{r};\delta_n)-\bar{\boldsymbol{s}}\|^2\right\} \quad (7a)$$

$$1)\ \dim(\hat{\boldsymbol{s}})=M\times 1,\ \ 2)\sum_{m=1}^{M}\hat{s}_m=1,\ \ 3)\ \bar{s}_{m'}=\begin{cases}1; m'=m\\0; \textit{else}\end{cases} \quad (7b)$$

where $\hat{s}_m=f_{NN}(\boldsymbol{r};\delta_n)$ is the symbol estimate by $\mathscr{T}_{NN}$, $\hat{\boldsymbol{s}}$ is the vector of $\hat{s}_m$ for all $m\in M$, and $\bar{s}_m$ is the binary label corresponding to s_m.

The proof is provided in Appendix A.A in the external document in [18]. The empirical risk in (7a) that guarantees equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ is the MSE risk function[2]. Therefore, the estimated message from $\mathscr{T}_{NN}$ is obtained by employing an *argmax* function at its output as in figure 1,

$$\hat{m}_{NN}=\arg\max_m f_{NN}(\boldsymbol{r})\xrightarrow{P}\arg\max_m\mathrm{P}[s_m|\boldsymbol{r}]=\hat{m}_{MAP} \quad (8)$$

Design Criteria: The design criteria in (7b) ensures that the derived $\mathscr{T}_{NN}$ estimates the posterior probabilities. (7b.1)

[1] Under the above conditions for complete training, the output of the $\mathscr{T}_{NN}$, $f_{NN}(\boldsymbol{r};\delta_n)$ converges to the output of a general function $f_{NN}(\boldsymbol{r})$ as in [20].

[2] Provided that the design criteria in (7b) are satisfied, it is straightforward to show the existence of other empirical risk functions (*e.g.*, Categorical Cross Entropy loss [21]) that achieve equivalence similar to the MSE risk.

ensures that the output layer of $\mathscr{T}_{NN}$ has M neurons as shown in figure 1, with each neuron corresponding to a posterior of a particular symbol, s_m, for $m{\in}[1,M]$, (7b.2) ensures that the sum of the values of the output neurons is equal to 1, to ensure a total probability of 1, where $\hat{s}_m$ is the value of the m^{th} output neuron. In practice, this criteria is guaranteed by using a softmax activation function at the output layer of $\mathscr{T}_{NN}$ given by, $\phi_L{=}\frac{\exp\{\hat{s}_m\}}{\sum_m \exp\{\hat{s}_m\}}$, and (7b.3) ensures that the $\mathscr{T}_{NN}$ is trained with a set of labels, $\bar{\mathbf{s}}{\in}[0,1]^M$, which is a one-hot encoded vector representation of the transmitted symbols, s_m, where the m^{th} element, $\bar{s}_m$ is equal to one and zero otherwise. Under these criteria, it is straightforward to show that the posterior probabilities of each symbol s_m is exactly equal to the empirical posterior mean of the one hot encoded symbols $\bar{s}_m$, i.e., $\mathrm{P}\left[s_m|\boldsymbol{r}\right]=\mathbb{E}[\bar{s}_m|\boldsymbol{r}]$. Therefore, to guarantee equivalence as in (4) it suffices to derive the risk function of the $\mathscr{T}_{NN}$ that computes the *empirical posterior mean.* Estimating the empirical mean is tractable using a neural network under complete training, since it can be estimated from the training data itself, and does not require knowledge of the channel distribution.

While, under incomplete training the equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ cannot be guaranteed, the performance of $\mathscr{T}_{NN}$ can be estimated in terms of its expected risk. This provides a measure of the disparity of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$, which leads to the following relationship on their error rates.

Lemma 2. *Under incomplete training, minimizing the empirical risk of $\mathscr{T}_{NN}$ in* (7a) *minimizes its BER. However, for known stationary channel distributions, $\mathscr{T}_{NN}$ cannot outperform $\mathscr{T}_{MAP}$ in terms of the average BER.*

$$BER_{NN} \geqslant BER_{MAP} \tag{9}$$

where BER_{NN} and BER_{MAP} are the average BER of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ over the entire channel distribution. The equality holds under equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$.

The proof is provided in Appendix A.B in the external document in [18].

V. BER Bounds for $\mathscr{T}_{NN}$

The *Expected Risk* of the $\mathscr{T}_{NN}$, $\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n))$ measures the error between the estimated symbols, $\hat{s}_m$ and the transmitted symbols, s_m over all possible recived symbols, $\boldsymbol{r}$ and hence is a measure of the average BER. Hence, $\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n))$ is a measure of the robustness of the $\mathscr{T}_{NN}$ to generalize to data that is not seen during training. Therefore, to derive tight data-dependent bound on the BER of $\mathscr{T}_{NN}$ for stationary channel distributions, we estimate the expected risk of $\mathscr{T}_{NN}$ given by,

$$\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n)) = \mathbb{E}_{\mathcal{D}}\left[\mathcal{L}(f_{NN}(\boldsymbol{r};\delta_n))\right] \tag{10}$$

where $\mathbb{E}_{\mathcal{D}}[.]$ is the expectation over the ensemble of possible datasets, $\mathcal{D}$ (for fixed sample size n). The difficulty here is that $\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n))$ cannot be computed during training and is a random variable due to the randomness in parameter initialization and choice of model-complexity [22] for $\mathscr{T}_{NN}$. Therefore, we derive expressions for the expected risk in terms of f_{NN} and f_{MAP} and use this to derive the lower and upper bounds on the BER, under complete and incomplete training.

Lemma 3. *Under complete training, $n{\to}\infty$ the expected risk of $\mathscr{T}_{NN}$ is distributed as a zero mean Gaussian as in* (11)*,*

$$\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta^*)){\sim}\mathcal{N}(0,\sigma_h^2),\ \ \sigma_h^2{=}\mathbb{E}\left[\sum_{m=1}^{M}\mathrm{var}\left\{\bar{s}_m|\boldsymbol{r}\right\}\right] \tag{11}$$

The proof is provided in Appendix B.A in the external document in [18]. Therefore, for any $\boldsymbol{r}$ during testing, we observe that the expected risk is unbiased, and has a variance, σ_h^2 that is independent of the parameters of $\mathscr{T}_{NN}$, δ^*. σ_h^2 only depends on the received vector, $\boldsymbol{r}$ and the labels, $\bar{\mathbf{s}}$, and is determined by the statistics of the channel. In Proposition 1 we further show that for an AWGN channel, the term σ_h^2 is determined only by the Signal-to-Noise ratio (SNR) and the constellation order M. Therefore, σ_h^2 is an irreducible term that appears due to the physical channel, and cannot be alleviated by further training or optimization of the $\mathscr{T}_{NN}$.

Proposition 1. *For an AWGN channel with SNR per bit, γ_b and for a decoded modulation scheme of order M, the term σ_h^2 in* (11) *is given by* (12)*,*

$$\sigma_h^2 = \frac{M\lambda A_M}{\gamma_b \log_2 M} \quad \textit{where,} \quad \lambda{=}\left(1{-}\frac{2}{\pi}\right), \quad \textit{and,} \tag{12}$$

$$\text{PSK: } A_M{=}\frac{1}{2},\ \text{PAM: } A_M{=}\frac{(M^2{-}1)}{6},\ \text{QAM: } A_M{=}\frac{(M{-}1)}{3}$$

The proof is provided in Appendix B.B in the external document in [18]. Alternately, even when a mathematical model does not exist for the channel, σ_h^2 can be calculated directly from the dataset using (11).

Lemma 4. *Under incomplete training (finite n), the expected risk associated of $\mathscr{T}_{NN}$ is distributed as a Gaussian as in,*

$$\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n^*)) \sim \mathcal{N}(\mu_n, \sigma_n^2 + \sigma_h^2) \quad \textit{where,} \tag{13a}$$

$$\mu_n = \mathbb{E}\left[\left(\bar{f}_{NN}(\boldsymbol{r}){-}f_{MAP}(\boldsymbol{r})\right)^2\right], \tag{13b}$$

$$\sigma_n^2 = \mathbb{E}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{NN}(\boldsymbol{r};\delta_n) - \bar{f}_{NN}(\boldsymbol{r})\right)^2\right]\right] \tag{13c}$$

where $\bar{f}_{NN}(\boldsymbol{r}){=}\mathbb{E}_{\mathcal{D}}[f_{NN}(\boldsymbol{r};\delta_n)]$ is the average function associated with $\mathscr{T}_{NN}$ obtained by averaging over all possible functions $f_{NN}(\boldsymbol{r})$ when $\mathscr{T}_{NN}$ is trained with n data samples.

The proof is provided in Appendix B.C in the external document in [18]. Under incomplete training, the expected risk in (13) has a bias, μ_n and a variance, $\sigma_n^2{+}\sigma_h^2$ that depends on the parameters of $\mathscr{T}_{NN}$, the number of training samples, n and the channel. The exact values of μ_n and σ_n^2 in (13) are determined by: a) using a subset of the input dataset as a validation set, and b) extracting the statistics (mean and variance) of the estimated distribution of the expected risk as per (13) over the validation set [17].

Corollary 1. Expected risk in (13) can be decomposed as,

$$\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n^*)){=}\mathcal{F}_n(N){+}\mathcal{N}(0,\sigma_h^2),\ \mathcal{F}_n(N){=}\mathcal{N}(\mu_n,\sigma_n^2) \tag{14}$$

where $\mathcal{F}_n(N)$ is referred to as the Structural Risk, and $N{=}|\delta_n^*|$ is the number of parameters of $\mathscr{T}_{NN}$ and is equal to the total number of weights and biases in $\mathscr{T}_{NN}$.

The proof is provided in Appendix B.D in the external document in [18]. Therefore, the expected risk in (13) has two components: a) the structural risk that depends on the amount of training, n and the number of parameters of $\mathscr{T}_{NN}$, and b) the term $\mathcal{N}(0,\sigma_h^2)$ that depends only on the channel. Moreover, we observe that when $\mathcal{F}_n(N){=}0$, (14) converges to (11). This is true when $n{\to}\infty$, since $\lim_{n\to\infty}\mu_n{=}0$ and $\lim_{n\to\infty}\sigma_n^2{=}0$, according to the law of large numbers and the capacity of NNs to serve as universal function approximators [19]. In practice, this is also approximately achieved by empirically finding a value n_{emp}, such that $\mathcal{F}_n(N){\approx}0$ for $n{\geqslant}n_{emp}$. The impact of the structure of $\mathscr{T}_{NN}$ and amount of training on the structural risk is studied in Section VI, which is used to empirically determine the value of N that achieves minimum expected risk for a given n.

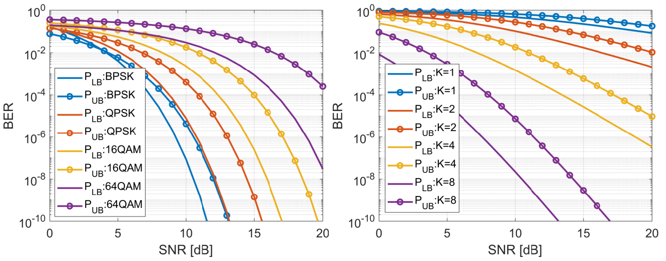
Theorem 1. *The BER of the SISO $\mathscr{T}_{NN}$, BER_{NN} trained with n samples with an expected risk of $\bar{\mathcal{R}}(f_{NN}(\boldsymbol{r};\delta_n^*)){\sim}\mathcal{N}(\mu_n,\sigma_n^2{+}\sigma_h^2)$, for an arbitrary modulation scheme of order M is bounded as in* (15)*,*

$$P_{LB} \leqslant BER_{NN} \leqslant P_{UB} \tag{15}$$

where P_{LB} is the equivalent BER of $\mathscr{T}_{MAP}$, BER_{MAP} for known stationary channel distributions, i.e., $P_{LB}{=}BER_{MAP}$. P_{UB} as derived in (16)*,*

$$P_{UB}=\left(\frac{1}{2}\right)e^{-\left(\frac{M(1-\mu_n)-1}{\sqrt{2M^2(\sigma_n^2+\sigma_h^2)}}\right)^2} \tag{16}$$

The bounds on the BER of $\mathscr{T}_{NN}$ is derived by using the expressions for the expected risk as in (11) and (13) under complete and incomplete training as provided in Appendix B.E in the external document in [18]. Since, the BER for $\mathscr{T}_{NN}$ is lower bounded as $P_{LB}{=}BER_{MAP}$, a $\mathscr{T}_{NN}$ achieves the minimum *average* BER for any $\boldsymbol{r}$, when it is equivalent to $\mathscr{T}_{MAP}$. Therefore, closed form expressions for BER_{MAP} and hence P_{LB} exist for known channel models such as AWGN and Rayleigh channels [8]. Under complete training, equivalence is guaranteed as per the design criteria and risk function of $\mathscr{T}_{NN}$ in Lemma 1, and consequently its BER only depends on the statistics of the channel. Under incomplete training (finite n), the BER of the $\mathscr{T}_{NN}$ is upper bounded by P_{UB}. Therefore, P_{UB} similar to the expected risk, is also dependent on the parameters of $\mathscr{T}_{NN}$ and the statistics of the channel. Moreover, the gap between P_{LB} and P_{UB} is a measure of the deviation of BER_{NN} from BER_{MAP} under incomplete training. Figure 2a shows the lower and upper bounds on the BER of $\mathscr{T}_{NN}$ as per (15) for AWGN channels, under various modulation schemes, for $n{=}10^3$. We observe that the gap between P_{LB} and P_{UB} increases with M, due to the lower noise margins for higher order constellations. For example, to achieve a target BER of 10^{-2}, $\mathscr{T}_{NN}$ would require at most 2.8dB and 4dB more SNR than the equivalent $\mathscr{T}_{MAP}$ for 16-QAM and 64-QAM signals respectively.



(a) P_{LB} and P_{UB} for SISO (b) P_{LB} and P_{UB} for MIMO

Fig. 2: Bounds on the BER of $\mathscr{T}_{NN}$ with $n{=}10^3$, σ_h^2 from (12), and μ_n, σ_n^2 determined from the validation set: a) for a SISO AWGN channel for all modulation schemes, and b) for a MIMO fading channel for 64-QAM modulation scheme.

For multi-antenna systems (*e.g.*, multiple-input, multiple-output (MIMO)), Maximum Ratio Combining (MRC) provides an optimal combining scheme that minimizes the error rate by exploiting the spatial diversity of the channel, in the absence of interference. Corollary 2 extends the bounds for BER_{NN} in (15) to MIMO systems where MRC is employed at the $\mathscr{T}_{NN}$ to combine the streams from multiple antennas.

Corollary 2. *Given the average SNR per channel, $\bar{\gamma}_k$, the upper bound on the BER of the $\mathscr{T}_{NN}$ with MRC combiner, for the MIMO Rayleigh fading channel is given by,*

$$P_{UB}{=}\frac{M{-}1}{M}\prod_{k=1}^{K}\frac{1}{(1{+}g_{NN}\bar{\gamma}_k)},\ g_{NN}{=}\frac{M(1{-}\mu_n){-}1}{M^2A_M} \tag{17}$$

The proof is provided in Appendix B.F in external document in [18]. Figure 2b shows the lower and upper bounds on the BER of $\mathscr{T}_{NN}$ as per (15) for MIMO fading channels, for 16-QAM modulated symbols. The gap between P_{LB} and P_{UB} increases with K, since the disparity between $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$ is aggregated when multiple streams are combined. For example, to achieve a target BER of 10^{-2}, the $\mathscr{T}_{NN}$ would require at most 3dB and 4dB more SNR than the equivalent $\mathscr{T}_{MAP}$ for $K{=}4$ and $K{=}8$ respectively.

VI. Evaluation and Results

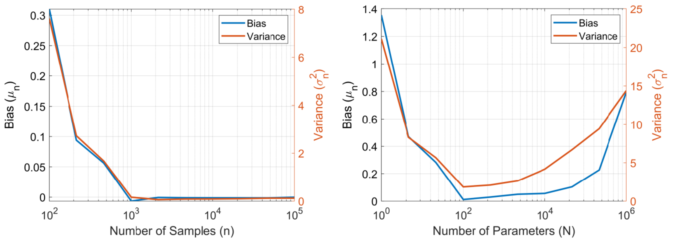
We analyze the impact of training and the model structure of $\mathscr{T}_{NN}$ on the equivalence with $\mathscr{T}_{MAP}$ using a practical simulation framework. The simulation parameters are detailed in table I. Experiments are conducted on 100 sets of training samples of size n drawn randomly from different channel distributions including AWGN and Rayleigh channels. The model structure of $\mathscr{T}_{NN}$ is parameterized by its number of trained parameters, $N{=}|\delta_n^*|$. In general, N is a measure of the depth and width of $\mathscr{T}_{NN}$ and determines its capacity to generalize beyond the training data [17]. 100 different neural networks for each choice of n and N are trained using different training sets and initializations of the $\mathscr{T}_{NN}$ parameters. μ_n and σ_n^2 are estimated as described in Section V, where the population expectation $\mathbb{E}_{\mathcal{D}}$ is estimated over the validation

TABLE I: Simulation Parameters

Parameters/Hyperparameters	Value/Model
Neural Network Type	Deep Feed-Forward NN
Empirical Risk ($\mathcal{R}$)	Mean-Squared Error (MSE)
Training Algorithm	Scaled Conjugate Gradient Descent(SCGD)
Activation Function (ϕ_l)	Hidden:Tan-Sig [24], Output:Softmax
Channel Models	AWGN, Rayleigh
Antenna Configuration	SISO, MIMO
Number of Hidden Layers	[3, 5]
Number of Input Neurons	2 (I,Q streams)
Number of Output Neurons	$M=[2,4,16,64]$
Number of Samples (n)	10^2 - 10^5
Number of Parameters (N)	10^0 - 10^5
Training,Validation,Testing	[60, 20, 20] %

set. $\mathscr{T}_{NN}$ is trained using SCGD [23] and its parameters, are initialized by randomly sampling from a distribution whose variance is inversely proportional to N, i.e, $\delta_0 \sim \mathcal{N}\left(0, \frac{1}{N}\mathbf{I}\right)$.

A. Model Parameters of $\mathscr{T}_{NN}$ for Equivalence

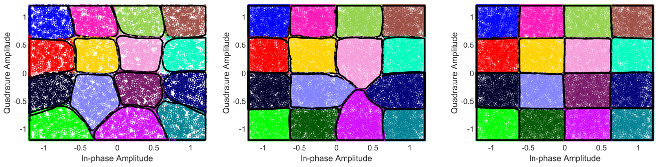


(a) With amount of training n (b) With number of parameters N

Fig. 3: Bias and Variance of the expected risk of $\mathscr{T}_{NN}$.

Designing computationally efficient a $\mathscr{T}_{NN}$, require determining the amount of training and its structure that guarantee minimum BER. Figure 3 shows the variation of the bias and variance of $\mathscr{T}_{NN}$ with the number of training samples, n and the number of parameters, N for received data drawn from AWGN channels averaged over SNR from 0-20dB. Figure 3a shows a decreasing trend in the average values of μ_n and σ_n^2 with increasing n, over all possible N. This indicates that the expected risk and consequently the BER decreases as more training samples are incorporated, However, we also observe that there is negligible improvement in μ_n and σ_n^2 for $n \geqslant 10^3$ samples. Therefore, we set $n_{emp}=10^3$ for which the structural risk is approximately zero, and consequently the complete training criteria is approximately achieved. Figure 3b shows a convex trend of μ_n and σ_n^2 with increasing N for a fixed $n=10^3$, with a minimum when $N=10^2$. It is clear that under-parameterized ($N \ll 10^2$) and over-parameterized ($N \gg 10^2$) $\mathscr{T}_{NN}$ result in large μ_n and σ_n^2 and Therefore, we observe that for a given n, there exists a specific structure of $\mathscr{T}_{NN}$ with $N=10^2$, for which the structural risk and consequently the BER is minimum.

Figure 4 shows the impact of the structure of $\mathscr{T}_{NN}$ (i.e., the choice of N) on the learned decision boundaries for an AWGN channel for 16-QAM constellation when only limited training samples are available, $n=10^3$. For AWGN channel, the optimum decision boundaries of $\mathscr{T}_{MAP}$ are the perpendicular bisectors of the lines connecting the constellations, which guarantees minimum error rate [8]. Therefore the learned decision boundaries by $\mathscr{T}_{NN}$, is a visual metric to explain its behaviour, and gauge its ability to achieve equivalence and generalize beyond the training data. When $\mathscr{T}_{NN}$ is under-parameterized ($N \ll 10^2$), the decision boundaries are under-fitted and are not the optimal boundaries. Therefore, more training samples are required to learn the optimum boundaries. When $\mathscr{T}_{NN}$ is over-parameterized ($N \gg 10^2$), the decision boundaries are overfitted to the specific data samples on which it is trained and the noise. This is because the $\mathscr{T}_{NN}$ configures the decision boundaries to also accommodate outliers in the training set, and therefore will not generalize to data not seen during training. Therefore, we observe that for a given n, there exists a specific structure of $\mathscr{T}_{NN}$ and choice of N for which the learned decision boundaries are optimum, and consequently result in the minimum BER. We also highlight that under complete training, $n \rightarrow \infty$), $\mathscr{T}_{NN}$ will always learn the optimum decision boundaries, due to its capacity to as universal approximators [19], regardless of the choice of N.



(a) Under fitted boundaries, $N \ll 10^2$ (b) Over fitted boundaries, $N \gg 10^2$ (c) Perfect fitted boundaries, $N=10^2$

Fig. 4: Decision boundaries of the $\mathscr{T}_{NN}$ for AWGN channel for 16-QAM constellation.

B. Impact of Training Schemes on Equivalence

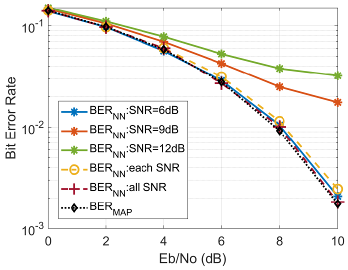
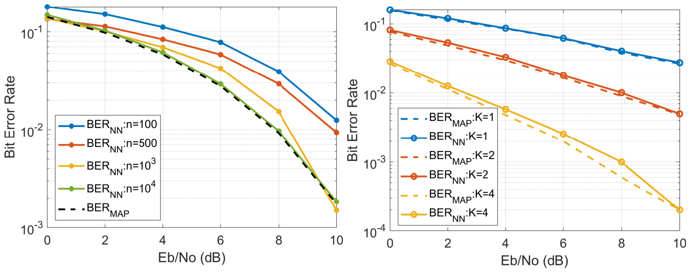


Fig. 5: BER of the $\mathscr{T}_{NN}$ for AWGN channel for 16-QAM modulation scheme with different training schemes.

The impact on the BER by the SNR at which the $\mathscr{T}_{NN}$ is trained is investigated in figure 5. It is clear that under complete training that is representative of the entire channel distribution, the $\mathscr{T}_{NN}$ will learn the optimum decision boundaries. Therefore, we observe that training different $\mathscr{T}_{NN}$ models for each SNR or training a single $\mathscr{T}_{NN}$ model over all SNR eventually learn the optimum decision boundaries and achieve equivalent BER as $\mathscr{T}_{MAP}$ under complete training (*e.g.*, $n=10^4 \gg 10^3$). These results corroborate the significance of training the $\mathscr{T}_{NN}$ using a dataset that is representative of the entire channel distribution. Alternatively, figure 5 also shows that for AWGN channels, provided that the $\mathscr{T}_{NN}$ is trained at a sufficiently low SNR (*e.g.*, SNR=6dB), it is still able to achieve equivalent BER as $\mathscr{T}_{MAP}$, because it learns the optimal decision boundaries. However, when the training SNR is high (SNR$\geqslant$9dB), the decision boundaries are configured for lower-noise margins and is not able to handle the noise at lower SNRs, and consequently results in high BER.

C. Error Performance of the $\mathscr{T}_{NN}$



(a) BER for AWGN channel for 16-QAM modulation scheme.

(b) BER for MIMO Fading channel for BPSK modulation scheme.

Fig. 6: BER comparison of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$: a) for SISO AWGN channel, and b) for MIMO Rayleigh fading channel with MRC employed at the $\mathscr{T}_{NN}$.

Figure 6 shows the BER of the $\mathscr{T}_{NN}$ for SISO and MIMO systems using AWGN and Rayleigh fading channels as representative examples of stationary channels with known distribution. Figure 6a confirms that with more training samples ($n \geqslant 10^3$), the BER of $\mathscr{T}_{NN}$ approaches that of $\mathscr{T}_{MAP}$. Figure 6b shows the performance of the $\mathscr{T}_{NN}$ for multiple-antenna systems when MRC is employed at the receiver to combine the streams from each antenna. It is also clear that the proposed feed-forward neural network based receiver is able to achieve equivalent BER as $\mathscr{T}_{MAP}$ under sufficient training over a variety of channels. This corroborates the theoretical equivalence of $\mathscr{T}_{NN}$ and $\mathscr{T}_{MAP}$, under sufficient training and choice of structure of $\mathscr{T}_{NN}$.

VII. Conclusion

Through rigorous theoretical and empirical analysis we show that a feed-forward deep NN-based receiver with the MSE risk is equivalent to the classical receiver, under complete rich training. We show that under incomplete training, the expected error is distributed as a Gaussian, and derive tight upper bounds on the BER of the NN-based receivers. Extensive practical simulations have corroborated that the BER of such NN-based receivers are within the derived theoretical bounds, and cannot outperform optimal classical receivers when a rigorous mathematical channel model is available. Moreover, we empirically derive the model of the NN-based receiver for which the error rate is minimum with minimum required training. These guarantees on error rate of NN-based receivers is a necessary step towards ensuring wide-spread adoption.

Acknowledgement

This work is funded by Air Force Research Laboratory Visiting Faculty Research Program (SA2020051003-V0205), Rome, NY. Approved for public release AFRL-2021-0278.

References

[1] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.

[2] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," 2018.

[3] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding." in Proc. Information Sciences and Systems (CISS), March 2017.

[4] T. Wang, C. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: opportunities and challenges," *China Communications,vol.*, vol. 14, no. 11, pp. 92–111, 2017.

[5] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[6] W. Guo, "Explainable artificial intelligence for 6g: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[7] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai," 2020.

[8] Proakis, *Digital Communications 5th Edition*. McGraw Hill, 2007.

[9] X. Tan, W. Xu, Y. Be'ery, Z. Zhang, X. You, and C. Zhang, "Improving massive mimo belief propagation detector with deep neural network," 2018, arxiv preprint.

[10] A. Klautau, N. González-Prelcic, A. Mezghani, and R. W. Heath, "Detection and channel equalization with deep learning for low resolution mimo systems," in *Proc. 52nd Asilomar Conference on Signals.* and Computers: Systems, 2018.

[11] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, p. 563–575, Dec 2017. [Online]. Available: http://dx.doi.org/10.1109/TCCN.2017.2758370

[12] F. A. Aoudia and J. Hoydis, "End-to-end learning of communications systems without a channel model," 2018.

[13] T. J. O'Shea, L. Pemula, D. Batra, and T. C. Clancy, "Radio transformer networks: attention models for learning to synchronize in wireless systems," in *Proc. Asilomar Conference on Signals.* Systems and Computers, October 2016.

[14] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. ten Brink, "Ofdm-autoencoder for end-to-end learning of communications systems," 2018.

[15] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "Comnet: Combination of deep learning and expert knowledge in ofdm receivers," 2018.

[16] Y. Yapici, I. Guvenc, and Y. Kakishima, "A map-based layered detection algorithm and outage analysis over mimo channels," *IEEE Transactions on Wireless Communications*, vol. PP, 10 2017.

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[18] M. Careem, A. Dutta, and N. Thawdar, "Proofs on equivalence of neural network receivers." [Online]. Available: https://www.dropbox.com/s/2dts0t9rk0qv57w/ICC_20_Proofs_On_Equivalence.pdf?dl=0

[19] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[20] G. V. Moustakides and K. Basioti, "Training neural networks for likelihood/density ratio estimation," 2019.

[21] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[22] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, "A modern take on the bias-variance tradeoff in neural networks," 2019.

[23] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525 – 533, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608005800565

[24] T. P. Vogl, J. K. Mangis, A. Rigler, W. T. Zink, and D. Alkon, "Accelerating the convergence of the back-propagation method," *Biological Cybernetics*, vol. 59, pp. 257–263, 2004.