Noise Sensitivity and Stability of Deep Neural Networks for Binary Classification

Citation for the original published paper (version of record):

Jonasson, J., Steif, J., Zetterqvist, O. (2023). Noise Sensitivity and Stability of Deep Neural Networks for Binary Classification. Stochastic Processes and their Applications, 165: 130-167.
http://dx.doi.org/10.1016/j.spa.2023.08.003

(article starts on next page)

Noise Sensitivity and Stability of Deep Neural Networks for Binary Classification

Johan Jonasson [*] Jeffrey E. Steif [†] Olof Zetterqvist [‡]

Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

Abstract

A first step is taken towards understanding often observed non-robustness phenomena of deep neural net (DNN) classifiers. This is done from the perspective of Boolean functions by asking if certain sequences of Boolean functions represented by common DNN models are noise sensitive or noise stable, concepts defined in the Boolean function literature. Due to the natural randomness in DNN models, these concepts are extended to annealed and quenched versions. Here we sort out the relation between these definitions and investigate the properties of two standard DNN architectures, the fully connected and convolutional models, when initiated with Gaussian weights.

Keywords: Boolean functions, Noise stability, Noise sensitivity, Deep neural networks, Feed forward neural networks

1 Introduction

The driving question of this paper is how robust a typical binary neural net classifier is to input noise, i.e. for a typical neural net classifier and a typical input, will tiny changes to that input make the classifier change its mind? When asking this, we take inspiration from phenomena observed for deep neural networks (DNN) used in practice and use that inspiration to give mathematically rigorous answers for some simple DNN models under one (of several possible) reasonable interpretations of the question. It is not a prerequisite for the reader to be familiar with DNNs to find the topic interesting and any Machine Learning lingo will be explained shortly.

DNNs have shown results that range from good to staggering in many different data-driven areas, e.g. for prediction and classification. One of many reasons for this is that with sufficiently large models, neural networks can approximate any function [**?**]. However, there is much to be discovered about these black box models, two concerns being about robustness and optimal model design. Studies have shown that DNNs are vulnerable to various attacks, where adding small noise to inputs can lead to significant differences in the output [**? ?**]. For example, an image that is clearly of a fish which a DNN classifier also strongly believes is a fish can be such that only changing it by a tiny amount of random noise suddenly makes the classifier assign high probability to that it is now a dog. This raises the question of how stable DNN models tend to be under small perturbations such as these. Obviously, any non-trivial classification function must have the property that for some input, only a tiny amount of change leads to a different output. However, how typical is an input ω such that the output $f(\omega)$ changes from tiny amounts of change in ω? In this formulation, one can clearly interpret the word "typical" in many different ways and also consider many different ways of defining what a tiny change is.

[*] jonasson@chalmers.se

[†] steif@chalmers.se

[‡] olofze@chalmers.se

To take some small but rigorous steps towards answers, we will in this paper focus on the setting where the input into the DNN is a vector ω of binary bits: $\omega \in \{-1,1\}^n$, and the output is binary classification, $f(\omega) \in \{-1,1\}$.

This point of view is not new and can be seen in Boolean networks. Here some research has been done for different noise settings, but to the best of our knowledge, it seems that none of it is close to what we propose, and little is strictly rigorous. To mention a few, [?] and [?] work in a Boolean network setting. In terms of DNNs, one is then considering a recurrent neural network model with the exact same weight matrix at each layer. The first of these papers considers small changes in the input, and the second paper considers small changes in the network structure. Both conclude that the final output (i.e. the fixed point) is robust to these changes.

The perhaps most natural way to talk about a "typical input", and the one that we are going to adopt, is to consider an input generated at random from a given probability distribution on $\{-1,1\}^n$ (e.g. if we are considering one of the standard benchmark problems of classifying handwritten digits, we would e.g. consider inputting handwritten digits drawn from a probability distribution reflecting how people actually write digits). Then one asks what the probability is that the input is such that the DNN model changes its classification by changing the input in a tiny way.

We prove results for a fully connected DNN architecture with input noise which are valid for arbitrary probability distributions over the input and the noise (as long as the noise with high probability actually produces a change of at least one input bit). However, since the concepts studied are usually understood to assume uniform input distribution and pure noise, i.e. each input bit changes with some tiny predetermined probability independently over the bits, the presentation will be made under these conditions. However, in Sections 3.1 and 3.2, it becomes apparent that the sensitivity properties for fully connected DNN models hold under the most general conditions possible on the input distribution and the noise as will be commented on there. The later section's results rely however on uniform input and pure noise.

In summary, we intend to analyse robustness of DNNs to noise from the perspective of Boolean functions, i.e. to consider those feed-forward DNNs that represent functions with input in $\{-1,1\}^n$ and output in $\{-1,1\}$ and analyse how sensitive these are to small random noise to random input. Doing this, we find ourselves in, or at least very close to, the setting of the research field of noise sensitivity and noise stability of Boolean functions, concepts introduced in [?]. The standard references nowadays to these concepts are the textbooks [? ?]. We will return with exact definitions and extensions of the concepts shortly, but in short, noise sensitivity means what we already said: a Boolean function f is noise sensitive if for large n, $\omega \in \{-1,1\}^n$ uniformly random and η that differs from ω by changing a tiny random amount of randomly chosen bits, then $f(\omega)$ and $f(\eta)$ are virtually uncorrelated. One says that f is noise stable if such tiny changes are very unlikely to change the output of f. (Clearly, a rigorous definition must be in terms of asymptotics as $n \to \infty$). The question in focus now becomes

Is a Boolean function represented by a given DNN noise sensitive or noise stable?

To further restrict the setting, the activation function will at all layers be assumed to be the sign function, and the linear transformation at a given node will always be without a bias term. Neither of these restrictions is common in practice, of course. Still, one can at least arguably claim that since the idea of feeding the output of a neuron into an activation function is to decide if that neuron fires or not, the sign function is "the ideal" activation (but, of course, not used in practice because of the difficulty in training).

All in all, precisely and in a way that explains the DNN lingo, each model we consider will be such that there is a given Boolean function h, a given so called depth T and given so called layer sizes $n_1, n_2, \ldots, n_T$. With those given, we are considering Boolean functions f on Boolean input strings $\omega \in \{-1,1\}^n = \{-1,1\}^{n_0}$ that can be expressed by a choice of matrices $W_1, \ldots, W_T$, $W_t \in \mathbf{R}^{n_t \times n_{t-1}}$ for $t = 1, \ldots, T$ and for $\omega = \omega_0 \in \{-1,1\}^n$ taking

$$\omega_t = \operatorname{sign}(W_t \omega_{t-1}),\ t = 1, \ldots, T \tag{1}$$

and

$$f(\omega) = h(\omega_T) \tag{2}$$

where the sign is taken point-wise. The most common h is of the form $\text{sign}(\mathbf{w}\,\omega_T)$ for some row vector $\mathbf{w}$, i.e. a prediction made from standard logistic regression on ω_T.

As already stated and from what is apparent from the restrictions made, i.e. Boolean input, sign activation functions, no bias terms and, in particular, uniform probability measure over inputs and noise, makes the setting fairly far removed from practical settings. Moreover, we will not consider models that have been fit to data in any way other than a loose motivation for one of our model choices when it comes to modelling randomness in the coefficients of the W_t-matrices.

To consider the matrices as random is natural when taking inspiration from DNNs in practice, since when one trains the model to fit with data, i.e. minimise the loss function at hand, one usually starts the optimisation algorithm by taking the initial coefficients to be random, often i.i.d. normal. Also during training, further randomness is often brought into the picture by the use of stochastic gradient descent. In the end of course, the training algorithm converges, but since there are usually many local minima for the loss function, the randomness in the coefficients at the start makes it random which local maximum one converges to. Furthermore, convergence is almost never reached and this is on purpose, since it is common practice to use early stopping, i.e. stop training well before convergence, to avoid overfitting. As already declared though, we will not consider any data and model fitting, only observe that any training of course produces correlation between the random weight matrix components and suggest and analyse a tractable model of such correlation.

Hence, in summary, we regard this paper as mostly a contribution to the field of noise sensitivity/stability of Boolean functions inspired by an interesting and important phenomenon of DNN prediction models, rather than to applied machine learning. Nevertheless, it is a first step towards an understanding of the non-robustness phenomena of DNNs and, to our knowledge, the first strictly rigorous contribution.

Observe that when considering the W_t's as random, the precise predictor function f that comes out of it is in itself a random object. This is not the case in the field of noise sensitivity/stability and one can now ask two different things: (i) will the predictor be noise sensitive when taking both the randomness in the predictor itself and the randomness in the input and the noise into account?, (ii) will the random predictor after the weights have been drawn, with high probability become noise sensitive in the usual sense? This leads us to extend the standard definitions of noise sensitivity and noise stability to also encompass these aspects.

The paper is structured as follows. Section 2 focuses on the relevant concepts and states the relations between them. The remaining sections each focus on selected examples of models of the family of DNN architectures given by (1) and (2) with natural assumptions on the randomness of the weights.

> **Section 3.1:** Fully connected DNN with T_n layers of equal width n. All weights are assumed to be standard normal and mutually independent. This is a standard configuration of the DNN at the start of training. We prove that as soon as $\lim_{n\to\infty} T_n = \infty$, the weights will with very high probability be such that the resulting classifier is very strongly sensitive to perturbations of the input no matter the input distribution and noise distribution. If T_n is bounded, the resulting DNN will produce a noise stable classifier.
>
> **Section 3.2:** All weights are once again standard normal, but some of them are now correlated: the columns of each W_t are multivariate normal with all correlation being ρ_n. The columns are mutually independent within and across the W_t's. We show that with ρ_n converging to 1 sufficiently fast, the resulting DNN becomes noise stable with high probability. If ρ_n converges to 1 slowly enough, the resulting classifier is with high probability strongly sensitive to perturbations.
>
> **Section 4:** $2k+1$-majority on "$2k+1$-trees with overlaps", i.e. the graph where the vertices/neurons in each generation share some children, see Figures 2a, 2c. It is proved that if the number

of children shared by two parents next to each other is $2k$ (corresponding to stride $s = 1$), then the resulting Boolean function is noise stable, whereas if the number of children shared is less than $2k$ (stride $s \geq 2$), then we get a noise sensitive Boolean function.

These models are convolutional neural nets where, in machine learning language, each filter represents a regular majority of the input bits. It will be observed that the results easily extend to the case where the weights of the filters are random under the only condition that there is at least some chance that a filter represents regular majority.

In the sections on the fully connected models, analysis of a certain Markov chain on $\{0, 1, \ldots, n-1, n\}$, which is symmetric around $n/2$ and absorbs in 0 and n, plays a central role. We believe that the structure of this Markov chain makes it interesting in its own right.

2 Different notions of noise sensitivity and noise stability

Let $\omega \in \{\pm 1\}^n$ be an i.i.d. (1/2,1/2) Boolean row vector and let $\{f_n\}$ be a sequence of Boolean functions from $\{\pm 1\}^n$ to $\{\pm 1\}$. Additionally let ω^ϵ be a Boolean row vector such that $\omega^\epsilon(i) = \omega(i)$ with probability $1-\epsilon$ and $\omega^\epsilon(i) = -\omega(i)$ with probability ϵ independently for different i. View ω^ϵ as a small perturbation of ω. In this context, we can now define noise sensitivity and noise stability of a sequence of Boolean functions f_n. In [**?**] these are defined as

Definition 2.1. The sequence $\{f_n\}$ is **noise sensitive** if for every $0 < \epsilon \leq 1/2$,

$$\lim_{n\to\infty} \operatorname{Cov}\left(f_n(\omega), f_n(\omega^\epsilon)\right) = 0$$

Definition 2.2. The sequence $\{f_n\}$ is **noise stable** if

$$\lim_{\epsilon\to 0} \sup_n \mathrm{P}(f_n(\omega) \neq f_n(\omega^\epsilon)) = 0.$$

The definition of noise stability is easily seen to be equivalent with the condition $\lim_{\epsilon\to 0} \limsup_n \mathrm{P}(f_n(\omega) \neq f_n(\omega^\epsilon)) = 0$. An example of a noise stable sequence is the weighted majority functions

$$\mathrm{maj}_{\theta^{(n)}}(\omega) = \operatorname{sign}\left(\sum_{j=1}^{n} \theta_j^{(n)} \omega(j)\right),$$

where $\theta_1^{(n)}, \ldots, \theta_n^{(n)}$ are arbitrary given constants [**?**]. An example of a noise sensitive sequence is the parity functions [**?**],

$$\mathrm{par}(\omega) = \prod_{j=1}^{n} \omega_j.$$

For noise sensitivity, there is a more general and often much stronger concept, which will turn out to be interesting here since most fully connected and sufficiently deep neural nets will turn out to be noise sensitive in a very strong way. If a sequence of functions $\{f_n\}$ is noise sensitive as defined above, then one can always find a sequence $\epsilon_n \leq 1/2$ tending to 0 with n slowly enough such that

$$\lim_{n\to\infty} \operatorname{Cov}(f_n(\omega), f_n(\omega^{\epsilon_n})) = 0.$$

Since $\operatorname{Cov}(f_n(\omega), f_n(\omega^\epsilon))$ is well known to be decreasing in ϵ on $[0, 1/2]$, the faster ϵ_n can be taken to decrease, the stronger the statement. This leads to the following definition,

Definition 2.3. Let $\epsilon_n \leq 1/2$ be non-increasing in n with $\epsilon_n \to 0$. The sequence $\{f_n\}$ is **quantitatively noise sensitive (QNS) at level** $\{\epsilon_n\}$ if,

$$\lim_{n\to\infty} \operatorname{Cov}\left(f_n(\omega), f_n(\omega^{\epsilon_n})\right) = 0.$$

The definitions of noise sensitivity and stability are by now standard when describing properties of deterministic sequences of Boolean functions. However, when dealing with randomness within the functions themselves we need a more general definition. This more general setup occurs in Boolean neural networks, where the network parameters Θ can be seen as random elements (usually depending on randomness in what training data is presented to the network and the initial values of the parameters before training). Let $\mathcal{F}_n$ be the set of all Boolean function from $\{\pm 1\}^n$ to $\{\pm 1\}$ and let π_n be a arbitrary probability measure on $\mathcal{F}_n$. Recall that for $0 \leq \epsilon \leq 1/2$ and each function f, we have $\mathrm{Cov}_{\omega,\omega^\epsilon}(f(\omega), f(\omega^\epsilon)) \geq 0$, and hence $\mathrm{Cov}_{f,\omega,\omega^\epsilon}(f(\omega), f(\omega^\epsilon)) \geq 0$. We can then define both quenched and annealed versions of noise sensitivity and noise stability as follows.

Definition 2.4. π_n is **quenched QNS at level** $\{\epsilon_n\}$ if for every $\delta > 0$ and $0 < \epsilon_n \leq 1/2$, there is an N such that for all $n \geq N$

$$\pi_n\{f_n \,:\, \mathrm{Cov}_{\omega,\omega^\epsilon}(f_n(\omega), f_n(\omega^{\epsilon_n})) \leq \delta\} \geq 1 - \delta.$$

Definition 2.5. π_n is **annealed noise QNS at level** $\{\epsilon_n\}$ if for every $0 < \epsilon_n \leq 1/2$

$$\lim_{n\to\infty} \mathrm{Cov}_{f_n,\omega,\omega^{\epsilon_n}}(f_n(\omega), f_n(\omega^{\epsilon_n})) = 0.$$

Definition 2.6. π_n is **quenched noise stable** if for every δ there is an $\epsilon > 0$ such that for all n,

$$\pi_n\{f_n : \mathrm{P}_{\omega,\omega^\epsilon}(f_n(\omega) \neq f_n(\omega^\epsilon)) < \delta\} \geq 1 - \delta$$

Definition 2.7. π_n is **annealed noise stable** if

$$\lim_{\epsilon\to 0} \sup_n \mathrm{P}_{f_n,\omega,\omega^\epsilon}(f_n(\omega) \neq f_n(\omega^\epsilon)) = 0$$

This notion of quenched noise sensitivity has arisen elsewhere; see [? ? ? ?]. Notice that if π_n has support on only one Boolean function, then these definitions are equivalent to the usual ones in Definition 2.1 and 2.2. In Theorem 2.1 we show which relations there are between these definitions.

Theorem 2.1. *Let $\mathcal{F}_n$ be the set of all Boolean functions on $\{-1,1\}^n \to \{-1,1\}$ and let π_n be a probability measure on $\mathcal{F}_n$. Then the following are true*

(i) $\{\pi_n\}$ is annealed QNS at level $\{\epsilon_n\}$ iff $\{\pi_n\}$ is quenched QNS at level $\{\epsilon_n\}$ and $\mathrm{Var}_{f_n}(\mathrm{E}_\omega[f_n(\omega)]) \to 0$ as $n \to \infty$.

(ii) $\{\pi_n\}$ is annealed noise stable iff $\{\pi_n\}$ is quenched noise stable.

Proof. To prove the first statement we use the conditional covariance formula that for any random variables X, Y and Z

$$\mathrm{Cov}(X,Y) = \mathrm{E}[\mathrm{Cov}(X,Y|Z)] + \mathrm{Cov}(\mathrm{E}[X|Z], \mathrm{E}[Y|Z])$$

to observe that

$$\mathrm{Cov}_{f,\omega,\omega^\epsilon}(f(\omega), f(\omega^{\epsilon_n})) = \mathrm{E}_f\left[\mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n}))\right] + \mathrm{Cov}_f(\mathrm{E}_\omega[f(\omega)], \mathrm{E}_{\omega^{\epsilon_n}}[f(\omega^{\epsilon_n})]). \qquad (3)$$

Now observe the following. First, the first term on the right is always non-negative. Secondly, since ω and ω^{ϵ_n} are equal in distribution, the second term on the right is equal to $\mathrm{Var}_f(\mathrm{E}_\omega[f(\omega)])$.

We now prove (i) starting with that quenched QNS at level $\{\epsilon_n\}$ and $\mathrm{Var}_f(\mathrm{E}_\omega[f_n(\omega)]) \to 0$ as $n \to \infty$ leads to annealed QNS at level $\{\epsilon_n\}$. Fix $\delta \in (0,1]$. We know that there exists an N such that for all $n > N$

$$\pi_n\{f : \mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) \leq \frac{\delta}{4}\} \geq 1 - \frac{\delta}{4}$$

and

$$\mathrm{Var}_f(\mathrm{E}_\omega[f(\omega)]) < \frac{\delta}{2}.$$

Using (3) this leads to

$$\mathrm{Cov}_{f,\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) < \mathrm{E}_f\left[\frac{\delta}{4}\mathrm{I}_{\mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega),f(\omega^{\epsilon_n}))\leq\frac{\delta}{4}} + \mathrm{I}_{\mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega),f(\omega^{\epsilon_n}))>\frac{\delta}{4}}\right] + \frac{\delta}{2} < \delta$$

This proves the first direction of (i). For the other direction, fix $\delta \in (0, 1]$. Since π_n is annealed QNS at level $\{\epsilon_n\}$ there exists an N such that $\forall n > N$

$$\mathrm{Cov}_{f,\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) < \delta^2.$$

Now, using (3) and the fact that $\mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) \geq 0$ for all f it must be that

$$\mathrm{Var}_f(\mathrm{E}_\omega[f(\omega)]) < \mathrm{Cov}_{f,\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) < \delta^2 < \delta.$$

Hence the variances converge to zero. Additionally, for such n we have that $0 \leq \mathrm{Cov}_{f,\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) \leq \delta^2$. Now using Markov's inequality and (3) we get

$$\pi_n\{f : \mathrm{Cov}_{\omega,\omega^{\epsilon_n}}(f(\omega), f(\omega^{\epsilon_n})) \geq \delta\} \leq \delta$$

which concludes (i).

Next we prove (ii). We start by showing that quenched noise stability leads to annealed noise stability. Fix $\delta > 0$. π_n being quenched noise stable means that there exist an $\epsilon > 0$ such that for all n

$$\pi_n\{f : \mathrm{P}(f(\omega) \neq f(\omega^\epsilon)) < \frac{\delta}{2}\} \geq 1 - \frac{\delta}{2}.$$

Hence

$$\mathrm{P}_{f,\omega,\omega^\epsilon}(f(\omega) \neq f(\omega^\epsilon)) = \mathrm{E}_f\left[\mathrm{E}_{\omega,\omega^\epsilon}[\mathrm{I}_{f(\omega)\neq f(\omega^\epsilon)}]\right] \leq$$
$$\mathrm{E}_f\left[\frac{\delta}{2}\mathrm{I}_{\mathrm{E}_{\omega,\omega^\epsilon}[I_{f(\omega)\neq f(\omega^\epsilon)}]\leq\frac{\delta}{2}} + \mathrm{I}_{\mathrm{E}_{\omega,\omega^\epsilon}[\mathrm{I}_{f(\omega)\neq f(\omega^\epsilon)}]>\frac{\delta}{2}}\right] < \delta.$$

This proves the first part. Now we prove that annealed noise stable implies quenched noise stable.

Fix $\delta > 0$ and pick an $\epsilon > 0$ sufficiently small such that $\mathrm{P}_{f,\omega,\omega^\epsilon}(f(\omega) \neq f(\omega^\epsilon)) < \delta^2$. Such an ϵ are guaranteed to exist since π_n are annealed noise stable. Then due to Markov´s inequality

$$\pi_n(f : \mathrm{P}_{\omega,\omega^\epsilon}(f(\omega) \neq f(\omega^\epsilon)) > \delta) < \delta$$

which gives us quenched noise stable. This proves (ii). □

As seen from statement (ii), π_n being annealed noise stable is equivalent to π_n being quenched noise stable. Therefore we will further on only refer it as π_n being noise stable.

3 Random Boolean feed forward neural networks

In this section, we investigate a Boolean function structure $f(\omega, \Theta)$ inspired by feed-forward neural networks. Let $\omega_0 \in \{-1, 1\}^n$ be the input bits as a column vector. Then we can recursively define $\omega_t = \mathrm{sign}(\theta_t\omega_{t-1})$ where $\theta_t \in R^{n\times n}$ and sign acts pointwise. In the deep learning literature, the θ's and sign would be referred to as the weights of f and the activation function respectively. In the Neural Network literature, each t corresponds to a layer where ω_t are seen as the bits, or nodes, at layer t. The iteration is done for $t = 1, \ldots, T$ for some predetermined number $T = T_n$ giving us the final output $h(\omega_T)$ where $h = h_n$ is some Boolean function $\{-1,1\}^n \to \{-1,1\}$. A known fact is that a typical Neural Network can approximate any function to arbitrary accuracy as long as the amount of tunable parameters (a.k.a. weights), $\Theta = \{\theta_1, \dots, \theta_T\}$, is large enough. In the Boolean setting, these models can still represent a huge number of Boolean functions. However, there are some limitations since no bias term is present in our Boolean Network, which typically is in Neural Networks in practice. Typically Θ is determined by some training algorithm based on the observation data, which maximises the likelihood of the model. This means that Θ is not deterministic since there is both randomness in the observed data and in the optimisation algorithm. Therefore we can consider a probability measure π_n on all f where the randomness comes from Θ.

Here we consider cases where the θ_t's are independent and, for each t, the columns of θ_t are independent and jointly normal $N(0,\Sigma)$ with two different versions of Σ. These cases, which thus induce their respective measures $\{\pi_n\}$ on the set of Boolean function, are

1. $\Sigma = \mathbf{I}_{n\times n}$.
2. $\Sigma = \rho + (1-\rho)\mathbf{I}_{n\times n}$. A useful way to construct such θ_t's is to define $\theta_t(i,j) = \sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)$ where $\nu_t(i)$, $\psi_t(i,j) \sim N(0,1)$ all being independent.

The case $\rho = 0$ represents a typical starting state for the network before training. After some training, we expect the parameters of the network to be dependent. The cases $\rho > 0$ are examples of such dependence. Of course, we do not expect that the true dependence after training is represented in this way. Our assumptions should therefore be viewed as a simplifying mathematical framework under which our theorems can be proved. Somewhat related to this question of the behaviour of the parameters, in [**?**], a particular statistical structure for the values of intermediate layers in some convolution codes was discovered.

The following lemma is crucial.

Lemma 3.1. Let $x, y \in \{-1,1\}^n$ be column vectors such that $|i : x(i) \neq y(i)| = nv$ for some $v \in [0,1]$ and let θ be a random row vector such that $\theta \sim N(0, I_{n\times n})$. Then

$$\mathrm{P}\left(\mathrm{sign}(\theta x) \neq \mathrm{sign}(\theta y)\right) = \frac{2}{\pi}\arctan\left(\sqrt{\frac{v}{1-v}}\right)$$

Proof. Let C be the set of indices where x and y differ. Notice that $|C| = nv$. Consider the line segment between x and y defined as $\frac{y+x}{2} + \tau\frac{y-x}{2}$, $\tau \in [-1,1]$. Then $\mathrm{sign}(\theta x) \neq \mathrm{sign}(\theta y)$ if there is a solution to

$$\theta^T\left(\frac{y+x}{2} + \tau\frac{y-x}{2}\right) = 0 \tag{4}$$

for some $\tau \in [-1,1]$.

Now, let $A_{\Lambda,k} = \{j : j \in \Lambda, x(j) = k\}$ for some set Λ. Since both x and y are in the hypercube $\{-1,1\}^n$, the condition can be rewritten as

$$\left|\sum_{j\in A_{C^c,1}}\theta(j) - \sum_{j\in A_{C^c,-1}}\theta(j)\right| \leq \left|\sum_{j\in A_{C,1}}\theta(j) - \sum_{j\in A_{C,-1}}\theta(j)\right|. \tag{5}$$

Define X and Y from the following equations

$$\sqrt{nv}X = \sum_{j\in A_{C,1}}\theta(j) - \sum_{j\in A_{C,-1}}\theta(j)$$

and

$$\sqrt{n-nv}Y = \sum_{j\in A_{C^c,1}}\theta(j) - \sum_{j\in A_{C^c,-1}}\theta(j).$$

Then it is easy to check that X and Y are standard independent normal and that the event in (5) can be rewritten as

$$-\sqrt{\frac{v}{1-v}}|X| \leq Y \leq \sqrt{\frac{v}{1-v}}|X|.$$

Due to symmetry around $Y = 0$, this results in

$$\mathrm{P}\left(-\sqrt{\frac{v}{1-v}}|X| \leq Y \leq \sqrt{\frac{v}{1-v}}|X|\right) = \int_{-\infty}^{\infty} \int_{-\sqrt{\frac{v}{1-v}}|x|}^{\sqrt{\frac{v}{1-v}}|x|} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dy dx$$

$$= 2\int_{-\arctan\left(\sqrt{\frac{v}{1-v}}\right)}^{\arctan\left(\sqrt{\frac{v}{1-v}}\right)} \int_0^{\infty} \frac{r}{2\pi} e^{-r^2/2} d\varphi dr = \int_{-\arctan\left(\sqrt{\frac{v}{1-v}}\right)}^{\arctan\left(\sqrt{\frac{v}{1-v}}\right)} \frac{1}{\pi} d\varphi = \frac{2}{\pi} \arctan\left(\sqrt{\frac{v}{1-v}}\right)$$

where the second equality is due to the substitution $x = r\cos(\varphi)$ and $y = r\sin(\varphi)$. This concludes the proof. □

In both noise parameter settings 1 and 2, we have independence between the weights leading into a node. This means, according to Lemma 3.1, that given ω_{t-1} and $(\omega^{\epsilon})_{t-1}$ differ at nv bits the probability that $\omega_t(i)$ and $(\omega^{\epsilon})_t(i)$ differ is $\frac{2}{\pi}\arctan(\sqrt{\frac{v}{1-v}})$ for all i. The difference between the two parameter settings is that in 2 the output bits of a layer are usually correlated. Also, due to symmetry, the probability of a disagreement at a fixed point, i.e. $\omega_t(i) \neq (\omega^{\epsilon})_t(i)$, at layer t depends only on the number of disagreements between ω_{t-1} and $(\omega^{\epsilon})_{t-1}$ and not where they disagree. This means that the number of disagreements at layer t, which we denote by $D_t = D_t^{\epsilon} = D_t^{\epsilon_n}$, can in both cases be seen as a Markov chain with $n+1$ states, where $D_t = 0$ and $D_t = n$ are absorbing states. Notice that D_t depends on the initial noise ϵ since D_0 corresponds to the number of bit disagreements created by the initial noise. However, for the sake of lighter notation we will not have a specific suffix showing this dependence.

3.1 Uncorrelated networks

In the uncorrelated case, D_t is binomially distributed according to $(D_t|D_{t-1} = nv) \sim \mathrm{Bin}(n, g(v))$ where $g(v) := \frac{2}{\pi}\arctan\left(\sqrt{\frac{v}{1-v}}\right)$. Note that if ϵ_n is of order $1/n$ or lower, then $\mathrm{P}(D_1 = 0)$ does not get to 0 at the same time as $P(D_1 > n/2) \to 0$, which by symmetry implies that there cannot be QNS at that level. Hence one must have at least $n\epsilon_n \to \infty$ for QNS to be possible.

The following theorem almost entirely considers the sensitivity properties of f_{n,T_n} and shows that under very mild assumptions for T_n larger than a specified function of n, $\{f_{n,T_n}\}$ is QNS in this the strongest possible sense (in particular that f_{n,T_n} is noise sensitive in the original sense as soon as $T_n \to \infty$). Indeed, since the distribution of θ_t is such that even if we know ω and ω^{ϵ}, any trace of that knowledge is forgotten after the first layer, f_{n,T_n} has very strong sensitivity properties even for fixed input and fixed noise. What we mean precisely with this is formulated separately in Theorem 3.2. Note in particular part (ii) implies that the input and noise distribution do not need to be uniform but can indeed be taken to reflect what is realistic in the application at hand, e.g. a distribution over images of real handwritten digits.

Parts of Theorem 3.2 are clearly stronger than their counterparts in Theorem 3.1, but we find it natural to state and prove the weaker statements first and then extend them by pointing out the fairly minor extra observations that need to made in the proof.

Theorem 3.1. *Consider the above fully connected network with i.i.d. normal entries in each θ_t and i.i.d. θ_t's. Let $1/2 \geq \epsilon_n \downarrow 0$ be such that $n\epsilon_n \to \infty$ and let $\lim_{n\to\infty} K_n/\log(1/\epsilon_n) = \infty$. Then*

(i) if $\lim_{n\to\infty} b_n = 0$, $\lim_{n\to\infty} nb_n = \infty$ and $T_n \in [K_n, e^{b_n n}]$, then for any Boolean functions $\{h_n\}$, the resulting $\{f_{n,T_n}\}$ is annealed QNS at level $\{\epsilon_n\}$ with respect to $\{\pi_n\}$,

(ii) if the h_n's are odd and $T_n \geq K_n$, then $\{f_{n,T_n}\}$ is annealed QNS at level $\{\epsilon_n\}$,

(iii) if $T_n \geq K_n$ then for any $\{h_n\}$, $\{f_{n,T_n}\}$ is quenched QNS at level $\{\epsilon_n\}$,

(iv) there are Boolean functions $\{h_n\}$ such that for T_n growing sufficiently fast with n, $\{f_{n,T_n}\}$ is not annealed noise sensitive.

In addition,

(v) if h_n is noise stable and T_n is bounded, then $\{f_{n,T_n}\}$ is annealed (and hence quenched) noise stable.

Remark. If one randomly chooses a sequence of Boolean functions uniformly among *all* Boolean functions, it is known that the sequence will asymptotically almost surely be noise sensitive; see Exercise 1.14 in [**?**]. While the Boolean functions arising in Theorem 3.1 here are also random, they have a very specific form.

Theorem 3.2. *Consider the above fully connected network with i.i.d. normal entries in each θ_t and i.i.d. θ_t's. Let $1/2 \geq \epsilon_n \downarrow 0$ be such that $n\epsilon_n \to \infty$ and let $\lim_{n\to\infty} K_n/\log(1/\epsilon_n) = \infty$. Then*

(i) under the assumptions of either (i) or (ii) in Theorem 3.1, for any fixed $\omega, \eta \in \{-1,1\}^n$ with $\omega \notin \{\eta, -\eta\}$,

$$\lim_{n\to\infty} \mathrm{Cov}_\Theta(f_{n,T_n}(\omega), f_{n,T_n}(\eta)) = 0.$$

(ii) under the assumptions of either (i) or (ii) in Theorem 3.1, for any probability measure $\mathbf{Q}_n$ on $\{-1,1\}^n \times \{-1,1\}^n$ such that $\lim_{n\to\infty} \mathbf{Q_n}(\{(\omega,\eta) : \eta \in \{\omega, -\omega\}\}) = 0$.

$$\lim_{n\to\infty} \mathrm{Cov}_{\mathbf{Q_n},\Theta}(f_{n,T_n}(\omega), f_{n,T_n}(\eta)) = 0.$$

(iii) Assume that h_n is odd and fix any $k \in \{1, 2, \ldots, n-1\}$ and $\delta > 0$. Fix also $\omega \in \{-1,1\}^n$ and let $M_k = M_k^{(n)}(\omega)$ be the number of η with $\eta(i) \neq \omega(i)$ for exactly k indexes i, such that $f_{n,T_n}(\eta) \neq f_{n,T_n}(\omega)$. Then for $T_n \geq K_n$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{M_k}{\binom{n}{k}} \notin \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right)\right) = 0.$$

Proof of Theorem 3.1. Recall $D_t = D_t^{\epsilon_n}$ from above. In the sequel in (i)-(iii), to not burden the notation, write just ϵ for ϵ_n with the understanding that $\epsilon = \epsilon_n$.

The conditional distribution of $(\omega_T, (\omega^\epsilon)_T)$ given $\mathcal{F}_{T-1} := \sigma(\omega_0, (\omega^\epsilon)_0, \theta_1, \ldots, \theta_{T-1})$ equals that of $(\omega_0, (\omega^{g(D_{T-1}/n)})_0)$. In other words; to determine the distribution of $(\omega_T, \omega_T^\epsilon)$ given $\mathcal{F}_{T-1}$, we only need to know D_{T-1}. Consequently, $\mathrm{E}[f(\omega_0)f((\omega^\epsilon)_0)|D_{T-1} = d] = \mathrm{E}[h(\omega_T)h((\omega^\epsilon)_T)|D_{T-1} = d] = \mathrm{E}[h(\omega_0)h((\omega^{g(d/n)})_0)]$. Let us study how D_t behaves, started from $D_0 \sim \mathrm{Bin}(n,\epsilon)$. First observe that $g(v)/v$ is decreasing on $(0, 1/2)$. This holds as $g(0) = 0$ and an easy computation shows $g''(v) < 0$ for $v \in [0, 1/2]$. Since $g'(1/2) = 2/\pi$, it follows from Taylor's formula that

$$g\left(\frac{1}{2} - \delta\right) > \frac{1}{2} - \frac{2}{3}\delta$$

for sufficiently small $\delta > 0$. Fixing such a δ, it then follows that $g(d/n) > (1 + 2\delta/3)d/n$ whenever $1 \leq d < n/2 - \delta n$. This means by Chernoff bounds and the fact that g is increasing that there is $\kappa = \kappa(\delta, \epsilon) > 0$ such that

- for $\epsilon n/2 < d < (1/2 - \delta)n/2$, $\mathrm{P}(D_{t+1} < (1 + 2\delta/3)d|D_t = d) < e^{-\kappa\sqrt{n}}$,
- for $d > (1/2 - \delta)n$, $\mathrm{P}(D_{t+1} < (1/2 - \delta)n|D_t = d) < e^{-\kappa n}$.

(Here the $\sqrt{n}$ in the exponent in the first point follows from that $g(1/n)$ is of order $1/\sqrt{n}$ and $\epsilon_n > 1/n$ for n large.) Combining these two points and the symmetry of g around 1/2,

$$\mathrm{P}\left(\exists t \in \left[\frac{\log \frac{1-2\delta}{\epsilon}}{\log\left(1+\frac{2\delta}{3}\right)}, T\right] : D_t \notin \left[\left(\frac{1}{2}-\delta\right)n, \left(\frac{1}{2}+\delta\right)n\right]\right) < Te^{-\kappa n}.$$

To prove (i), let $L_n = e^{b_n n}$. Take now $T \in \left[1 + \frac{\log \frac{1-2\delta}{\epsilon_n}}{\log\left(1+\frac{2\delta}{3}\right)}, L_n\right]$. Conditionally on $D_{T-1} = d \in [(1/2-\delta)n, (1/2+\delta)n]$, we have that $(\omega_T, (\omega^\epsilon)_T)$ equals in distribution $(\omega_0, (\omega^\alpha)_0)$ for $\alpha = g(d/n) \in [1/2-\delta, 1/2+\delta]$. Equation (4.2) in Section 4.3 in [**?**] implies that for $\rho > 0$, there exists $\delta > 0$ so that for all Boolean functions h and for all $\alpha \in [1/2-\delta, 1/2+\delta]$,

$$\mathrm{Cov}(h(\omega), h(\omega^\alpha)) < \rho.$$

Hence for $\rho > 0$ and $\delta > 0$ sufficiently small it follows from the above that $\mathrm{Cov}(f(\omega_0), f((\omega^\epsilon)_0)|D_{T-1} \in [(1/2-\delta)n, (1/2+\delta)n]]) < \rho/2$. Since $\mathrm{P}(D_{T-1} \in [(1/2-\delta)n, (1/2+\delta)n]) > 1 - L_n e^{-\kappa n} > 1 - \rho/8$ for large n, we get

$$\mathrm{Cov}(f(\omega_0), f((\omega^\epsilon)_0)) < \rho.$$

To see this, let B be the event $\{D_{T-1} \in [(1/2-\delta)n, (1/2+\delta)n]\}$ and recall that $\mathrm{Cov}(f(\omega_0), f((\omega^\epsilon)_0)) = \mathrm{E}[\mathrm{Cov}(f(\omega_0), f((\omega^\epsilon)_0)|I_B)] + \mathrm{Var}(\mathrm{E}[f(\omega_0)|I_B])$. Since $\mathrm{P}(B^c) < \rho/8$, it is easy to see that the first term is bounded by $\mathrm{Cov}(f(\omega_0), f((\omega^\epsilon)_0)|B) + \rho/4$ and the second term is bounded by $\rho/2$.

This proves that for some constant $K = K(\rho, \epsilon)$, $T \in [K, L_n]$ and n sufficiently large, $\mathrm{Cov}(f(\omega), f(\omega^\epsilon)) < \rho$. In particular if $K_n \to \infty$ and $T \in [K_n, L_n]$, then f is annealed QNS. This proves (i).

Next consider (ii) and (iii). This amounts to considering what happens to D_t in the long run. We have argued that with overwhelming probability, D_t will quickly approach $n/2$ and stay there for a very long time. However, after an even longer time, D_t will end up in one of the absorbing states 0 or n. Then the above argument for annealed QNS does not hold (since it is no longer true that $D_t \in [(1/2-\delta)n, (1/2+\delta)n]$ with high probability).

Let $\{D'_t\}$ be a copy of $\{D_t\}$ but started with $D'_0 \sim \mathrm{Bin}(n, 1/2)$. We can couple (D_0, D'_0) so that $D_0 \le D'_0$, so do that.

Since g is increasing, the distribution of D_t given $D_{t-1} = d$ is increasing in d. Hence, since we have coupled so that $D_0 \le D'_0$, there is a further coupling (D, D') such that $D'_t \ge D_t$ for every t. Use such a coupling from now on and note that a consequence is that if at some point $D'_t = D_t$, then also $D'_s = D_s$ for all $s \ge t$. By the above with $t_0 = 1 + \log(4\epsilon)/\log(6/7)$, $P(\forall t \in [t_0, L_n] : D_t, D'_t \in [n/4, 3n/4]) > 1 - L_n e^{-\kappa n} > 1 - e^{-\kappa n/2}$ for large n. For all $d, d' \in [n/4, 3n/4]$ and $d \le d'$,

$$\begin{aligned}\mathrm{E}[D'_t - D_t | D_{t-1} = d, D'_{t-1} = d'] &= n\left(g\left(\frac{d'}{n}\right) - g\left(\frac{d}{n}\right)\right)\\ &< \frac{3}{4}(d' - d),\end{aligned}$$

where the last inequality follows on observing that $g'(v) < 3/4$ for $v \in [1/4, 3/4]$. This gives for $t \in [t_0, L_n]$,

$$\begin{aligned}\mathrm{E}[D'_t - D_t] &\le ne^{-\kappa n/2} + \mathrm{E}\left[D'_t - D_t | D_{t-1}, D'_{t-1} \in \left[\frac{n}{4}, \frac{3n}{4}\right]\right]\\ &\le 2ne^{-\kappa n/2} + \frac{3}{4}\mathrm{E}[D'_{t-1} - D_{t-1}].\end{aligned}$$

For such t, $\mathrm{E}[D'_{t-1} - D_{t-1}] \ge 1/n$ and the right hand side is smaller than $(4/5)\mathrm{E}[D'_{t-1} - D_{t-1}]$ for large n. By induction we thus get $\mathrm{E}[D'_{t_0+t} - D_{t_0+t}] \le \max(1/n, (4/5)^t n)$ for $t_0 + t \le L_n$. This gives $\mathrm{E}[D'_{t_0+t} - D_{t_0+t}] \le 1/n$ whenever $L_n \ge t \ge 9\log n$ and n large. Hence $\mathrm{E}[D'_t - D_t] \le 1/n$ for all

$L_n \geq t \geq 10 \log n$ and n large. By Markov's inequality, $\mathrm{P}(D'_t \neq D_t) \leq 1/n$ for all $L_n \geq t \geq 10 \log n$ and n large. Since if $D'_t = D_t$ for some t, then $D'_s = D_s$ for all $s \geq t$, we get $\mathrm{P}(D'_t \neq D_t) \leq 1/n$ for all $t \geq 10 \log n$.

Thus, taking $T \geq 10 \log n$, $\mathrm{P}(D_T \neq D'_T) < 1/n$. This means that the total variation distance between the distribution of D_T and D'_T is less than $1/n$, i.e. $\sum_{d=0}^n |\mathrm{P}(D'_T = d) - \mathrm{P}(D_T = d)| < 2/n$. This gives for $T \geq 10 \log n + 1$

$$\begin{aligned}
\mathrm{E}[f(\omega_0)f((\omega^{1/2})_0] &= \mathrm{E}[h(\omega_T)h((\omega^{1/2})_T)] \\
&= \sum_d \mathrm{E}[h(\omega_T)h((\omega^{1/2})_T)|D'_{T-1} = d]P(D'_{T-1} = d) \\
&\geq \sum_d \mathrm{E}[h(\omega_T)h((\omega^{\epsilon})_T)|D_{T-1} = d]\mathrm{P}(D_{T-1} = d) - \sum_d |\mathrm{P}(D'_{T-1} = d) - \mathrm{P}(D_{T-1} = d)| \\
&\geq \mathrm{E}[h(\omega_T)h((\omega^{\epsilon})_T)] - \frac{2}{n} \\
&= \mathrm{E}[f(\omega_0)f((\omega^{\epsilon})_0)] - \frac{2}{n},
\end{aligned}$$

where the first inequality follows from that $|h| \leq 1$. To now prove quenched QNS for all h and $T \geq 10 \log n$ observe that we now have

$$\begin{aligned}
\mathrm{E}[\mathrm{Cov}(f(\omega_0), f((\omega^{\epsilon})_0)|\Theta)] &= \mathrm{E}[\mathrm{Cov}(f(\omega_0), f((\omega^{\epsilon})_0)|\Theta) - \mathrm{Cov}(f(\omega_0), f((\omega^{1/2})_0)|\Theta)] \\
&= \mathrm{E}[\mathrm{E}[f(\omega_0)f((\omega^{\epsilon})_0)|\Theta] - \mathrm{E}[f(\omega_0)f((\omega^{1/2})_0)|\Theta]] \\
&= \mathrm{E}[f(\omega_0)f((\omega^{\epsilon})_0)] - \mathrm{E}[f(\omega_0)f((\omega^{1/2})_0)] \\
&< \frac{2}{n}.
\end{aligned}$$

Since for any fixed f, $\mathrm{Cov}(f(\omega), f(\omega^{\epsilon})) \geq 0$, this implies quenched QNS for $T \geq 10 \log n$. Combining with (i), this gives (iii).

If h is also odd and $T \geq 10 \log n$, we have $h(-(\omega^{1/2})_T) = h((-\omega^{1/2})_T) = -h((\omega^{1/2})_T)$ and since $(\omega, \omega^{1/2}) =_d (\omega, -\omega^{1/2})$, we have $\mathrm{E}[h(\omega_T)h((\omega^{1/2})_T)] = 0$. Thus

$$\mathrm{E}[f(\omega_0)f((\omega^{\epsilon})_0] \leq \frac{2}{n}.$$

Since h odd implies $\mathrm{E}[f(\omega_0)] = 0$, this proves annealed QNS for all odd h and $T \geq 10 \log n$ and combining with (i) we get (ii).

To prove (iv), we need an example of an h that is not odd and where annealed noise sensitivity does not hold for large T. To achieve this, first observe that at each t, $\mathrm{P}(D_t \in \{0, n\}|D_{t-1}) \geq 2^{-n+1}$. Hence, as $D_t \in \{0, n\}$ implies that $D_{t'} = D_t$ for all $t' \geq t$, for T_n such that $T_n 2^{-n} \to \infty$, $\mathrm{P}(D_{T_n} \notin \{0, n\}) \to 0$. This entails $\mathrm{P}((\omega^{\epsilon})_{T_n} = \pm \omega_{T_n}) \to 1$. Now let h_n be any even function. Then $\mathrm{P}(h_n(\omega_{T_n}) \neq h_n((\omega^{\epsilon})_{T_n})) \to 0$ and so in fact f_n is annealed noise stable. If also $-1 < \liminf_n \mathrm{E}[h_n(\omega)] \leq \limsup_n \mathrm{E}[h_n(\omega)] < 1$, then $\{f_n\}$ is not annealed noise sensitive.

For (v), assume for simplicity that T_n equals the constant T; going from this to general bounded T_n is easy and left to the reader. Fix a small $\delta > 0$. Since h_n is stable one can pick $\rho > 0$ small enough that $\sup_n \mathrm{P}(h_n(\omega) \neq h_n(\omega^{\xi})) < \delta$ whenever $\xi < \rho$. Pick such a ρ and let $\epsilon = \rho^{2^T}$. We have $g(v) = (2/\pi) \arctan \sqrt{v/(1-v)} < (2/3)\sqrt{v}$ for $v < \rho$ and ρ small enough. By Chernoff bounds we then have that there is a $\kappa > 0$ independent of n such that

$$\mathrm{P}(D^{\epsilon}_{T-1} \geq n\rho^2) < e^{-\kappa n}.$$

Given $D_{T-1} = nv$ for $v < \rho^2$, the conditional distribution of $(\omega_T, (\omega^{\epsilon})_T)$ is that of $(\omega, \omega^{g(v)})$ and since $g(v) < \rho$, we get by noise stability of h_n that

$$\mathrm{P}(f_{n,T_n}(\omega) \neq f_{n,T_n}(\omega^{\epsilon})|D^{\epsilon}_{T-1} < \rho^2) < \delta.$$

Summing up, this now gives

$$\mathrm{P}(f_{n,T_n}(\omega) \neq f_{n,T_n}(\omega^\epsilon)) < e^{-\kappa n} + \delta$$

This easily implies noise stability.

$\square$

Proof of Theorem 3.2. Starting with (i), this follows from the simple observation that if d is the number of disagreements between ω and η, then ω_1 is a vector i.i.d. fair coin flips and given ω_1, η_1 differs from ω_1 in uniformly random positions whose number is binomial with parameters n and $g(d/n)$. Hence all arguments from (i) or (ii) in Theorem 3.1 can be copied from above from $t = 1$.

Part (ii) is an immediate corollary of (i) on observing that

$$\mathrm{Cov}_{\mathbf{Q_n},\Theta}(f_{n,T_n}(\omega), f_{n,T_n}(\eta)) = \mathrm{E}_{\mathbf{Q}_n}[\mathrm{Cov}_\Theta(f_{n,T_n}(\omega), f_{n,T_n}(\eta))] + \mathrm{Cov}_{\mathbf{Q}_n}(\mathrm{E}_\Theta[f_{n,T_n}(\omega)], \mathrm{E}_\Theta[f_{n,T_n}(\eta)])$$

and by the invariance and symmetry properties of Θ, $\mathrm{E}_\Theta[f_{n,T_n}(\omega)]$ is independent of ω and thus a constant and hence $\mathrm{Cov}_{\mathbf{Q}_n}(\mathrm{E}_\Theta[f_{n,T_n}(\omega)], \mathrm{E}_\Theta[f_{n,T_n}(\eta)]) = 0$.

Now for (iii) fix ω, let A_k be the set of η' that each differs from ω at k positions and take $\eta \in A_k$. We have by (i), since h_n odd implies $\mathrm{E}[f_{n,T_n}(\omega)] = \mathrm{E}[f_{n,T_n}(\eta)] = 0$, that $|\mathrm{E}[f_{n,T_n}(\omega) f_{n,T_n}(\eta)]| < \delta^4$, i.e. $\mathrm{P}(f_{n,T_n}(\omega) \neq f_{n,T_n}(\eta)) \in (1/2 - 1/2\delta^4, 1/2 + 1/2\delta^4)$, for n large.

Now let $\bar{M}_k = \bar{M}_k^{(n)}(\omega) = |\{(\eta, \xi) \in A_k : f_{n,T_n}(\eta) \neq f_{n,T_n}(\xi)\}|$. It follows that

$$\mathrm{E}[\bar{M}_k] \in \left((1-\delta^4)\frac{1}{2}\binom{\binom{n}{k}}{2}, (1+\delta^4)\frac{1}{2}\binom{\binom{n}{k}}{2}\right) \subseteq \left((1-\delta^4)\frac{1}{4}\binom{n}{k}^2, (1+\delta^4)\frac{1}{4}\binom{n}{k}^2\right).$$

Let M be the maximum value that $\bar{M}_k$ can take on, so that

$$M \leq \left(\frac{\binom{n}{k}}{2}\right)^2 = \frac{1}{4}\binom{n}{k}^2.$$

Since $M - \bar{M}_k$ is nonnegative, it follows from Markov's inequality that

$$\mathrm{P}\left(\bar{M}_k \leq (1-\delta^2)\frac{1}{4}\binom{n}{k}^2\right) < \delta^2.$$

Also if $\bar{M}_k > (1-\delta^2)\frac{1}{4}\binom{n}{k}^2$, we have $X_k^+ \in ((1-\delta)\frac{1}{2}\binom{n}{k}, (1+\delta)\frac{1}{2}\binom{n}{k})$, where X_k^+ is the number of $\eta \in A_k$ with $f_{n,T_n}(\eta) = 1$. This gives $M_k \in ((1-\delta)\frac{1}{2}\binom{n}{k}, (1+\delta)\frac{1}{2}\binom{n}{k})$ and hence for n sufficiently large,

$$\mathrm{P}\left(\frac{M_k}{\binom{n}{k}} \notin \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right)\right) < \delta^2.$$

This proves (iii).

$\square$

(The property described in the statement (ii) in Theorem 3.4 could be referred to as f_{n,T_n} being annealed noise sensitive with respect to $\mathbf{Q}_n$.)

3.2 Correlated networks

In this section we will investigate the noise sensitivity of a deep network where the network weights θ are sampled from a correlated normal distribution, i.e. Case 2 in the introduction.

More precisely, the model is that the random matrices $\theta_1, \theta_2, \ldots, \theta_T$ are independent and for given t, the columns $\theta_t(\cdot, 1), \ldots, \theta_t(\cdot, n)$ are independent. However, each column $\theta_t(\cdot, j)$ of each θ_t is now assumed to be n-dimensional Gaussian with expectation 0 and covariance matrix Σ, where $\Sigma_{i,i} = 1$, $i = 1, \ldots, n$ and $\Sigma_{i,i'} = \rho$, $1 \leq i < i' \leq n$. Here $\rho = \rho_n$ is a positive given correlation and we are interested in providing conditions on ρ that ensure noise stability or noise sensitivity of $f = f_n = f_{n,T_n}$ defined as in the previous section.

To model the vectors $\theta_t(\cdot, j)$, we let $\theta_t(i,j) = \sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)$, where the $\nu_t(j)$'s and $\psi_t(i,j)$'s are all independent standard Gaussian.

Since the entries in any given row of θ_t are i.i.d. standard normals, Lemma 3.1 still says that $\mathrm{P}(F_t(i)|D_{t-1} = d) = g(d/n)$, where $F_t(i) = \{\omega_t(i) \neq (\omega^\epsilon)_t(i)\}$. However the events $F_t(i)$ and $F_t(i')$ are not, as in Section 2, conditionally independent given D_{t-1}.

Recall the proof of Lemma 3.1, where the following observation was made. Here we recall that $C = \{j : \omega_{t-1}(j) \neq (\omega^\epsilon)_{t-1}(j)\}$, $A_{\Lambda,k} = \{j : j \in \Lambda, \omega(j) = k\}$ and

$$F_t(i) = \left\{ \left| \sum_{j \in A_{C^c,1}} \theta(i,j) - \sum_{j \in A_{C^c,-1}} \theta(i,j) \right| \leq \left| \sum_{j \in A_{C,1}} \theta(i,j) - \sum_{j \in A_{C,-1}} \theta(i,j) \right| \right\}.$$

Using the above representations of $\theta_t(i,j)$ this becomes

$$\begin{aligned} F_t(i) = \Bigg\{ &\left| \sum_{j \in A_{C^c,1}} \left(\sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)\right) - \sum_{j \in A_{C^c,-1}} \left(\sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)\right) \right| \\ &\leq \left| \sum_{j \in A_{C,1}} \left(\sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)\right) - \sum_{j \in A_{C,-1}} \left(\sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)\right) \right| \Bigg\}. \end{aligned}$$

Make the following substitutions

$$\begin{aligned} \sqrt{d}U_t^C &= \sum_{j \in A_{C,1}} \nu_t(j) - \sum_{j \in A_{C,-1}} \nu_t(j) \\ \sqrt{n-d}U_t^{C^c} &= \sum_{j \in A_{C^c,1}} \nu_t(j) - \sum_{j \in A_{C^c,-1}} \nu_t(j) \\ \sqrt{d}V_t^C(i) &= \sum_{j \in A_{C,1}} \psi_t(i,j) - \sum_{j \in A_{C,-1}} \psi_t(i,j) \\ \sqrt{n-d}V_t^{C^c}(i) &= \sum_{j \in A_{C^c,1}} \psi_t(i,j) - \sum_{j \in A_{C^c,-1}} \psi_t(i,j) \end{aligned}$$

and notice that U_t^C, $U_t^{C^c}$, $V_t^C(i)$ and $V_t^{C^c}(i)$ are all independent standard Gaussians for all t and i. This gives, with $v = d/n$,

$$F_t(i) = \left\{ \left| \sqrt{\frac{\rho}{1-\rho}} U_t^{C^c} + V_t^{C^c}(i) \right| \leq \sqrt{\frac{v}{1-v}} \left| \sqrt{\frac{\rho}{1-\rho}} U_t^C + V_t^C(i) \right| \right\}$$

and hence also

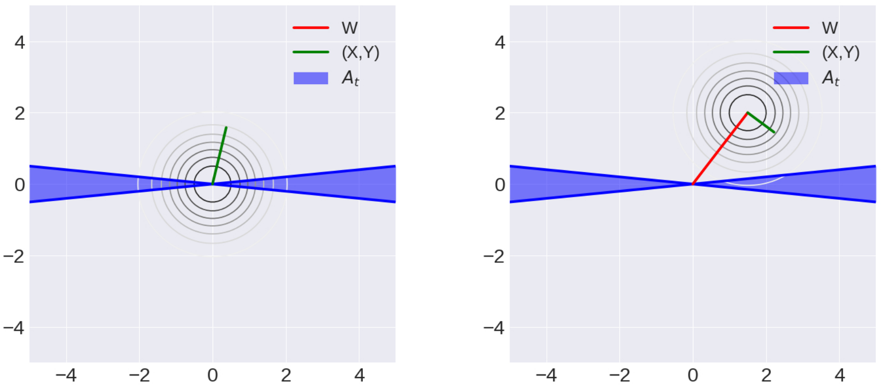
$$\mathrm{P}(F_t(i)|\omega_{t-1}, (\omega^\epsilon)_{t-1}) = \mathrm{P}\left(\left| \sqrt{\frac{\rho}{1-\rho}} U_t^{C^c} + V_t^{C^c}(i) \right| \leq \sqrt{\frac{v}{1-v}} \left| \sqrt{\frac{\rho}{1-\rho}} U_t^C + V_t^C(i) \right| \right).$$

The dependence between different i's is captured by the common variables U_t^C and $U_t^{C^c}$. Conditioning on $W_t = \sqrt{\rho/(1-\rho)}(U_t^C, U_t^{C^c})$ in addition to the condition $D_{t-1} = d$, one gets that $F_t(1), \ldots, F_t(n)$ are conditionally independent and, writing $w = (w^C, w^{C^c})$,

$$\mathrm{P}(F_t(i)|D_{t-1} = d, W_t = w) = \mathrm{P}\left(|Y| \leq \sqrt{\frac{v}{1-v}} |X| \right), \tag{6}$$

where X and Y are independent normals with unit variance and means w^C and w^{C^c} respectively. Write $g_w(v)$ for the right hand side of (6). Summing up, by letting $\tilde{D}_t = D_t/n$, we have shown that the conditional distribution $(D_t|\tilde{D}_{t-1} = v, W = w)$ is binomial with parameters n and $g_w(v)$.

Indeed given $\tilde{D}_{t-1} = v$, D_t can be determined in two steps. First sample W from the two dimensional independent Gaussian with covariance $\sqrt{\rho/(1-\rho)}I_2$. Then given $W = w$, D_t is determined as the number of independent two-dimensional Gaussian's (X_i, Y_i) with mean w and X_i and Y_i of unit variance and independent that end up in the region $A_t = \{(x, y) : |y| \leq \sqrt{\frac{v}{1-v}}|x|\}$. The probability of a given sample in the second step ending up in A_t is $g_w(v)$. From this it is easy to see that $g_w(v)$ is increasing in v for all values of w. An illustration of A_t and W is seen in figure 1.



(a) $W = (0, 0)$. Corresponds to the uncorrelated case.

(b) $W = (1.5, 2)$. One realisation of the correlated case.

Figure 1: An illustration of how D_t is sampled given $\tilde{D}_{t-1} = v$. First sample W from a independent two dimensional Gaussian with covariance matrix $\sqrt{\rho/(1-\rho)}I_2$. Then given $W = w$, D_t is determined as the number of independent Gaussian's with mean w and unit variance that end up in the region A_v.

We will need the following definition.

Definition 3.1. A sequence of Boolean function $\{h_n\}$ has a **sharp threshold at 1/2** if $\{h_n\}$ is such that for ω being i.i.d. Bernoulli(p_n) and if $\{p_n - 1/2\}$ is bounded away from 0, then $\lim_{n\to\infty} \mathrm{P}(h_n(\omega) = \mathrm{sign}(p_n - 1/2)) = 1$.

One among many examples of Boolean functions that has a sharp threshold at 1/2 is the majority function. It can be shown that this also is true for the weighted majority function when the weights $\theta_i > 0$ are such that $\lim_{n\to\infty} \frac{\max_i \theta_i}{\sqrt{n}\min_i \theta_i} = 0$.

Let $\tau_c^\epsilon = \tau_c^{\epsilon_n} = \min\{t : g_{W_t}(\tilde{D}_{t-1}) \geq c\}$ with $\tau_c^\epsilon = \infty$ if $g_{W_t}(\tilde{D}_{t-1}) < c$ for all t. The following lemma is crucial. It shows that noise sensitivity or noise stability almost entirely come down to if $\tilde{D}$ started from arbitrary small ϵ hits 1/2 before 0 (which gives rise to sensitivity results) or vice versa (giving rise to stability results). In both of these cases, the rest of the arguments come down to adopting suitable conditions on h_n to go along with that. Therefore and since we are also of the opinion that $\tilde{D}$ is a very natural Markov process that is interesting in its own right, we will later on state the results for $\tilde{D}$ explicitly along with sensitivity/stability results.

Lemma 3.2. Let $1/2 \geq \epsilon_n \downarrow 0$. Assume $\{h_n\}$ has a sharp threshold at 1/2, $\rho_n > \delta$ for some $\delta > 0$. Then the following statements hold.

(i) If for all ϵ_n, $\lim_{n\to\infty} \mathrm{P}(\tau_{1/2}^{\epsilon_n} \geq T_n) = 0$ and $\{h_n\}$ is odd, then $\{f_{n,T_n}\}$ is annealed, and hence quenched, QNS at level $\{\epsilon_n\}$.

(ii) If for all $\epsilon > 0$ sufficiently small, $\lim_{n\to\infty} \mathrm{P}(\tau_{2\epsilon}^{\epsilon} \leq T_n) = 0$, then $\{f_{n,T_n}\}$ is annealed, and hence quenched, noise stable.

Proof. Let $T = T_n$ and $f = f_{n,T}$. As before, write $\theta_t(i,j) = \sqrt{\rho}\nu_t(j) + \sqrt{1-\rho}\psi_t(i,j)$, which in row vector form becomes $\theta_t(i,\cdot) = \sqrt{\rho}\nu_t + \sqrt{1-\rho}\psi_t(i,\cdot)$, so that

$$\omega_T(i) = \mathrm{sign}\left((\sqrt{\rho}\nu_T + \sqrt{1-\rho}\psi_T(i,\cdot)) \cdot \omega_{T-1}\right).$$

This means that conditionally on ω_{T-1} and ν_T, $\omega_T(i)$ are i.i.d. Bernoulli with probability $p(\nu_T, \omega_{T-1})$ $:= \mathrm{P}(\omega_T(i) = 1|\nu_T, \omega_{T-1}) = \mathrm{P}((\sqrt{\rho}\nu_T + \sqrt{1-\rho}\psi_T(i,\cdot)) \cdot \omega_{T-1} > 0)$, which equals

$$\mathrm{P}\left(X > -\sqrt{\frac{\rho}{1-\rho}}\frac{1}{\sqrt{n}}\nu_T \cdot \omega_{T-1}\right)$$

where X is a standard normal random variable. Notice that $p(\nu_T, \omega_{T-1}) > \frac{1}{2}$ if and only if $\nu_T \cdot \omega_{T-1} > 0$. Fix $\gamma_1 > 0$. Since $\rho_n > \delta$ for some δ, there is for every $\gamma_1 > 0$ an $\gamma_2 > 0$ independently of ρ and ω_{T-1} such that for n large

$$\mathrm{P}\left(p(\nu_T, \omega_{T-1}) \in \left[\frac{1}{2} - \gamma_2, \frac{1}{2} + \gamma_2\right] \Big| \omega_{T-1}\right) < \gamma_1. \tag{7}$$

Let us now also include the process $(\omega^\epsilon)_t = (\omega^{\epsilon_n})_t$. By (6), $\mathrm{P}(F_t(i)|\tilde{D}_{t-1} = v)$ is increasing in v, which means that the conditional distribution $(\tilde{D}_t|\tilde{D}_{t-1} = v)$ is stochastically increasing in v. Hence inductively for $t \leq T-1$, $(\tilde{D}_{T-1}|\tilde{D}_{t-1} = v)$ is stochastically increasing in v. Additionally, as in the proof of Lemma 3.1, writing B_T for the event $\{\mathrm{sign}(\nu_T \cdot \omega_{T-1}) \neq \mathrm{sign}(\nu_T \cdot (\omega^\epsilon)_{T-1}\}$,

$$\mathrm{P}\left(B_T|\omega_{T-1}, (\omega^\epsilon)_{T-1}\right) = g(\tilde{D}_{T-1}) \tag{8}$$

from which it follows by symmetry that for all $\eta \in [0, 1/2]$

$$\mathrm{P}\left(B_T|\tilde{D}_t = \frac{1}{2} - \eta\right) + \mathrm{P}\left(B_T|\tilde{D}_t = \frac{1}{2} + \eta\right) = 1.$$

It follows directly that if n is even $\mathrm{P}(B_T|\tilde{D}_t = 1/2) = 1/2$ and if n is odd, since $\mathrm{P}(B_T|\tilde{D}_t = 1/2 - \eta) <$ $\mathrm{P}(B_T|\tilde{D}_t = 1/2 + \eta)$, that $\mathrm{P}(B_T|\tilde{D}_t = 1/2 + 1/2n) \geq 1/2$. Assume for simplicity for the rest of the proof of (i) that n is even; for the case with n odd just replace any conditioning on $\tilde{D}_t = 1/2$ with conditioning on $\tilde{D}_t = 1/2 + 1/2n$.

Since $(\tilde{D}_{T-1}|\tilde{D}_{t-1} = v)$ is increasing in v, we now get

$$\mathrm{P}(B_T|\tau_{1/2}^\epsilon = t) \geq \mathrm{P}\left(B_T|\tilde{D}_t = \frac{1}{2}\right) = \frac{1}{2}.$$

Now let $A_T = B_T \setminus C_T$, where

$$C_T = \{p(\nu_T, \omega_{T-1}) \in [1/2 - \gamma_2, 1/2 + \gamma_2]\} \cup \{p(\nu_T, (\omega^\epsilon)_{T-1}) \in [1/2 - \gamma_2, 1/2 + \gamma_2]\}.$$

By (7) and the fact that $A_T \subset B_T$, it now follows for $t < T$

$$\mathrm{P}\left(A_T|\tau_{1/2}^\epsilon = t\right) \geq \frac{1}{2} - \mathrm{P}(C_T|\tau_{1/2}^\epsilon = t) > \frac{1}{2} - 2\gamma_1.$$

By the assumptions on $\{h_n\}$, for sufficiently large n and $t < T$, $\mathrm{E}[f(\omega)f(\omega^\epsilon)|A_T, \tau_{1/2}^\epsilon = t] <$ $-1 + \gamma_1$.We then get

$$\begin{aligned}\mathrm{E}\left[f(\omega)f(\omega^\epsilon)|\tau^\epsilon_{1/2}=t\right] &= \mathrm{E}\left[f(\omega)f(\omega^\epsilon)|A_T,\tau^\epsilon_{1/2}=t\right]\mathrm{P}\left(A_T|\tau^\epsilon_{1/2}=t\right)\\ &+\mathrm{E}\left[f(\omega)f(\omega^\epsilon)|A_T^c,\tau^\epsilon_{1/2}=t\right]\mathrm{P}\left(A_T^c|\tau^\epsilon_{1/2}=t\right)\\ &<(-1+\gamma_1)\left(\frac{1}{2}-2\gamma_1\right)+\frac{1}{2}+2\gamma_1<5\gamma_1.\end{aligned}$$

Finally this means that for n sufficiently large,

$$\begin{aligned}&0\le \mathrm{E}\left[f(\omega)f(\omega^\epsilon)\right]\\ &=\sum_{t=0}^{T-1}\mathrm{E}\left[f(\omega)f(\omega^\epsilon)|\tau^\epsilon_{1/2}=t\right]\mathrm{P}\left(\tau^\epsilon_{1/2}=t\right)+\mathrm{E}\left[f(\omega)f(\omega^\epsilon)|\tau^\epsilon_{1/2}\ge T\right]\mathrm{P}\left(\tau^\epsilon_{1/2}\ge T\right)\\ &<5\gamma_1+\mathrm{P}\left(\tau^\epsilon_{1/2}\ge T\right).\end{aligned}$$

Since h is assumed odd, so is f and hence $\mathrm{Cov}(f(\omega),f(\omega^\epsilon))=\mathrm{E}[f(\omega)f(\omega^\epsilon)]$. Since γ_1 is arbitrary, (i) follows.

To prove (ii), fix γ_1, pick ϵ such that $g(2\epsilon)<\gamma_1$ and pick γ_2 such that (7) holds. Notice that due to (8) and the fact that g is increasing, we have

$$\mathrm{P}\left(B_T|\tilde{D}_{T-1}\le 2\epsilon\right)\le g(2\epsilon)<\gamma_1.$$

Hence by (7)

$$\mathrm{P}(B_T\cup C_T|\tilde{D}_{T-1}\le 2\epsilon)<3\gamma_1.$$

Additionally for n large and ϵ sufficiently small, due to h_n having a sharp threshold at $1/2$,

$$\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|B_T^c\cap C_T^c,\tilde{D}_{T-1}\le 2\epsilon\right)<\gamma_1.$$

This results in

$$\begin{aligned}&\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|\tilde{D}_{T-1}\le 2\epsilon\right)\\ &=\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|B_T\cup C_T,\tilde{D}_{T-1}\le 2\epsilon\right)\mathrm{P}\left(B_T\cup C_T|\tilde{D}_{T-1}\le 2\epsilon\right)\\ &+\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|B_T^c\cap C_T^c,\tilde{D}_{T-1}\le 2\epsilon\right)\mathrm{P}\left(B_T^c\cap C_T^c|\tilde{D}_{T-1}\le 2\epsilon\right)\\ &<3\gamma_1+\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|B_T^c\cap C_T^c,\tilde{D}_{T-1}\le 2\epsilon\right)<4\gamma_1.\end{aligned}$$

Since $\lim_{n\to\infty}\mathrm{P}(\tau^\epsilon_{2\epsilon}\le T)=0$, we have for n sufficiently large that $\mathrm{P}(\tilde{D}_{T-1}>2\epsilon)\le \mathrm{P}(\tau^\epsilon_{2\epsilon}\le T)<\gamma_1$.

It now follows that for n large,

$$\begin{aligned}&\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)\right)\\ &<\mathrm{P}\left(f(\omega)\ne f(\omega^\epsilon)|\tilde{D}_{T-1}\le 2\epsilon\right)+\mathrm{P}\left(\tilde{D}_{T-1}>2\epsilon\right)\\ &<5\gamma_1\end{aligned}$$

Since γ_1 is arbitrary, this concludes (ii).

□

We are now ready to start the proof that $\tilde{D}$ hits $1/2$ before 0 and $f = f_{n,T_n}$ is annealed (and thus quenched) QNS for a large range of ρ if T_n is suitably large. The results are stated in Theorem 3.3.

We are going to make several observations concerning the behaviour of $g_W(v)$. Since we will for the most parts only be considering a single t, we will in such cases drop t from the notation.

Before going on, recall that the sum of the squares of two independent standard normal random variables is exponential with mean 2. This also means that $r^2 := ||W||_2^2$ is exponential with mean $2\rho/(1-\rho)$.

Lemma 3.3. Let $r = ||w||_2$. Then

$$g_w(v) \geq \frac{2}{\pi} \arctan\left(\sqrt{\frac{v}{1-v}}\right) e^{-r^2/2} = g(v)e^{-r^2/2}.$$

Proof. Let $w = (a,b)$. Then

$$g_w(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\sqrt{\frac{v}{1-v}}|x|}^{\sqrt{\frac{v}{1-v}}|x|} e^{-((x-a)^2+(y-b)^2)/2}dydx = \frac{e^{-\frac{a^2+b^2}{2}}}{2\pi}\int_{-\infty}^{\infty}\int_{-\sqrt{\frac{v}{1-v}}|x|}^{\sqrt{\frac{v}{1-v}}|x|} e^{-(x^2+y^2)/2}e^{xa+yb}dydx$$

$$\geq \frac{e^{-\frac{r^2}{2}}}{2\pi}\int_{-\infty}^{\infty}\int_{-\sqrt{\frac{v}{1-v}}|x|}^{\sqrt{\frac{v}{1-v}}|x|} e^{-(x^2+y^2)/2}dydx = \frac{2}{\pi}\arctan\left(\sqrt{\frac{v}{1-v}}\right)e^{-\frac{r^2}{2}}$$

where the inequality follows from symmetry of the integrated function around the origin and that $e^x + e^{-x} \geq 2$. The last equality follows from the computation in the proof of Lemma 3.1. □

Lemma 3.4. Assume that $v \leq 1/2$ and that $w = (w^C, w^{C^c})$ satisfies $|w^{C^c}| \leq \sqrt{v/(1-v)}|w^C| - 2$. Then $g_w(v) > 1/2$.

Proof. Let $X \sim N(w^C, 1)$ and $Y \sim N(w^{C^c}, 1)$ be independent and write $X = w^C + \xi$ and $Y = w^{C^c} + \eta$ for independent standard normal ξ and η. We have

$$g_w(v) = \mathrm{P}\left(|Y| \leq \sqrt{\frac{v}{1-v}}|X|\right) = \mathrm{P}((X,Y) \in A_v)$$

where

$$A_v = \left\{(x,y) : |y| \leq \sqrt{\frac{v}{1-v}}|x|\right\}.$$

Since $v \leq 1/2$, it is easily seen that the L_2 distance between (w^C, w^{C^c}) and A_v^c is smaller or equal to $\sqrt{2}$ whenever $|w^{C^c}| \leq \sqrt{v/(1-v)}|w^C| - 2$. Hence

$$g_w(v) \geq \mathrm{P}(\xi^2 + \eta^2 \leq 2) = 1 - e^{-1} > \frac{1}{2}.$$

□

Lemma 3.5. Let as above $W = W_t = (W^C, W^{C^c})$ for $t < T$, where W^C and W^{C^c} are independent normals with means 0 and variance $\rho/(1-\rho)$. Then for $v \leq 1/2$,

$$\mathrm{P}\left(g_W(v) > \frac{1}{2}\right) > \frac{1}{2\pi}\sqrt{\frac{v}{1-v}}e^{-\frac{8(1-\rho)}{\rho}\frac{1-v}{v}}.$$

Proof. By Lemma 3.4, $\mathrm{P}(g_W(v) > 1/2) \geq \mathrm{P}(|W^{C^c}| \leq \sqrt{v/(1-v)}|W^C| - 2)$. Writing (r, φ) for the polar coordinates of W, it is straightforward by back substitution to see that it is sufficient for $|W^{C^c}| \leq \sqrt{v/(1-v)}|W^C| - 2$ that

$$r^2 > 16(1-v)/v$$

and

$$|\varphi| < \arctan\left(\frac{1}{2}\sqrt{\frac{v}{1-v}}\right)$$

and the latter in turn occurs whenever $|\varphi| < \frac{1}{4}\sqrt{v/(1-v)}$. Since r^2 and $|\varphi|$ are independent and exponential$((1-\rho)/2\rho)$ and uniform on $[0, \pi/2]$ respectively, we thus get

$$\mathrm{P}\left(g_W(v) > \frac{1}{2}\right) > e^{-\frac{8(1-\rho)}{\rho}\frac{1-v}{v}}\frac{2}{\pi}\frac{1}{4}\sqrt{\frac{v}{1-v}}$$

which is the desired bound. □

We are now ready to state the first main results of this section.

Theorem 3.3. *Let $\{\epsilon_n\}$ be such that $1/2 \geq \epsilon_n \downarrow 0$ and $n\epsilon_n \to \infty$. Assume that for some $\delta > 0$ independent of n, $\delta < \rho < 1 - \frac{4(\log\log n)^3}{\log n}$ and $T_n \geq e^{4(\log\log n)^2}$. Then the following statements hold.*

(i) $\lim_{n\to\infty} \mathrm{P}(\tau_{1/2}^{\epsilon_n} > T_n) = 0$,

(ii) If h_n is odd and has a sharp threshold at 1/2, then $\{f_{n,T_n}\}$ is annealed QNS and hence also quenched QNS at level $\{\epsilon_n\}$.

(iii) If h_n is odd and T_n grows sufficiently large with n, then $\{f_{n,T_n}\}$ is annealed QNS and hence also quenched QNS at level $\{\epsilon_n\}$.

As in the uncorrelated case, we can prove even stronger versions.

Theorem 3.4. *Let $\{\epsilon_n\}$ be such that $1/2 \geq \epsilon_n \downarrow 0$ and $n\epsilon_n \to \infty$. Assume that for some $\delta > 0$ independent of n, $\delta < \rho < 1 - \frac{4(\log\log n)^3}{\log n}$ and $T_n \geq e^{4(\log\log n)^2}$. Then the following statements hold.*

(i) Fix $\omega, \eta \in \{-1,1\}^n$ arbitrarily such that $\eta \notin \{\omega, -\omega\}$. Then if h_n is odd and either has a sharp threshold at 1/2 or T_n is sufficiently large, then

$$\lim_{n\to\infty} \mathrm{Cov}_\Theta(f_{n,T_n}(\omega), f_{n,T_n}(\eta)) = 0.$$

(ii) Let $\mathbf{Q}_n$ be any probability measure on $\{-1,1\}^n \times \{-1,1\}^n$ such that $\lim_{n\to\infty} \mathbf{Q_n}(\eta \in \{\omega, -\omega\}) = 0$, Then if h_n is odd and either has a sharp threshold at 1/2 or T_n is sufficiently large, then

$$\lim_{n\to\infty} \mathrm{Cov}_{\mathbf{Q_n},\Theta}(f_{n,T_n}(\omega), f_{n,T_n}(\eta)) = 0.$$

(iii) Assume that h_n is odd and either has a sharp threshold at 1/2 or T_n is sufficiently large. Fix any $k \in \{1, 2, \ldots, n-1\}$ and $\delta > 0$. Fix also $\omega \in \{-1,1\}^n$ and let $M_k = M_k^{(n)}(\omega)$ be the number of η with $\eta(i) \neq \omega(i)$ for exactly k indexes i, such that $f_{n,T_n}(\eta) \neq f_{n,T_n}(\omega)$. Then for $T_n \geq K_n$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{M_k}{\binom{n}{k}} \notin \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right)\right) = 0.$$

Proof of Theorem 3.3. Let us first outline the strategy of the proof of (i). We will run the $\tilde{D}$-process for a predetermined time s. In doing so, we will prove that (a) regardless of $\epsilon \in (0, 1/2]$, the probability that $\tilde{D}$ hits 0 during time s is very small regardless of the value of $\tilde{D}_0$ as long as it is nonzero, and (b) the probability that $\tilde{D}$ hits 1/2 during time s is of much higher order. Having done that allows us to repeatedly run the process for s units of time and use the Markov property to draw the conclusion that with very high probability, $\tilde{D}$ hits 1/2 before 0 for sufficiently fast growing T_n. The final part (c) upper bounds the time it takes to hit either 1/2 or 0.

To avoid confusion we point out that it will be easy for the reader to see that many of the bounds given are far from optimal and thus to some extent arbitrary; they are simply good enough for their purpose.

Let $s = s_n = \log_2 \log n$ and fix $\epsilon > 0$. We start by giving an upper bound on $P(\tilde{D}_s = 0)$. We have for large n and all $t < T_n$,

$$\begin{aligned}
\mathrm{P}(\tilde{D}_{t+1} = 0 | \tilde{D}_t > 0) &\leq \mathrm{P}\left(\tilde{D}_{t+1} = 0 \middle| \tilde{D}_t = \frac{1}{n}\right) \\
&\leq \mathrm{P}\left(\frac{||W||_2^2}{2} > \frac{2\log n}{\log\log n}\right) + \left(1 - g\left(\frac{1}{n}\right) e^{-\frac{2\log n}{\log\log n}}\right)^n \\
&< e^{-\frac{4(\log\log n)^3}{\log n}\frac{2\log n}{\log\log n}} + \left(1 - \frac{1}{n^{2/3}}\right)^n \\
&< e^{-8(\log\log n)^2} + e^{-n^{1/3}} \\
&< e^{-7(\log\log n)^2},
\end{aligned}$$

where the second inequality uses Lemma 3.3 for the second term. By the Markov property of $\tilde{D}$ and Bonferroni, we get

$$\mathrm{P}(\tilde{D}_s = 0) < (\log_2 \log n) e^{-7(\log\log n)^2} < e^{-6(\log\log n)^2}. \tag{9}$$

This concludes part (a) in the sketch.

Next, we lower bound $\mathrm{P}(g_{W_s}(D_{s-1}) > 1/2)$. There is $\kappa > 0$ such that for any $v \in [1/n, 1/2)$,

$$\begin{aligned}
\mathrm{P}\left(\tilde{D}_{t+1} \geq \frac{\sqrt{v}}{20} \middle| \tilde{D}_t = v\right) &\geq \mathrm{P}\left(g_W(v) > \frac{\sqrt{v}}{20}\right) \mathrm{P}\left(\tilde{D}_{t+1} \geq \frac{\sqrt{v}}{20} \middle| g_W(v) > \frac{\sqrt{v}}{10}\right) \\
&> \mathrm{P}\left(\frac{||W||_2^2}{2} \leq 1\right)\left(1 - e^{-\kappa\sqrt{n}}\right) \\
&= \left(1 - e^{-\frac{1-\rho}{\rho}}\right)\left(1 - e^{-\kappa\sqrt{n}}\right) \\
&> 1 - e^{-\frac{1}{2}\frac{4(\log\log n)^3}{\log n}} \\
&> \frac{(\log\log n)^3}{\log n}.
\end{aligned}$$

Here the second inequality is Lemma 3.3 and Chernoff bounds. This gives, provided that $\tilde{D}_0 \geq 1/n$,

$$\mathrm{P}\left(\forall t = 0, \ldots, s-2 : \tilde{D}_{t+1} \geq \min\left(\frac{1}{2}, \frac{\sqrt{\tilde{D}_t}}{20}\right)\right) > \left(\frac{(\log\log n)^3}{\log n}\right)^{\log_2 \log n} > e^{-2(\log\log n)^2}. \tag{10}$$

If the event in the left hand side occurs, then

$$\tilde{D}_{s-1} \geq \frac{1}{400}\left(\frac{1}{n}\right)^{2^{-\log_2 \log n}} = \frac{1}{400}e^{-1} > \frac{1}{1200}.$$

By Lemma 3.5,

$$\mathrm{P}\left(g_{W_s}(\tilde{D}_{s-1}) > \frac{1}{2} \middle| \tilde{D}_{s-1} > \frac{1}{1200}\right) > \frac{1}{40\pi}e^{-\frac{9600(1-\rho)}{\rho}} \geq \frac{1}{40\pi}e^{-\frac{9600(1-\delta)}{\delta}} =: a.$$

Combining with (10), it follows that

$$\mathrm{P}\left(g_{W_s}(\tilde{D}_{s-1}) > \frac{1}{2}\Big|\tilde{D}_0 > 0\right) > a\, e^{-2(\log\log n)^2} > e^{-3(\log\log n)^2}. \tag{11}$$

Comparing (9) and (11) we see that the conditional probability that $g_{W_s}(\tilde{D}_{s-1})$ reaches $1/2$ before absorbing at zero given $\tilde{D}_0 > 0$ is at least $1 - e^{-3(\log\log n)^2}$, finishing part (b) of the sketch. Finally the time to either hitting $1/2$ or absorbing at 0 is dominated by a geometric random variable with parameter $e^{-3(\log\log n)^2}$, so with probability going to 1 this will happen before time T_n whenever $T_n > e^{4(\log\log n)^2}$. Adding the fact that since $\epsilon_n > 1/n$ for n large, $\mathrm{P}(\tilde{D}_0 = 0) < e^{-\sqrt{n}}$, this concludes the proof of (i). Now (ii) immediately follows from part (i) of Lemma 3.2.

For (iii), observe that by (i) and since $\tilde{D}$ absorbs when it hits 0 or 1, we get by symmetry of $\tilde{D}$ that $\lim_{n\to\infty}\mathrm{P}(\tilde{D}_{T_n} = 0) = \lim_{n\to\infty}\mathrm{P}(\tilde{D}_{T_n} = 1) = 1/2$ for $\{T_n\}$ growing sufficiently large. Since h_n is odd (iii) follows for such $\{T_n\}$.

□

Proof of Theorem 3.4. The proof is very similar to the proof of Theorem 3.2. A quick glance back at the proof of (i) in Theorem 3.3 shows that $\tilde{D}$ with high probability hits $1/2$ before 0 also if the input $(\omega, \omega^{\epsilon_n})$ is replaced with (ω, η). Now (i) follows from exactly the same proof as that of part (i) of Lemma 3.2. Finally, both (ii) and (iii) follows from (i) in the same way as (ii) and (iii) follow from (i) in Theorem 3.2. □

Next we move to proving that for sufficiently large ρ, $\{f_{n,T_n}\}$ is noise stable. First we need an upper bound on $g_w(v)$:

Lemma 3.6. Let $w \in \mathcal{R}^2$ and let (r,θ) be its polar coordinates and let $\varphi = \arctan\sqrt{v/(1-v)}$. Assume that $|\theta| \geq \varphi$ and $|\theta - \pi| \geq \varphi$. Then

$$g_w(v) \leq e^{-\frac{1}{2}r^2\sin^2(|\theta|-\varphi)}.$$

Proof. As in the proof of Lemma 3.4, let

$$A_v = \left\{(x,y) : |y| \leq \sqrt{\frac{v}{1-v}}|x|\right\},$$

so that

$$g_w(v) = \mathrm{P}((w^C + \xi, w^{C^c} + \eta) \in A_v),$$

where ξ and η are independent standard normal. Without loss of generality, we assume that $\theta \in [0, \pi/2]$. By assumption, $w \notin A_v$. and the Euclidean distance from w to A_v is $r\,\sin(\theta - \varphi)$, so for $w + (\xi,\eta)$ to be in A_v, it is necessary that $\xi^2 + \eta^2 \geq r^2\sin^2(\theta-\varphi)$. The left hand side is exponential with mean 2, so

$$g_w(v) \leq e^{-\frac{1}{2}r^2\sin^2(\theta-\varphi)}.$$

□

Theorem 3.5. *Assume that $\rho > 1 - \log\log n/18\log n$. Then the following holds.*

(i) For $T = T_n \geq (\log n)^{1/4}$ and for all $\delta > 0$, it holds for all sufficiently small $\epsilon > 0$ that if $\tilde{D}_0 = v < \epsilon^2/2$, then

$$\limsup_n \mathrm{P}\left(\{\tilde{D}_T > 0\} \cup \left\{\exists t \in \{1,\ldots,T\} : \tilde{D}_t > \frac{\epsilon^2}{2}\right\}\right) < \delta.$$

In particular,

$$\limsup_n \mathrm{P}\left(\exists t \geq 1 : \tilde{D}_t > \frac{\epsilon^2}{2}\right) < \delta.$$

(ii) f_{n,T_n} *is annealed and quenched noise stable if either* h_n *has a sharp threshold at 1/2 or* $T_n \geq (\log n)^{1/4}$.

Proof. This proof will start with showing that with high probability, $\tilde{D}$ immediately drops to less than $1/(\log n)^{1/2}$ and then stays there for a time of order at least $(\log n)^{1/16}$. Having done that, it will be shown that during that time, $\tilde{D}$ will in fact have hit 0. As in previous proofs, many inequalities used are obviously far from optimal, but simply good enough for their purpose.

Consider now the first step of $\tilde{D}$. Write (r, θ) for the polar coordinates of $W = W_1$. We have by Lemma 3.6, since $\arctan \sqrt{(\epsilon^2/2)/(1-(\epsilon^2/2)^2)} < \epsilon$

$$\begin{aligned}
\mathrm{P}\left(g_W(v) < e^{-\frac{18\epsilon^3 \log n}{\log\log n}}\right) &> \mathrm{P}\left(g_W\left(\frac{\epsilon^2}{2}\right) < e^{-\frac{18\epsilon^3 \log n}{\log\log n}}\right) \\
&> \mathrm{P}\left(e^{-\frac{1}{2}r^2 \sin^2(|\theta|-\epsilon)} < e^{-\frac{18\epsilon^3 \log n}{\log\log n}}\right) \\
&> \mathrm{P}\left(|\theta| > 3\epsilon, |\theta - \pi| > 3\epsilon, \frac{1}{2}r^2 > \frac{18\epsilon \log n}{\log\log n}\right) \\
&> \left(1 - \frac{18\epsilon}{\pi}\right) e^{-\epsilon} \\
&> 1 - 4\epsilon
\end{aligned}$$

and given that $g_W(v) < e^{-\frac{18\epsilon^3 \log n}{\log\log n}}$, we get for large n,

$$\tilde{D}_1 < 2e^{-\frac{18\epsilon^3 \log n}{\log\log n}} < \frac{1}{(\log n)^{1/2}}$$

with conditional probability at least $1 - \epsilon$. Taken together, the last two inequalities give,

$$\mathrm{P}\left(\tilde{D}_1 \geq \frac{1}{(\log n)^{1/2}}\right) < 5\epsilon. \tag{12}$$

Let $E_t = \{\tilde{D}_t < 1/(\log n)^{1/2}\}$; this notation will be convenient as we are going to do much conditioning on this event from here on.

Next consider the distribution of $\tilde{D}_{t+1}$ given $\tilde{D}_t = \alpha^2/2 < 1/(\log n)^{1/2}$. Let as before $(r, \theta) = (r_t, \theta_t)$ be the polar coordinates of W_t, $t = 1, 2, \ldots$.

Let

$$B = \left\{e^{-\frac{1}{2}r^2 \sin^2(|\theta|-\alpha)} < \frac{1}{2(\log n)^{1/2}}\right\}.$$

It is sufficient for B to occur that $|\theta| \geq 3/(\log n)^{1/4}$ and $r^2 \geq (\log n)^{1/2} \log\log n$. Since $1 - \rho < \log\log n/300 \log n$, this gives for large n,

$$\mathrm{P}\left(B \,\middle|\, \tilde{D}_t = \frac{\alpha^2}{2}\right) = \mathrm{P}(B) > \left(1 - \frac{6}{\pi(\log n)^{1/4}}\right) e^{-\frac{\log\log n}{36 \log n}(\log n)^{1/2} \log\log n} > 1 - \frac{3}{(\log n)^{1/4}}. \tag{13}$$

When B occurs,

$$g_W\left(\frac{\alpha^2}{2}\right) < \frac{1}{2(\log n)^{1/2}}$$

according to Lemma 3.6. Hence for n large by Chernoff bounds,

$$\mathrm{P}\left(\tilde{D}_{t+1} < \frac{1}{(\log n)^{1/2}} \middle| \tilde{D}_t = \frac{\alpha^2}{2}, B\right) > 1 - e^{-\frac{n}{\log n}} > 1 - \frac{1}{(\log n)^{1/4}}.$$

Since $\alpha^2/2$ is an arbitrary number in $[0, 1/(\log n)^{1/2})$, we get

$$\mathrm{P}\left(\tilde{D}_{t+1} < \frac{1}{(\log n)^{1/2}}\Big|\tilde{D}_t < \frac{1}{(\log n)^{1/2}}, B\right) > 1 - \frac{1}{(\log n)^{1/4}}$$

and hence on combining with (13),

$$\mathrm{P}\left(\tilde{D}_{t+1} < \frac{1}{(\log n)^{1/2}}\Big|E_t\right) > 1 - \frac{4}{(\log n)^{1/4}}. \tag{14}$$

Next let $A = A_{t+1} = \{|\theta_{t+1}| - \epsilon > \pi/4, r_{t+1}^2 > 6 \log n\}$. Then since $1 - \rho < \log\log n/18 \log n$, provided that $\epsilon < \pi/12$,

$$\mathrm{P}(A) > \frac{1}{3}e^{-\frac{\log\log n}{6}} > \frac{2}{(\log n)^{1/5}}.$$

If A occurs, Lemma 3.6 implies for $v < \epsilon^2/2$

$$g_W(v) < e^{-\frac{1}{4}r^2} < e^{-\frac{3}{2}\log n} = \frac{1}{n^{3/2}}.$$

That entails in turn that

$$\mathrm{P}(\tilde{D}_{t+1} = 0|A) > \left(1 - \frac{1}{n^{3/2}}\right)^n > 1 - \frac{1}{\sqrt{n}}.$$

Hence

$$\begin{aligned}\mathrm{P}\left(\tilde{D}_{t+1} = 0|E_t\right) &> \mathrm{P}\left(\tilde{D}_{t+1} = 0|A \cap E_t\right)\mathrm{P}\left(A|E_{t-1}\right) = \mathrm{P}\left(\tilde{D}_{t+1} = 0|A\right)\mathrm{P}\left(A\right) \\ &> \frac{2}{(\log n)^{1/5}}\left(1 - \frac{1}{\sqrt{n}}\right) > \frac{1}{(\log n)^{1/5}}.\end{aligned} \tag{15}$$

By taking (14) and (15) together, it now easily follows that the conditional probability given E_1 that $\tilde{D}$ for $t \geq 1$ hits 0 before $(\log n)^{1/2}$ exceeds $1 - 1/(\log n)^{1/4-1/5} = 1 - 1/(\log n)^{1/20}$ and with probability exceeding $1 - e^{-(\log n)^{1/20}}$ it happens before time $(\log n)^{1/4}$.

Combining this with (12) and taking $\epsilon < \delta/6$ we get

$$\limsup_n \mathrm{P}\left(\{\tilde{D}_{T_n} > 0\} \cup \{\exists t \in \{1, \ldots, T_n\} : \tilde{D}_t > \frac{1}{(\log n)^{1/2}}\}\right) < \delta.$$

This obviously implies

$$\limsup_n \mathrm{P}\left(\{\tilde{D}_{T_n} > 0\} \cup \{\exists t \in \{1, \ldots, T_n\} : \tilde{D}_t > \epsilon\}\right) < \delta.$$

For (ii) assume first that h_n has a sharp threshold at $1/2$ and T_n is arbitrary. Fix $\kappa > 0$. Then (i) tells us that for sufficiently small $\epsilon > 0$ and sufficiently large n, $P(\forall t : \tilde{D}_t < 2\epsilon) > 1 - \kappa/2$. In particular

$$\limsup_n \mathrm{P}(\tau_{2\epsilon}^{\epsilon} \leq T_n) < \frac{\kappa}{2}.$$

Since the final output is $f_{n,T_n}(\omega) = h_n(\omega_T)$ and h_n has a sharp threshold at $1/2$, the result follows from Lemma 3.2.

Finally if $T_n \geq (\log n)^{1/4}$, then by (i), $\lim_{\epsilon\to 0} \mathrm{P}(\tilde{D}_{T_n} = 0) = 1$, i.e. $\lim_{\epsilon\to 0} \mathrm{P}(\omega_{T_n} = (\omega^\epsilon)_{T_n}) = 1$, which implies noise stability.

□

4 Convolutional treelike networks

Above we have studied a Boolean representation of randomised feed forward neural networks. Another common architecture of neural networks is convolutional neural networks, ConvNet, where the output of each layer is the convolution between the input and some filter of size d through some activation function. The learnable parameters in the model will be those in the filter. If we study a one dimensional input and if no padding is used, the output of each convolution is of length $(n-d)/s+1$ where s is the stride and n the number of inputs. As previously, we consider the activation function to be the sign function. We also limit ourselves to a filter size of length $d=3$ (but the technique used will with a few observations easily generalise to some other settings, which we point out and state at the end of this section). Stacking many of these convolutional layers our networks can be seen as a graph $G_n=(V_n,E_n)$. From this graph we induce a Boolean function f_n that goes from $\{-1,1\}^n$ to {-1,1}. The graphs that we consider here, and soon give proper definitions of, can be seen in Figure 2.

With filter size 3, the value of each node at layer t in G_n is the weighted 3-majority function, where the weights are the parameters in the filter θ_t. These weights are the same for all triples of nodes next to each other on the same layer on which the filter is applied. A weighted 3-majority function can only express either a regular 3-majority, negative 3-majority, a dictator function or a negative dictator function. Clearly, reversing the signs of all values in a layer does not have anything to do with stability questions, so we may assume that the filter at a given layer either expresses a dictator or a regular majority. In each setting analysed here, we will at first assume that each layer expresses a regular majority, i.e. $\theta_t=(1,1,1)\ \forall t$. It will then be easy to see that the arguments used can be easily extended to the setting where some layers are dictator layers under the very mild assumption that the distribution of the θ_t's is such that there is a probability bounded away from 0 that θ_t expresses a regular majority. Each θ_t is here assumed to be independent across t. First, notice that if the stride $s>2$, G_n would be the iterated 3-majority function with no overlap, which is known to be sensitive, so we will only consider $s=1$ and $s=2$.

This leaves us with four different structures of interest. There corresponding graph G_n are illustrated in Figure 2 and have the following formal definitions of them:

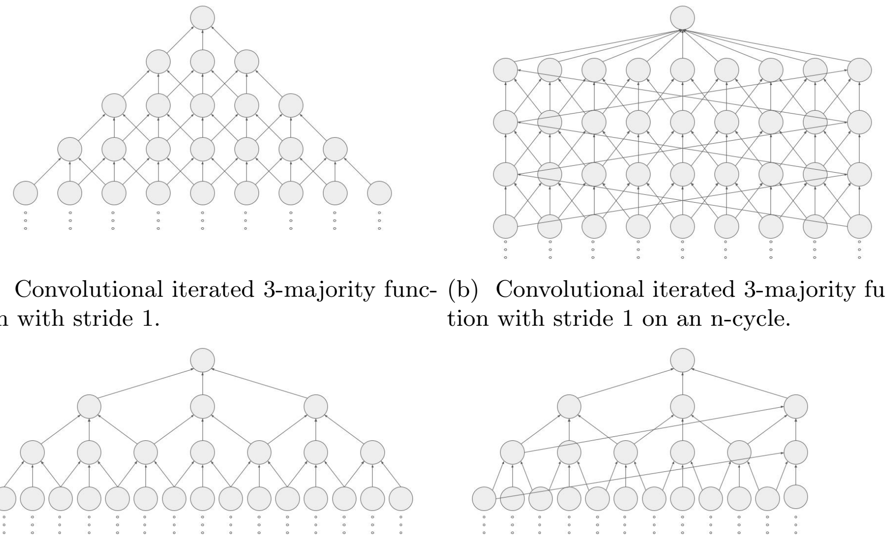
(a) Convolutional iterated 3-majority with stride 1, $G_n^{(1)}=(V,E)$, where $V=\{v_{n,0},v_{n-1,-1},v_{n-1,0},v_{n-1,1},v_{n-2,-2},v_{n-2,-1},v_{n-2,0},v_{n-2,1},v_{n-2,2},\dots,v_{0,-n},\dots,v_{0,n}\}$ and $E=\{(v_{k,i},v_{k-1,j}):k=n,\dots,1,\ j=i-1,i,i+1\}$.

(b) Convolutional iterated 3-majority with stride 1 on an n-cycle, $G_n^{(1')}=(V,E)$, where $V=\{v_{K,-n},v_{K-1,-(n-1)},\dots,v_{K-1,n},\dots,v_{0,-n},\dots,v_{0,n}\}$, $E=\{(v_{K,1},v_{K-1,i}):i=1,\dots,n\}\cup\{v_{k,i},v_{k-1,j}:k=K-1,\dots,1,\ |i-j|\le 1 \mod n\}$.

(c) Convolutional iterated 3-majority with stride 2, $G_n^{(2)}=(V,E)$, where $V=\{v_{n,1},v_{n-1,1},v_{n-1,2},v_{n-1,3},\dots,v_{0,1},\dots,v_{0,2^{n+1}-1}\}$, $E=\{(v_{n-k,i},v_{n-k-1,j}):k=0,\dots,n-1,\ |j-2i|\le 1\}$.

The graph in (d) will not be treated any more than that in the end it will be obvious that the results for (c) are valid there too, so the formal definition is skipped.

Given one of these G_n, let f_n be the Boolean function induced by G_n where the value of each node v, $f(v)$, is evaluated as the majority of the evaluation of the three connected nodes below it (where "below" refers to the standard tree convention of thinking of nodes $v_{k+1,i}$ as sitting immediately below the nodes $v_{k,i}$). The nodes of layer 0 are considered to output the input ω into the ConvNet, where $\omega(j)$, $j\in L_0$, corresponds to bit j in the Boolean input vector. For example, for a node $v_{k,i}\in G_n^1$, $f(v_{k,i})$ can be evaluated recursively using $f(v_{k,i})=\operatorname{sign}(f(v_{k-1,i})+f(v_{k-1,i+1})+f(v_{k-1,i+2}))$ and $f(v_{0,j})=\omega(j)$, $\forall k,j:\ 1\le k\le n,\ 1\le j\le 2n+1$. We will denote the set of nodes at layer k as L_k, $k=0,\dots,n$. In terms of the formal definition, $L_k=\{v_{k,i}:v_{k,i}\in V\}$. Note that the layers/generations are numbered bottom up, which is unconventional in the graph sense, but is the "right thing" to do in the neural net sense. For any i such that $v_{k,i}$ and $v_{k,i+1}$ are both in V, we say that those two vertices



(a) Convolutional iterated 3-majority function with stride 1.

(b) Convolutional iterated 3-majority function with stride 1 on an n-cycle.

(c) Convolutional iterated 3-majority function with stride 2.

(d) Convolutional iterated 3-majority function with stride 2 on an n-cycle.

Figure 2: The graphs $G_n^{(1)}$,$G_n^{(1')}$,$G_n^{(2)}$ and $G_n^{(2')}$.

are next to each other or a closest pair or, sometimes, neighbours in L_k even though they are strictly speaking not neighbours in G_n. For $G_n^{(1')}$ and $G_n^{(2')}$, the nodes $v_{k,1}$ and $v_{k,n}$ are also said to be next to each other.

For each node $u \in G_n$, in layer k, let D_u denote its set of descendants, i.e. the set of nodes in layers $k-1, \dots, 0$ defined recursively that the descendants in L_{k-1} are u's three neighbours (in the graph sense) in L_{k-1} and then the descendants in L_{k-j} are the set nodes there that have an edge to at least one descendant in L_{k-j-1}. Observe that to determine $f(u)$ it is sufficient to study the subgraph D_u. When v is a descendant of u, we equivalently say that u is an ancestor of v. Let A_u be the set of ancestors of u, i.e. A_u is the set of nodes v such that $u \in D_v$. Write $D_{u,k} = D_u \cap L_k$, and $A_{u,k} = A_u \cap L_k$, and write $A^+_{u,k}$ for the union of $A_{u,k}$ and the set of nodes in L_k that are next to a node in $A_{u,k}$. If $v \in A_u$, we say that v is a parent of u or that u is a child of v if the graphical distance between u and v is 1. Also define $D_\Lambda = \cup_{u\in\Lambda} D_u$ for a set of nodes Λ.

The following subsections prove the different noise properties of f_n induced by the different convolutional graphs.

4.1 Convolutional iterated 3-majority with stride 1

Let $G_n = G_n^{(1)} = (V_n, E_n)$ i.e. the 3-iterated majority network with stride 1, Figure 2a. Then the induced Boolean functions $\{f_n\}$ requires $N = 2n+1$ input bits which we label in accordance with how the network is defined: $\omega(-n), \dots, \omega(n)$.

Write $f = f_n$. The key observation to make is that if $f(v_{0,i}) = \omega(i)$ and $f(v_{0,i+1}) = \omega(i+1)$ are equal, then $f(v_{t,i}) = f(v_{t,i+1})$ for all t such that $v_{t,i}$ and $v_{t,i+1}$ exist. If only one of them, say $v_{t,i}$ exists, then $f(v_{t,i}) = f(v_{t-1,i}) = f(v_{t-1,i+1})$. In particular if $\omega(-1) = \omega(0)$ or $\omega(i+1)$ or $\omega(0) = \omega(1)$, then $f(\omega) = f(v_{n,0}) = \omega(0)$. Obviously this generalises the statement that if $f(v_{s,0})$ and $f(v_{s,1})$ are equal, then $f(v_{t,0}) = f(v_{t,1})$ for all $t \geq s$.

If we instead have $\omega(-1) \neq \omega(0) \neq \omega(1)$ and $\omega(2) = \omega(1)$, then $f(v_{1,0}) = f(v_{1,1}) = \omega(1) = \omega(2)$

and it then follows in the same way that $f(\omega) = \omega(1)$.

An inductive structure suggests itself. Let K be the smallest positive integer such that either $\omega(-K) = \omega(-(K-1)) \neq \omega(-(K-2) \neq \ldots \neq \omega(K-1)$ or $\omega(-(K-1)) \neq \omega(-(K-2) \neq \ldots \neq \omega(K-1) = \omega(K)$, if such an K exists. In other words K is the distance from 0 to a closest pair of input bits with the same value. (There may be two such pairs. If so, by parity $\omega(-K) = \omega(-(K-1)) = \omega(K-1) = \omega(K)$ with the input alternating between $-(K-1)$ and $K-1$.) If no such i exists, set $K = n+1$.

Assume without loss of generality that, unless $K = n+1$, $\omega(K-1) = \omega(K)$. It follows from the definition of K, $\omega(-(K-1)) \neq \omega(-(K-2)) \neq \ldots \neq \omega(K-2) \neq \omega(K-1)$, This clearly holds also for $K = n+1$. If $K \leq n$, we then get $f(v_{1,-(K-2)}) \neq f(v_{1,-(K-3)}) \neq \ldots \neq f(v_{1,K-3}) \neq f(v_{1,K-2}) = f(v_{1,K-1})$. This means that in layer 1, the closest pair of input bits with the same value is one step closer to 0 than in layer 0. It now follows from induction that $f_n(\omega) = f(v_{n,0}) = \omega(K)$. Of particular importance here is to observe that in particular $f_n(\omega)$ does not depend on $\omega(j)$, $j \notin [-K, K]$.

It remains to understand what happens if $K = n+1$, i.e when the whole input is alternating: $\omega(-n) = \omega(-(n-1)) = \ldots \neq \omega(n-1) \neq \omega(n)$. However then f will be alternating at all layers and $f(\omega) = \omega(-n) = \omega(n)$.

We have established the following lemma.

Lemma 4.1. (Closest pair lemma)

The induced Boolean function $f =$ "iterated 3-majority with stride 1" on $G_n^{(1)}$ is the same as the function g with the following description:

Given the input bit vector ω of length $2n+1$, let K be the smallest positive integer i such that $\omega(-i) = \omega(-(i-1))$ or $\omega(i-1) = \omega(i)$ and set $g(\omega) = \omega(-K)$ or $g(\omega) = \omega(K)$ in the respective cases. If no such pair exists, set $g(\omega) = \omega(n)$.

With this lemma we can state the following theorem.

Theorem 4.1. *The iterated 3-majority function with stride* 1 *is noise stable.*

Proof. According to Lemma 4.1 we can translate the iterated 3-majority with stride 1 function to the closest pair function.

Let K be the distance between the closest pair and the mid input vertex 0. Since each bit is i.i.d., K is geometric with parameter 3/4 truncated at $n+1$:

$$\mathrm{P}(K = k) = \frac{3}{4}\left(\frac{1}{4}\right)^{k-1} = 3\left(\frac{1}{4}\right)^{k},$$

$k = 1, \ldots, n$ and $\mathrm{P}(K = n+1) = (1/4)^n$.

Since f_n does not depend on ω outside $[-K, K]$, for $f_n(\omega^\epsilon) \neq f_n(\omega)$ to hold there must at least be a disagreement in the interval $[-K, K]$, i.e. a $j \in [-K, K]$ with $\omega(j) \neq \omega^\epsilon(j)$. This gives

$$\mathrm{P}(f_n(\omega) \neq f_n(\omega^\epsilon) | K = k) \leq 1 - (1-\epsilon)^{2k+1},$$

$k = 1, \ldots, n+1$ and $\mathrm{P}((f_n(\omega) \neq f_n(\omega^\epsilon) | K = n+1) \leq 1 - (1-\epsilon)^{2n+1}$.

Putting this into Definition (2.2) we see that

$$\begin{aligned}
\lim_{\epsilon \to 0} \limsup_n \mathrm{P}(f_n(\omega^\epsilon) \neq f_n(\epsilon)) &\leq \lim_{\epsilon \to 0} \limsup_n \sum_{k=1}^{n} 3\left(\frac{1}{4}\right)^k \left(1 - (1-\epsilon)^{2k+1}\right) \\
&\leq \lim_{\epsilon \to 0} \sum_{k=1}^{\infty} 3\left(\frac{1}{4}\right)^k \left(1 - (1-\epsilon)^{2k+1}\right) \\
&= \lim_{\epsilon \to 0} 3\left[\frac{4}{3} - \frac{4(1-\epsilon)^4}{4 - (1-\epsilon)^2}\right] = 0
\end{aligned}$$

This concludes the proof.

□

Let us now consider when the filter weights θ_t are random. Then θ_t represents either a regular majority or a dictator function. If θ_t represents a dictator function each node u at layer t only depends on one node v at layer $t-1$ and the values $f_n(u)$ and $f_n(v)$ are the same. This means we can effectively skip that layer on noticing that either the leftmost or the rightmost node of layers $t-1, t-2, \ldots, 0$ can no longer affect anything and can be removed. Removing all dictator layers results in an ordinary stride 1 iterated 3-majority with fewer layers. Hence Theorem 4.1 applies and we can conclude the following.

Corollary 4.1. *The iterated 3-majority function with stride 1 is annealed and quenched noise stable under any probability distribution on θ_t.*

4.2 Convolutional iterated 3-majority with stride 1 on the n-cycle

Consider now $G_n^{(1')}$, Figure 2b, the network where each layer has n nodes with circular convention so that the first node at each layer is considered next to the last node in that layer. We assume that n is odd. Each output from node i in layer t is the majority of nodes $i-1$, i, $i+1$ in layer $t-1$. The final output of the network is the majority of ω_T for some $T = T_n$. The starting configuration is $\omega_0 \in \{-1, 1\}^n$.

Theorem 4.2. *For any $T = T_n$, the convolutional iterated 3-majority function with stride 1 on an n-cycle is noise stable.*

Proof. Let $\ldots, X(-1), X(0), X(1), X(2), \ldots$ be independent Bernoulli(1/2) random variable and model $(\omega, \omega^\epsilon)$ in terms of these by taking $\omega = \omega_0 = (X(1), \ldots, X(n))$ and $\omega^\epsilon = (X^\epsilon(1), \ldots, X^\epsilon(n))$. Let $\tau_0 = \min\{i > 1 : X(i-1) = X(i) = X^\epsilon(i-1) = X^\epsilon(i)\}$. Then recursively, let

$$\tau_j = \min\{i > \tau_{j-1} + 1 : X(i) = X(i-1) = X^\epsilon(i-1) = X^\epsilon(i) = -X(\tau_{j-1})\},\ j = 1, 2, \ldots.$$

Observe that if one writes $X_0 = X$, one can then define $X_1, X_2, \ldots$ in complete analogy with how $\omega_1, \omega_2, \ldots$ are defined, i.e. $X_{t+1}(i) = \operatorname{sign}(X_t(i-1) + X_t(i) + X_t(i+1))$, $t = 0, 1, 2, \ldots$.

Let $V_0 = [1, \tau_0]$ and then $V_j = [\tau_{j-1} + 1, \tau_j]$. Let $S = \max\{j \geq 0 : \tau_j \leq n\}$ (with $S = -1$ if $\tau_0 > n$). Let also $\tilde{V} = [\tau_S + 1, n]$ (with τ_{-1} taken to be 0). Note that $\tilde{V} \subseteq V_{S+1}$ (and $\tilde{V}$ may also be empty). In short, this divides $[1, n]$ into chunks, where each chunk ends with two consecutive indexes $i-1$ and i where $X_t(i-1) = X_t(i) = X_t^\epsilon(i-1) = X_t^\epsilon(i)$ for all t plus a chunk $\tilde{V}$ at the end, which is empty precisely if $\tau_S = n$. On the circle, i.e. where ω and ω^ϵ are defined, V_0 and $\tilde{V}$ are next to each other in a natural way.

Note that X_T and ω_T are equal on each V_j, $j = 1, \ldots, S$. They may differ on V_0 and $\tilde{V}$, but this difference will not need to be controlled in any other way than the simple observation that their cardinalities are bounded.

Stability will be proven by proving that, for sufficiently small $\epsilon > 0$, the number of bits where ω_T and $(\omega^\epsilon)_T$ differ is with high probability smaller than the difference between the number of 1's in ω_T and $n/2$.

For subsets I of $[n]$ and $y \in \{-1, 1\}^n$, let $y(I) = (y(i))_{i \in I}$. Let $s(y(I)) = \sum_{i \in I} y(i)$ and $s(y) = s(y([n]))$. Let $C_j = s((\omega^\epsilon)_T(V_j)) - s(\omega_T(V_j))$, $j = 0, 1, 2, \ldots, S$, $C_{\tilde{V}} = s((\omega^\epsilon)_T(V)) - s(\omega_T(\tilde{V}))$ so that $C_{\tilde{V}} + \sum_{j=0}^S C_j = s((\omega^\epsilon)_T) - s(\omega_T)$. Let also $C'_j = s((X^\epsilon)_T(V_j)) - s(X_T(V_j))$, $j = 1, 2, \ldots$ and note that the C'_j's are i.i.d. and that $C_j = C'_j$ for $j = 1, \ldots, S$, so that $s((\omega^\epsilon)_T(V_j)) - s(\omega_T(V_j))$ also equals $C_0 + C_{\tilde{V}} + \sum_{j=1}^S C'_j$.

Clearly all moments of $|V_j|$ are uniformly bounded. Let $\nu = \mathrm{E}[|V_1|]$ and let $m = \lfloor n/\nu \rfloor$. Fix arbitrarily small $\rho > 0$ and $\delta > 0$. By the Central Limit Theorem, there is a constant $K_1 < \infty$

independent of n such that with $K^- = m - K_1\sqrt{n}$ and $K^+ = m + K_1\sqrt{n}$, for large n

$$\mathrm{P}\left(\sum_{j=0}^{K^-} |V_j| \geq n\right) < \frac{1}{8}\delta,\ \mathrm{P}\left(\sum_{j=0}^{K^+} |V_j| \leq n\right) < \frac{1}{8}\delta$$

so that

$$\mathrm{P}(K^- \leq S \leq K^+) > 1 - \frac{1}{4}\delta. \tag{16}$$

The C_j''s by symmetry have mean 0. Take some $j \geq 1$ and fix it until (17). Let $F = F_j$ be the event that there exists an index $i \in V_j$ with $X(i) \neq X^\epsilon(i)$. On F_j^c, $X_T(i) = (X^\epsilon)_T(i)$ for all T and $i \in V_j$. Thus $C_j' = 0$ on F^c. Let

$$\chi_\epsilon = \chi_j^\epsilon = \min\{r > 0 : X(\tau_{j-1} + 2r) \neq X^\epsilon(\tau_{j-1} + 2r) \text{ or } X(\tau_{j-1} + 2r + 1) \neq X^\epsilon(\tau_{j-1} + 2r + 1)\}$$

$$\chi = \chi_j = \min\{r > 0 : X(\tau_{j-1}+2r) = X^\epsilon(\tau_{j-1}+2r) = X(\tau_{j-1}+2r+1) = X^\epsilon(\tau_{j-1}+2r+1) \neq X(\tau_{j-1})\}.$$

Then χ and χ_ϵ are geometric random variables that cannot take on the same value and $\chi \geq |V_j|/2$. Also, $F \subset G := \{\chi_\epsilon < \chi\}$ and it is standard that G is independent of $\min(\chi, \chi_\epsilon)$ and

$$\mathrm{P}(G) = \frac{2\epsilon - \epsilon^2}{2\epsilon - \epsilon^2 + \frac{1}{4}(1-\epsilon)^2} < 10\epsilon.$$

Hence

$$\begin{aligned}\mathrm{Var}(C_j') &\leq \mathrm{E}[{C_j'}^2] \leq 4E[|V_j|^2\mathbf{1}_F] \leq 4\mathrm{E}[\chi^2\mathbf{1}_G] \\ &= 4\mathrm{E}[\mathrm{E}[\chi^2\mathbf{1}_G|\min(\chi,\chi^\epsilon)]] \\ &= 4\mathrm{E}[\mathrm{P}(G|\min(\chi,\chi^\epsilon))\mathrm{E}[\chi|\min(\chi,\chi^\epsilon),G]] \\ &= 4\mathrm{P}(G)\mathrm{E}[(\min(\chi,\chi^\epsilon)+Y)^2] \\ &< 40\epsilon\mathrm{E}[(\min(\chi,\chi^\epsilon)+Y)^2],\end{aligned}$$

where Y is a copy of χ that is independent of (χ, χ_ϵ). This gives

$$\mathrm{Var}(C_j') < 160\epsilon\mathrm{E}[\chi^2] < 16000\epsilon. \tag{17}$$

It follows that

$$\mathrm{Var}\left[\sum_{j=1}^{K^-} C_j'\right] < 16000\epsilon m.$$

By the Central Limit Theorem, for sufficiently large constant M_5

$$\mathrm{P}\left(\left|\sum_{j=1}^{K^-} C_j'\right| > M_5\epsilon^{1/2}\sqrt{n}\right) < \frac{\delta}{4}.$$

Taking ϵ sufficiently small, we get

$$\mathrm{P}\left(\left|\sum_{j=1}^{K^-} C_j'\right| < \frac{1}{2}\rho\sqrt{n}\right) > 1 - \frac{1}{4}\delta. \tag{18}$$

Kolmogorov's inequality gives

$$\mathrm{P}\left(\max_K \Big|\sum_{j=K^-}^{K} C_j'\Big| > n^{1/3}\right) < \frac{1}{4}\rho \tag{19}$$

for n large. Adding that $\mathrm{P}(|C_0 + C_{\tilde{V}}| > a_n) \leq \mathrm{P}(|V_0| + |\tilde{V}| > a_n) \to 0$ for any $a_n \to \infty$ to (16), (18) and (19) and summarising and recalling $C_j = C_j'$ for $1 \leq j \leq S$, we get

$$\mathrm{P}\left(\left|C_0 + C_{\tilde{V}} + \sum_{j=1}^{S} C_j'\right| < \rho\sqrt{n}\right) = \mathrm{P}\left(|s(\omega^\epsilon)_T) - s(\omega_T)| < \rho\sqrt{n}\right) > 1 - \delta \tag{20}$$

for ϵ sufficiently small and n sufficiently large.

The second part is very similar, but slightly easier. Let $\xi_{-1}^+ = 0$. Define for $j = 0, 1, 2, \ldots$ recursively

$$\xi_j^- = \min\{i > \xi_{j-1}^+ + 1 : X_0(i-1) = X_0(i) = -1\},\ \xi_j^+ = \min\{i > \xi_j^- + 1 : X_0(i-1) = X_0(i) = 1\}.$$

Let $U_j = [\xi_{j-1}^+ + 1, \xi_j^+]$, $j = 0, 1, 2, \ldots$, and let $\mu = \mathrm{E}[|U_j|]$. Let $R = \max\{j : \xi_j^+ \leq n\}$. Let also $\tilde{U} = [\xi_R^+ + 1, n]$. Since $|U_j|$ is clearly bounded stochastically by two times a sum of two independent geometric(1/4) random variables, all moments of $|U_j|$ are finite. Let $m = \lfloor n/\mu \rfloor$ and fix an arbitrarily small $\delta > 0$. The Central Limit Theorem gives that for a sufficiently large constant K_2,

$$\mathrm{P}(L^- < R < L^+) > 1 - \frac{1}{4}\delta, \tag{21}$$

where $L^- = m - K_2\sqrt{n}$ and $L^+ = m + K_2\sqrt{n}$. Let $D_j = s(\omega_T(U_j)) = \sum_{i=\xi_{j-1}+1}^{\xi_j} \omega_T(i)$, $j = 0, 1, 2, \ldots, R$, $D_{\tilde{U}} = s(\omega_T(\tilde{U}))$ and $D_j' = s(X_T(U_j))$, $j = 1, 2, \ldots$. The D_j are i.i.d. with mean 0 (by symmetry) and, since $|D_j| \leq |U_j|$, finite and clearly nonzero variance. Since $X_T = \omega_T$ on $U_1, \ldots, U_R$, it also holds that $D_j' = D_j$ for $j = 1, \ldots, R$ and $s(\omega_T) = D_0 + D_{\tilde{U}} + \sum_{j=1}^{R} D_j'$. By the Central Limit Theorem and sufficiently small $\rho > 0$,

$$\mathrm{P}\left(\left|\sum_{j=1}^{L^-} D_j'\right| > 3\rho\sqrt{n}\right) > 1 - \frac{1}{4}\delta. \tag{22}$$

Also, for large n by Kolmogorov's inequality

$$\mathrm{P}\left(\max_L \Big| \sum_{j=L^-}^{L} D_j' \Big| < n^{1/3}\right) > 1 - \frac{1}{4}\delta. \tag{23}$$

Summing up (21), (22) and (23), we get for large n and sufficiently small $\rho > 0$,

$$\mathrm{P}\left(\Big|\sum_{j=1}^{R} D_j\Big| > 2\rho\sqrt{n}\right) > 1 - \frac{3}{4}\delta. \tag{24}$$

Also, for any $a_n \to \infty$, $P(|U_1| + |U_{R+1}| < a_n) \to 1$. This gives in conjunction with (24) for sufficiently large n,

$$\mathrm{P}\left(|s(\omega_T)| > 2\rho\sqrt{n}\right) > 1 - \delta.$$

Taking ρ also small enough to satisfy (20), we get for sufficiently large n

$$\mathrm{P}(\mathrm{maj}(\omega_T) \neq \mathrm{maj}((\omega^\epsilon)_T)) < 2\delta.$$

This proves noise stability.

□

Remark. With similar arguments as for the non-cycle case, it is easy to see that the outputs from each layer will soon freeze in a configuration that is easily determined by the following observation. Suppose $\omega_0(k-1) = \omega_0(k) = 1$ and let $l = \min\{j \geq k+2 \ : \ \omega_0(j-1) = \omega_0(j) = -1\}$. Let also $m = \max\{k+2 \leq j \leq l-2 \ : \ \omega_0(j-1) = \omega_0(j) = 1\}$. Then for $t \geq d$ (note that $l-m$ is even), we have $\omega_t(k+1) = \ldots = \omega_t((l+m-2)/2) = 1$, $\omega_t((l+m)/2) = \ldots = \omega_t(l) = -1$, where $d = (1/2)\max\{j-i : k+2 \leq i \leq j \leq \ell : \omega_0(i) = -1, \omega_0(i+1) = 1, \ldots, \omega_0(j-1) = -1, \omega_0(j) = 1\}$. Of course the analogous thing with signs reversed holds. Taking $\bar{d}$ as the maximum of all such d's over the whole input, $\omega_t = \omega_{\bar{d}}$ for all $t \geq \bar{d}$. It is easy to see that $\bar{d}$ is of order $\log n$.

As in Section 4.1, the result can be generalised to random θ_t. This is easily done by noticing that if θ_t represents a dictator function, each note at layer t only has one input from layer $t-1$ with no overlap between nodes. This means that we can effectively skip that layer and indeed the whole graph can be collapsed into a less deep graph with the same width where each layer expresses regular 3-majority. The result is summarised in the following corollary.

Corollary 4.2. *The iterated 3-majority with stride 1 and with random θ_t chosen according to any probability distribution on the n-cycle is quenched and annealed noise stable.*

4.3 Convolutional iterated 3-majority with stride 2

Here $G_n = G_n^{(2)}$, i.e. the graph in Figure 2c, and f_n induced by G_n are considered, but as we will see in the end, the results are easily extended to $G_n^{(2')}$. There are a few crucial observations to make about G_n. First that at each layer k, $|A_{j,k}| \leq 2$ for every input node $j \in L_0$. This follows, by the definition of G_n, from the easily checked fact that the union of the sets of parents of two neighbouring nodes are either two neighbouring nodes or a single node. Secondly, for two nodes $u, u' \in L_k$ $D_u \cap D_{u'} = \emptyset$ if u and u' are not next to each other since the set of children of u and u' are disjoint and no child of u is next to any child of u'. This means $f(u)$ and $f(u')$ are measurable with respect to two disjoint subsets of ω and are hence independent. We can now formulate the following lemma.

Lemma 4.2. Fix an arbitrary layer L_k and let $u_j = v_{k,j}$ be the nodes in L_k enumerated from left to right. Also, let $S \subseteq \{1, \ldots, |L_k|\}$. Then

$$\mathrm{P}(\forall j \in S : f(u_j) = 1) \geq \frac{1}{2^{|S|}}.$$

Proof. Since $f(u_j)$ is an increasing function of ω, this follows from Harris inequality.

□

Now, let $\mathcal{A}$ be some randomised algorithm to determine the value of $f_n(\omega)$ by querying necessary values from ω one by one. Let $J_{\mathcal{A}}$ be the set of ω_i that are queried to determine $f(\omega)$. As in ([**?**]) p 90, we make the following definition

Definition 4.1. The revealment of a randomised algorithm $\mathcal{A}$ for a Boolean function f, denoted δ_A, is defined by

$$\delta_{\mathcal{A}} = \max_{j \in 1, \ldots, 2^{n+1}-1} \mathrm{P}(j \in J_{\mathcal{A}})$$

and the revealment of a Boolean function f is defined as

$$\delta_f = \inf_{\mathcal{A}} \delta_{\mathcal{A}}.$$

The following crucial fact holds [**?**] p 93.

Theorem 4.3. *If the revealments satisfy*

$$\lim_{n \to \infty} \delta_{f_n} = 0$$

then $\{f_n\}$ is noise sensitive.

We can now state the following theorem.

Theorem 4.4. *The sequence of convolutional iterated* 3*-majority function on* $G_n^{(2)}$ *with stride* 2 *is noise sensitive. This also holds on* $G_n^{(2')}$*.*

Proof. Let n be fixed, and for now, a multiple of three. We recursively define an algorithm $\mathcal{A}(m, W)$ that for all $m = 0, \dots, n/3$ and all $W \subseteq L_{3m}$, finds $f(w)$, $w \in W$ in random order. The algorithm goes as follows.

Starting step: $\mathcal{A}(0, W)$: Query all bits in $W \subseteq L_0$ in a random uniform order.

Inductive step: $\mathcal{A}(m+1, W)$: Let o be a uniform random permutation of W. Now recursively find $f(o(\ell))$, $\ell = 1, \dots, |W|$ by querying nodes in $D_{o(\ell),3m}$ with $\mathcal{A}(m, D_{o(\ell),3m})$ with the modification that when and if, in the process of doing so, encountering a node $v \in D_{o(\ell),3m}$ such that $f(v)$ has no information to give on $f(o(\ell))$, then skip the query of v. When querying each $f(o(\ell))$, do not use any information gained when querying $f(o(1)), \dots, f(o(\ell-1))$.

In other words, by this definition a node $v \in D_{W,3m}$ that has two ancestors, a_1 and a_2, in W may have been queried to find $f(a_1)$ previously, but the knowledge of $f(v)$ is not used if one later needs to query $f(a_2)$ until the turn comes to v in $\mathcal{A}(m, D_{a_2,m})$ (if $f(v)$ can influence $f(a_2)$ at that point.) We refer to this as the algorithm is *forgetful*; when querying each $f(o(\ell))$, no input is known at the start of doing that. (Of course, when the turn comes to v, do not query $f(v)$ again, but simply do not use the knowledge of $f(v)$ until this precise moment.)

Note that by forgetfulness, the permutation o is independent of which input bits are queried in the end. Refer to this as the algorithm being *permutation independent*.

Let $R_{m,W,j}$ be the event that $\mathcal{A}(m, W)$ queries j and let $q_m = \max_W \max_j \mathrm{P}(R_{m,W,j})$.

Fix an arbitrary leaf j and assume that $W \subseteq L_{3(m+1)}$ and $W \cap A_{j,3(m+1)} \neq \emptyset$. We claim that $\mathrm{P}(R_{m+1,W,j}) \leq \mathrm{P}(R_{m,D_{W,3m},j})$. This follows directly from the forgetfulness of $\mathcal{A}(m+1, W)$ which implies that since the full algorithm $\mathcal{A}(m, D_{W,3m})$ queries every node in $D_{W,3m}$ whereas its modification when applied in $\mathcal{A}(m+1, W)$ generally does not and then the forgetfulness implies that deterministically, for every possible input ω, the modification only queries a subset of the input nodes of the ones queried by the full algorithm.

In order to ever query j by $\mathcal{A}(m+1, W)$ it must be the case that

(a) At least one $a \in A_{j,3m}$ is queried when querying w for some $w \in A_{a,3(m+1)}$ and

(b) for at least one such a, j is queried for finding $f(a)$.

For an ancestor $a \in A_{j,3(m+1)}$, let o'_a be the permutation of $D_{a,3m}$ that is included by recursion in Algorithm $\mathcal{A}(m, D_{a,3m})$. Let E_a be the event that $o'_a(u) < o'_a(h) < o'_v(v)$ for all $u \in D_{a,3m} \setminus A^+_{j,3m}$, $h \in A_{j,3m} \cap D_{a,3m}$ and $v \in A^+_{j,3m} \cap D_{a,3m} \setminus A_{j,3m}$. Let $E = \bigcap_{a \in A_{j,3(m+1)}} E_a$.

Let us lower bound $\mathrm{P}(E_a)$. Now $D_{a,3m}$ can contain either one or two ancestors of j and $A^+_{j,3m}$ could be included in $D_{a,m}$ or one node in $A^+_{j,3m} \setminus A_{j,3m}$ could be outside $D_{a,m}$. The "worst case" for the present purpose is when j has two ancestors, $h_1, h_2 \in A_{j,3m}$ and $A^+_{j,3m} \subset D_{a,3m}$. Since $D_{a,3m}$ contains 15 nodes we get in this case

$$\mathrm{P}(E_a) = \frac{1}{6\binom{15}{4}}.$$

It is easy to see that in the other cases, $\mathrm{P}(E)$ is larger than the right hand side. Since W may contain two ancestors of j, we get

$$\mathrm{P}(E) \geq \frac{1}{\left(6\binom{15}{4}\right)^2}.$$

Since the fact that probability of $\mathcal{A}(m, D_{W,3m})$ querying j is independent of of the order in which the nodes in $D_{W,3m}$ are queried (i.e. the permutation independence) and the above claim,

$$\mathrm{P}(R_{m+1,W,j} | E^c) \leq \mathrm{P}(R_{m,D_{W,3m},j}) \leq q_m.$$

Let F be the event that $f(u) = f(v)$ for all $u, v \in D_{A_{j,3(m+1)},3m} \setminus A^{+}_{j,3m}$. Then, according to Lemma 4.2, $\mathrm{P}(F) \geq 1/2^{18}$. On E, for j to be queried, $F^c \cap R_{m,D_{W,3m},j}$ must necessarily occur. Also, F and $R_{m,D_{W,3m},j}$ are independent. This holds since $R_{m,D_{W,3m},j}$ by forgetfulness depends only on the leafs in $D_{A_{j,3m}}$ and the randomness in the implicit random permutations when $\mathcal{A}(m,a)$, $a \in A_{j,3m}$ are carried out, whereas F depends on none of that. Clearly the two events are also independent of E. Hence we get

$$\mathrm{P}(R_{m+1,W,j}) \leq \mathrm{P}(E^c)q_m + \mathrm{P}(E)\mathrm{P}(F^c)\mathrm{P}(R_{m,D_{W,3m},j}|E) \leq (\mathrm{P}(E^c) + \mathrm{P}(E)\mathrm{P}(F^c))q_m.$$

Since this bound is independent of W and j,

$$q_{m+1} \leq (\mathrm{P}(E^c) + \mathrm{P}(E)\mathrm{P}(F^c))q_m \leq \left(1 - \frac{1}{36\binom{15}{4}^2 2^{18}}\right) q_m.$$

Hence

$$q_{n/3} \leq \left(1 - \frac{1}{36\binom{15}{4}^2 2^{18}}\right)^{n/3} \to 0$$

as desired. Cases when n is not a multiple of 3 can be completed by e.g. querying everything in layers $0, \ldots, 3(n/3 - \lfloor n/3 \rfloor)$ and then using the above algorithm from there. This gives

$$q_{n/3} \leq \left(1 - \frac{1}{36\binom{15}{4}^2 2^{18}}\right)^{\lfloor n/3 \rfloor} \to 0$$

which in combination with Theorem 4.3 proves the theorem.

□

The arguments in the proof are easily adapted to when θ_t are random and we get the following corollary.

Corollary 4.3. *The iterated 3-majority function with stride 2 is annealed and quenched noise sensitive if the distribution of θ_t is such that the probability of representing a majority is bounded away from zero.*

Proof. We show that with a very high probability, $\{f_n\}$ will be such that the revealment converges to zero as n grows. As previously discussed, θ_t represents either a majority or dictator function. Using the same algorithm as in Theorem 4.4, q_m is now random depending on the filter structure, and we get that for a fixed iteration m

$$q_{m+1} \leq \left(1 - \frac{1}{36\binom{15}{4}^2 2^{18}}\right) q_m$$

if the layers $3m$, $3m+1$ and $3m+2$ all represent majority layers. If this is not the case, i.e. if at least one of the layers represents a dictator, we instead just use that $q_{m+1} \leq q_m$. Letting

$$C_n = |\{m \in \{1, \ldots \lfloor n/3 \rfloor\} : \theta_{3m}, \theta_{3m+1} \textit{ and } \theta_{3m+2} \textit{ represents majority functions.}\}|$$

we get

$$q_{n/3} \leq \left(1 - \frac{1}{36\binom{15}{4}^2 2^{18}}\right)^{C_n}.$$

Since the probability of a given θ_t representing a majority is bounded away from zero, $C_n \to \infty$ in probability as n grows and hence the right hand side converges to 0. This proves quenched noise sensitivity. Finally, using Theorem 2.1 and the fact that $E_\omega[f_n(\omega)] = 0$ for all θ_t, annealed noise sensitivity follows.

□

4.4 Extensions to convolutional iterated $2k+1$ majority with overlap

Obviously the problems just treated for convolutional iterated 3-majority are equally interesting with 3 replaced with $2k+1$ for some integer $k \geq 2$. In this case the sensitivity is trivial if the stride $s > 2k+1$ since this would correspond to the regular $2k+1$ iterated majority with no charred nodes. For smaller s the corresponding graph $G_{n,k,s}$ is defined as $G_{n,k,s} = (V_{n,k,s}, E_{n,k,s})$, where

$$V_{n,k,s} = \{v_{n,0}, v_{n-1,-k}, v_{n-1,-(k-1)}, \dots, v_{n-1,k}, v_{n-2,-2k}, \dots, v_{n-2,2k}, \dots, v_{0,-kn}, \dots, v_{0,kn}\} \text{ and}$$
$$E_{n,k,s} = \{(v_{t,i}, v_{t-1,j}) : t = n, \dots, 1, |i-j| \leq k, v_{t,i} \text{ and } v_{t-1,j} \text{ exist}\}$$

if $s = 1$ and

$$V_{n,k,s} = \{v_{n,1}, v_{n-1,1}, v_{n-1,2}, \dots, v_{n-1,2k+1}, v_{n-2,1}, \dots, v_{n-2,\frac{2ks^2-2k+s-1}{(s-1)}}, \dots, v_{0,1}, \dots, v_{0,\frac{2ks^n-2k+s-1}{(s-1)}}\} \text{ and}$$
$$E_{n,k,s} = \{(v_{t,i}, v_{t-1,j}) : t = 1, \dots, n, j - s(i-1) \in [1, 2k+1], v_{t,i} \text{ and } v_{t-1,j} \text{ exist}\}$$

if $1 < s < 2k+1$.

When $s > 1$, the crucial condition for the algorithm in Theorem 4.4 for proving noise sensitivity is that each input node has at most two ancestors at any generation. This can in fact be generalised to the $2k+1$ iterated majority with stride $s \geq 2$. This is shown in Lemma 4.3.

Lemma 4.3. Let $G_{n,k,s}$ be the corresponding graph to the convolutional iterated $2k+1$-majority with stride $s \geq 2$. Then $|A_{t,S}| \leq 2k$, for every t and every S of $m \leq 2k$ nodes next to each other.

Proof. Notice that if $s > 2k$, the corresponding graph $G_{n,k,s}$ would be such that $|A_{t,j}| = 1$ for each t and j, making the result trivial. Therefore, assume $2 \leq s \leq 2k$. Now observe that given a node $v_{t,i}$ at layer $t > 0$ has $2k+1$ children which are $\{v_{t-1,s(i-1)+j} : j \in \{1, \dots 2k+1\}\}$. So, for a node $v_{t+1,i}$ to be a parent to $v_{t,j}$ it must be that $j = s(i-1) + i'$ for some $i' \in \{1, \dots 2k+1\}$. Consequently, the parents to $v_{t,j}$ is $v_{t+1,i}$ such that

$$\left\lceil \frac{j-2k}{s} \right\rceil + 1 \leq i \leq \left\lfloor \frac{j-1}{s} \right\rfloor + 1$$

such that $v_{t,i}$ exists. Let $S = \{v_{t,u}, v_{t,u+1}, \dots, v_{t,u+m}\}$ where $u > 0$ and $0 \leq m \leq 2k-1$ such that $v_{t,u}$ and $v_{t,u+m}$ exist. Then

$$|A_{t+1,S}| = \left\lfloor \frac{u+m-1}{s} \right\rfloor - \left\lceil \frac{u-2k-1}{s} \right\rceil + 1 \leq \left\lfloor \frac{u+2k-2}{2} \right\rfloor - \left\lceil \frac{u-2k-1}{2} \right\rceil + 1 = 2k$$

The lemma now follows by using this argument recursively over t.

□

From Lemma 4.3 it follows that $|A_t, j| \leq 2k$ for all t and j. With that condition satisfied, the proof of Theorem 4.4 easily generalises. Without supplying further details, we state the following theorem.

Theorem 4.5. *The convolutional iterated $2k+1$-majority function with stride s is noise sensitive if $s \geq 2$.*

The results for noise stability also go through very easily if the stride is 1 when θ_t are non-random. This by noticing that if there are $k+1$ sequentially equal input bits with the same sign, no information can be transferred from one side to the other. In the non cyclic case, input bits "outside" such a sequence has no influence on the final result. So by a high probability, the function is determined by the central bits, which with a high probability does not have a disagreement between ω and ω^ϵ.

A similar result can be stated for the cyclic case where the proof unfolds in a similar fashion but with a slightly different definition on τ_j. All in all, we can state the following theorem.

Theorem 4.6. *The convolutional iterated $2k+1$-majority function with stride 1 is noise stable both with and without cyclical convention.*

When considering random weights for $s \geq 2$, noise sensitivity still holds as long as the distribution of θ_t is such that with probability bounded away from 0, θ_t represents ordinary majority. The arguments are identical to that of $k = 1$. Thus, we can state the following corollary.

Corollary 4.4. *If θ_t is such that the probability of representing a majority is bounded away from zero, then the convolutional iterated $2k+1$-majority function with stride $s \geq 2$ and random weights θ_t is annealed and hence also quenched noise sensitive.*

Remark. For stride one, the arguments with random weights do not generalise well when $k > 1$.

5 Open problems and research directions

We have taken inspiration from observed non-robustness phenomena for DNN classifiers and have determined when a few common DNN architectures give rise to noise sensitive or noise stable classifiers. Going on focusing on the non-robustness phenomena, there are numerous further avenues to be explored. First of all, of course, there are many other DNN architectures that can be considered. Can noise sensitivity/stability results be achieved for (some of) them? If so, will that help us to design powerful DNN architectures that are robust to noise?

Besides that, here are a first few questions concerning the fully connected DNN models.

- In this paper, the activation function is the sign function for all layers. Will replacing them with activation functions used in practice, such as the arctan function, the sigmoid function, the ReLU function, etc, cause different properties with respect to noise sensitivity/stability?
- Above, all layers are of equal width. Does it make a difference if the layers are allowed to be of different widths? Does it then matter how unequal the widths are, e.g. of vastly different orders as functions of n? Here we believe that having varying widths, does not matter no matter how much they vary, and that would be a fairly easy task to prove.
- In Section 3.2 it is shown that under weak assumptions, if $1 - \rho_n$ shrinks as $\log n/n$ or faster, then the resulting Boolean function will almost certainly be noise stable. On the other hand, if $1 - \rho_n$ shrinks as $(\log n)^3/n$, then we typically get a noise sensitive sequence of functions. Is there a cutoff between the two cases and if so, where between $\log n/n$ and $(\log n)^3/n$ is it?
- What happens if one has access to data and train the network to fit with that? This could be interpreted in many different ways. For example the setting could be the following. Suppose that data are generated by a particular fully connected DNN of the kind in Section 3.1 and suppose that this particular DNN expresses a noise stable function. Assume further that we get data-points one by one, where each data-point is a uniform random input together with its output. According to Theorem 3.1, the network will almost certainly be noise sensitive at the start of training and in the end, after having seen all possible inputs, the DNN has learnt to express the true noise stable function behind the data. Where along this process does the DNN turn from sensitive to stable? Can it also flip back and forth between noise sensitive and noise stable along the way?

Concerning the convolutional models, there is also more to be done for $k \geq 2$, i.e. for filters of size at least 5. We have already observed that when the filter size is at least 5, a filter can express many things besides (anti)majority and (anti)dictator function. For example with five input bits, a filter could express “the first input bit unless all the other bits agree on the opposite”. For stride $s \geq 2$, we showed that if the filter with at least some probability expresses a majority, then the resulting network is noise sensitive. However if regular majority is never expressed, then what?

Also, for $k \geq 2$ and stride 1, it remains open if the resulting network is noise stable as we believe it is.

The noise sensitivity properties in Section 3.1 and 3.2 are very strong and one can even input two strings from any joint distribution over $\{-1,1\}^n \times \{-1,1\}^n$ such that with probability tending to 1 with n the two strings neither equal nor completely opposite. With probability 1/2, the function described by the resulting DNN will produce different results for the two different strings. However, this is in a nonadversarial setting, i.e. the joint distribution of the input strings is independent of the network weights. Suppose an adversary gets information about the network weights and one of the input strings. Can he produce the second string by flipping bits of his own choice of the first string, with no other restriction than the expected number of flips has to be very small, so that the output of the second string differs from that of the first string? None of the results in this paper are concerned with questions such as these, and it would certainly be very interesting to look at that.

Achnowledgement

The first and third authors were supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The second author acknowledges the support of the Swedish Research Council, grant no. 2020-03763.