The Probabilistic Fault Tolerance of Neural Networks in the Continuous Limit

Anonymous authors
Paper under double-blind review

Abstract

The loss of a few neurons in a brain rarely results in any visible loss of function. However, the insight into what “few” means in this context is unclear. How many random neuron failures will it take to lead to a visible loss of function? In this paper, we address the fundamental question of the impact of the crash of a random subset of neurons on the overall computation of a neural network and the error in the output it produces. We study fault tolerance of neural networks subject to small random neuron/weight crash failures in a probabilistic setting. We give provable guarantees on the robustness of the network to these crashes. Our main contribution is a bound on the error in the output of a network under small random Bernoulli crashes proved by using a Taylor expansion in the continuous limit, where close-by neurons at a layer are similar. The failure mode we adopt in our model is characteristic of neuromorphic hardware, a promising technology to speed up artificial neural networks, as well as of biological networks. We show that our theoretical bounds can be used to compare the fault tolerance of different architectures and to design a regularizer improving the fault tolerance of a given architecture. We design an algorithm achieving fault tolerance using a reasonable number of neurons. In addition to the theoretical proof, we also provide experimental validation of our results and suggest a connection to the generalization capacity problem.

1 Introduction

Understanding the inner working of artificial neural networks (NNs) is currently one of the most pressing questions in learning theory. As of now, neural networks are the backbone of the most successful machine learning solutions. They are deployed in safety-critical tasks in which there is little room for mistakes. Nevertheless, such issues are regularly reported since attention was brought to the NNs vulnerabilities over the past few years.

Fault tolerance as a part of theoretical NNs research. Understanding complex systems requires understanding how they can tolerate failures of their components. This has been a particularly fruitful method in systems biology, where the mapping of the full network of metabolite molecules is a computationally quixotic venture. Instead of fully mapping the network, biologists improved their understanding of biological networks by studying the effect of deleting some of their components, one or a few perturbations at a time. Biological systems in general are found to be fault tolerant, which is thus an important criterion for biological plausibility of mathematical models.

Neuromorphic hardware (NH). Current Machine Learning systems are bottlenecked by the underlying computational power. One significant improvement over the now prevailing CPU/GPUs is neuromorphic hardware. In this paradigm of computation, each neuron is a physical entity, and the forward pass is done (theoretically) at the speed of light. However, components of such hardware are small and unreliable, leading to small random perturbations of the weights of the model. Thus, robustness to weight faults is an overlooked concrete Artificial Intelligence (AI) safety problem. Since we ground the assumptions of our model in the properties of NH and of biological networks, our fundamental theoretical results can be directly applied in these computing paradigms.

Research on NN fault tolerance. In the 2000s, the fault tolerance of NNs was a major motivation for studying them. In the 1990s, the exploration of microscopic failures was fueled by the hopes of developing neuromorphic hardware (NH). Taylor expansion was one of the tools used for the study of fault tolerance. Another line of research proposes sufficient conditions for robustness. However, most of these studies are either empirical or are limited to simple architectures. In addition, those studies address the worst case, which is known to be
more severe than a random perturbation. Recently, fault tolerance was studied experimentally as well. DeepMind proposes to focus on neuron removal (25) to understand NNs. NVIDIA (21) studies error propagation caused by micro-failures in hardware (3). In addition, mathematically similar problems are raised in the study of generalization (29; 30) and robustness (42).

The quest for guarantees. Existing NN approaches do not guarantee fault tolerance: they only provide heuristics and evaluate them experimentally. Theoretical papers, in turn, focus on the worst case and not on errors in a probabilistic sense. It is known that there exists a set of small worst-case perturbations, adversarial examples (5), leading to pessimistic bounds not suitable for the average case of random failures, which is the most realistic case for hardware faults. Other branch of theoretical research studies robustness and arrives at error bounds which, unfortunately, scale exponentially with the depth of the network (29). We define the goal of this paper to guarantee that the probability of loss exceeding a threshold is lower than a pre-determined small value. This condition is sensible. For example, self-driving cars are deemed to be safe once their probability of a crash is several orders of magnitude less than that of human drivers (40; 15; 56). In addition, current fault tolerant architectures use mean as the aggregation of copies of networks to achieve redundancy. This is known to require exponentially more redundancy compared to the median approach and, thus, hardware cost. In order to apply this powerful technique and reduce costs, certain conditions need to be satisfied which we will evaluate for neural networks.

Contributions. Our main contribution is a theoretical bound on the error in the output of an NN in the case of random neuron crashes obtained in the continuous limit, where close-by neurons compute similar functions. We show that, while the general problem of fault tolerance is NP-hard, realistic assumptions with regard to neuromorphic hardware, and a probabilistic approach to the problem, allow us to apply a Taylor expansion for the vast majority of the cases, as the weight perturbation is smooth enough, introducing the continuous limit (39) to prove the properties of NNs: it requires neighboring neurons at each layer to be similar. This makes the moments of the error linear-time computable. To our knowledge, the tightness of the bounds we obtain is a novel result. In turn, the bound allows us to build an algorithm that enhances fault tolerance of neural networks. Our algorithm uses median aggregation which results in only a logarithmic extra cost – a drastic improvement on the initial NP-hardness of the problem. Finally, we show how to apply the bounds to specific architectures and evaluate them experimentally on real-world networks, notably the widely used VGG (38).

Outline. In Sections 2-4, we set the formalism, then state our bounds. In Section 5, we present applications of our bounds on characterizing the fault tolerance of different architectures. In Section 6 we present our algorithm for characterizing fault tolerance. In Section 7 we present our experimental evaluation. Finally, in Section 8 we discuss the consequences of our findings. Full proofs are available in the supplementary material. Code is provided at the anonymized repo github.com/iclr-2020-fault-tolerance/code. We abbreviate Assumption 1 → A1, Proposition 1 → P1, Theorem 1 → T1, Definition 1 → D1.

2 DEFINITIONS OF PROBABILISTIC FAULT TOLERANCE

In this section, we define a fully-connected network and fault tolerance formally.

Notations. For any two vectors $x, y \in \mathbb{R}^n$ we use the notation $(x, y) = \sum_{i=1}^{n} x_i y_i$ for the standard scalar product. Matrix $\gamma$-norm for $\gamma = (0, +\infty]$ is defined as $\|A\|_{\gamma} = \sup_{x \neq 0} \|Ax\|_{\gamma}/\|x\|_{\gamma}$. We use the infinity norm $\|x\|_{\infty} = \max |x_i|$ and the corresponding operator matrix norm. We call a vector $0 \neq x \in \mathbb{R}^n$ $q$-balanced if $\min |x_i| \geq q \max |x_i|$. We denote $[n] = \{1, 2, ..., n\}$. We define the Hessian $H_{ij} = \partial^2 y(x)/\partial x^i \partial x^j$ as a matrix of second derivatives. We write layer indices down and element indices up: $W_{ij}$. For the input, we write $x_i \equiv x^i$. If the layer is fixed, we omit its index. We use the element-wise Hadamard product $(x \odot y)_i = x_i y_i$.

Definition 1. (Neural network) A neural network with $L$ layers is a function $y_L: \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ defined by a tuple $(L, W, B, \varphi)$ with a tuple of weight matrices $W = (W_1, ..., W_L)$ (or their distributions) of size $W_i: n_i \times n_{i-1}$, biases $B = (b_1, ..., b_L)$ (or their distributions) of size $b_i \in \mathbb{R}^{n_i}$ by the expression $y_{l} = \varphi(z_{l})$ with pre-activations $z_{l} = W_{l} y_{l-1} + b_{l}$, $l \in [L]$, $y_{0} = x$ and $y_{L} = z_{L}$. Note that the last layer is linear. We additionally require $\varphi$ to be $L$-Lipschitz. We assume that the network was trained

\[1\text{-Lipschitz } \varphi \text{ s.t. } |\varphi(x) - \varphi(y)| \leq |x - y|. \] If $\varphi$ is $K$-Lipschitz, we rescale the weights to make $K = 1$: $W_{ij}^{1/K} \to W_{ij}^{1/1}$. This is the general case. Indeed, if we rescale $\varphi(x) \to K \varphi(x)$, then, $y_{l-1} \to K y_{l-1}$, and in the sum $z_{l} = \sum W_{ij}^{1/1}K \cdot K y_{l-1} \equiv z_{l}$
dependent on
We use this model because it mimics essential properties of NH (41). Components fail relatively
The recent review study (41) identifies three types of faults: permanent, transient, and intermittent.

Definition 4. (Output error for a weight distribution) The error in case of weight failure with
definition 2 thus covers all these cases.

Next, we use tail bounds to guarantee (ε, δ)-FT. This definition means that with high probability
1 – δ additional loss due to faults does not exceed ε. Expectation over the crashes U ~ D|x can be
interpreted in two ways. First, for a large number of neural networks, each having permanent crashes,
ΕΔ is the expectation over all instances of a network implemented in the hardware multiple times.
For a single network with intermittent crashes, EΔ is the output of this one network over repetitions.

Interpretation. To evaluate the fault tolerance of a network, we compute the first moments of ΔL+1.
Next, we use tail bounds to guarantee (ε, δ)-FT. This definition means that with high probability
1 – δ additional loss due to faults does not exceed ε. Expectation over the crashes U ~ D|x can be
interpreted in two ways. First, for a large number of neural networks, each having permanent crashes,
ΕΔ is the expectation over all instances of a network implemented in the hardware multiple times.
For a single network with intermittent crashes, EΔ is the output of this one network over repetitions.

Proposition 1. The task of evaluating EΔk for any k = 1, 2,... with constant additive or multiplicative
error for a neural network with ϕ ∈ C∞, Bernoulli neuron crashes and a constant number of
layers is NP-hard.

We provide a theoretical alternative for the practical case of neuromorphic hardware. We overcome
NP-hardness in Section 4 by providing an approximation dependent on the network, and not a constant
factor one: for weights W we give Δ and Δ dependent on W such that Δ(W) ≤ EΔ ≤ Δ(W).
In addition, we only consider some subclass of all networks.

3 THE HARDNESS OF FAULT TOLERANCE

In this section, we show why fault tolerance is a hard problem. Not only it is NP-hard in the most
general setting but, also, even for small perturbations, the error of the output of can be unacceptable.

3.1 NP-HARDNESS

A precise assessment of an NN’s fault tolerance should ideally diagnose a network by looking at the
outcome of every possible failure, i.e. at the Forward Propagated Error (23) resulting from removing
every possible subset of neurons. This would lead to an exact assessment, but would be impractical in
the face of an exponential explosion of possibilities as by Proposition 1 (proof in the supplementary
material).

Proposition 1. The task of evaluating EΔk for any k = 1, 2,... with constant additive or multiplicative
error for a neural network with ϕ ∈ C∞, Bernoulli neuron crashes and a constant number of
layers is NP-hard.

We provide a theoretical alternative for the practical case of neuromorphic hardware. We overcome
NP-hardness in Section 4 by providing an approximation dependent on the network, and not a constant
factor one: for weights W we give Δ and Δ dependent on W such that Δ(W) ≤ EΔ ≤ Δ(W).
In addition, we only consider some subclass of all networks.

3.2 PESSIMISTIC SPECTRAL BOUNDS

By Definition 4 the fault tolerance assessment requires to consider a weight perturbation W + U
given current weights W and the loss change yL+1(W + U) – yL+1(W) caused by it. Mathematically,

Empirical Risk Minimization – the standard task 1/k ∑k=1n ω(yL(xk), y∗k) → min
3The loss is bounded for the proof of Algorithm 1’s running time to work
Assumption 1. The probability of failure

Assumption 2. The number of neurons at each layer

Taylor expansion.

Table 1: Correspondence between discrete and continuous quantities. When an

regular (discrete) NN is a function mapping vectors to vectors, a continuous NN is an operator mapping functions to functions.

this means calculating a local Lipschitz coefficient $K$ connecting $\|y_{L+1}(W + U) - y_{L+1}(W)\| \leq K\|U\|$. In the literature, there are known spectral bounds on the Lipschitz coefficient for the case of input perturbations. These bounds use the spectral norm of the matrix $\| \cdot \|_2$ and give a global result, valid for any input. This estimate is loose due to its exponential growth in the number of layers, as $\|W\|_2$ is rarely $< 1$. See Proposition 2 for the statement:

**Proposition 2** ($K$ using spectral properties). $\|y_L(x_2) - y_L(x_1)\|_2 \leq \|x_2 - x_1\|_2 \cdot \prod_{l=1}^{L} \|W_l\|_2$

The proof can be found in (29) or in the supplementary material. It is also known that high perturbations under small input changes are attainable. Adversarial examples (5) are small changes to the input resulting in a high change in the output. This bound is equal to the one of (23), which is tight in case if the network has the fewest neurons. In contrast, in Section 4 we derive our bound in the limit $n \to \infty$.

We have now shown that even evaluating fault tolerance of a given network can be a hard problem. In order to make the analysis practical, we use additional assumptions based on the properties of neuromorphic hardware.

### 4 Realistic Simplifying Assumptions for Neuromorphic Hardware

In this section, we introduce realistic simplifying assumptions grounded in neuromorphic hardware characteristics. We first show that if faults are not too frequent, the weight perturbation would be small. Inspired by this, we then apply a Taylor expansion to the study of the most probable case.

**Assumption 1.** The probability of failure $p = \max\{|p_l| l \in \{L\}\}$ is small: $p \lesssim 10^{-4}, 10^{-3}$

This assumption is based on the properties of neuromorphic hardware (35). Next, we then use the internal structure of neural networks.

**Assumption 2.** The number of neurons at each layer $n_l$ is sufficiently big, $n_l \gtrsim 10^2$

This assumption comes from the properties of state-of-the-art networks (1).

The best and the worst fault tolerance. Consider a 1-layer NN with $n = n_0$ and $n_L = n_1 = 1$ at input $x_i = 1$: $y(x) = \sum x_i / n$. We must divide $1/n$ to preserve $y(x)$ as $n$ grows. This is the most robust network, as all neurons are interchangeable. Here $\mathbb{E}\Delta = -p$ and $\text{Var}\Delta = p/n$, variance decays with $n$. In contrast, the worst case $y(x) = x_1$ has all but one neuron unused. Therefore $\mathbb{E}\Delta = p$ and $\text{Var}\Delta = p$, variance does not decay with $n$.

The next proposition shows that under a mild additional regularity assumption on the network, Assumptions 1 and 2 are sufficient to show that the perturbation of the norm of the weights is small.

**Proposition 3.** Under $\mathcal{A}_L$ and if $\{W^i_l\}_{i=1}^{n_l}$ are $q$-balanced, for $\alpha > p$, the norm of the weight perturbation $U^i_l$ at layer $l$ is probabilistically bounded as: $\delta_0 = P(\|U_l^i\|_1 \geq \alpha\|W_l^i\|) \leq \exp(-n_l \cdot q \cdot d_{KL}(\alpha|p_l))$ with KL-divergence between numbers $a,b \in (0,1)$, $d_{KL}(a,b) = a \log a/b + (1-a) \log (1-b)/(1-a)$ and $W^i_l$ from $\mathcal{E}^{(i)}$.

Inspired by this result, next, we compute the error $\Delta$ given a small weight perturbation $U$ using a Taylor expansion.\footnote{The inspiration for splitting the loss calculation into favorable and unfavorable cases comes from (27).}

\footnote{In order to certify fault tolerance, we need a precise bounds on the remainder of the Taylor approximation. For example, for ReLU functions, Taylor approximation fails. The supplementary material contains another counter-example to the Taylor expansion of an NN. Instead, we give sufficient conditions for which the Taylor approximation indeed holds.}
As the width $n$ increases, networks $NN_n$ have a continuous limit \( NN_n \rightarrow NN_c \), where $NN_c$ is a continuous neural network \( \{1\} \), and $n = \min\{n_i\}$. That network $NN_c$ has globally bounded operator derivatives $D_k$ for orders $k = 1, 2$. We define $D_{12} = \max\{D_1, D_2\}$.

See Figure 1 for a visualization of $A3$ and Table 1 for the description of $A3$. The assumption means that with the increase of $n$, the network uses the same internal structure which just becomes more fine-grained. The continuous limit holds in the case of explicit duplication, convolutional networks and corresponding explicit regularization. The supplementary material contains a more complete explanation.

The derivative bound for order 2 is in contrast to the worst-case spectral bound which would be exponential in depth as in Proposition 2. This is consistent with experimental studies \( \cite{11} \) and can be connected to generalization properties via minima sharpness \( \cite{17} \).

Proposition 4. Under $A3$ derivatives are equal the operator derivatives of the continuous limit:

$$\frac{\partial^k y_L}{\partial y_{i1} \ldots \partial y_{il}^x} = \frac{1}{n_l^k} \frac{\delta^k y_L}{\delta y_{i1}(1) \ldots \delta y_{il}(1)} + o(1), \ n_l \rightarrow \infty$$

For example\( \cite{7} \) consider $y(x) = 1/n_1 \sum_{i=1}^{n_1} \varphi \left( \sum_{i=0}^{n_0} x_{i0} / n_0 \right)$ at $x_i \equiv 1$. Factors $1/n_0$ and $1/n_1$ appear because the network must represent the same $y^*$ as $n_0, n_1 \rightarrow \infty$. Then, $\partial y / \partial x_i = \varphi'(1) / n_1$ and $\partial^2 y / \partial x_i \partial x_j = \varphi''(1) / n_1^2$.

Theorem 1. For crashes at layer $l$ and output error $\Delta_L$ at layer $L$ under $A3$ with $q = 1/n_l$ and $r = p + q$, the mean and variance of the error can be approximated as

$$\mathbb{E} \Delta_L = p_l \sum_{i=1}^{n_l} \frac{\partial y_L}{\partial x_{i1}^L} \big|_{x_{i1}=0} + \Theta_\pm(1) D_2 r^2, \ \text{Var} \Delta_L = p_l \sum_{i=1}^{n_l} \left( \frac{\partial y_L}{\partial x_{i1}^L} \big|_{x_{i1}=0} \right)^2 + \Theta_\pm(1) D_3 r^3$$

By $\Theta_\pm(1)$ we denote any function taking values in $[-1, 1]$.

The full proof of the theorem is in the supplementary material. The remainder terms are small as both $p$ and $q = 1/n_l$ are small quantities under $A3$. In addition, $\mathbb{E} \Delta = O(1)$ remains constant, and $\text{Var} \Delta_L = O(1/n_l)$. This is the standard rate in case we estimate the mean of a random variable by averaging over $n_l$ independent samples, and our previous example in the beginning of the Section shows that it is the best possible rate. Our result shows sufficient conditions under which neural networks allow for such a simplification. In the next sections we use the obtained theoretical evaluation to develop a regularizer increasing fault tolerance, and say which architectures are more fault-tolerant.

5 Probabilistic Guarantees on Fault Tolerance Using Tail Bounds

In this section, we apply the results from the previous sections to obtain a probabilistic guarantee on fault tolerance. We identify which kinds of architectures are more fault-tolerant.

\( ^6 \)A necessary condition for $D_k$ to be bounded is to have a reasonable bound on the derivatives of the ground truth function $y^*(x)$. We assume that this function is sufficiently smooth.

\( ^7 \)The proposition is illustrated in proof-of-concept experiments with explicit regularization in the supplementary material. There are networks for which the conclusion of $A3$ would not hold, for example, a network with $w^{ij} = 1$. However, such a network does not approximate the same function as $n$ increases since $y(x) \rightarrow \infty$, violating $A3$.

\( ^8 \)The derivative $\partial y_L / \partial \xi^i(\xi) \equiv -\partial y_L(y_i + \xi \circ y_i) / \partial y_i \cdot y_i^x$ is interpreted as if $\xi^i$ was a real variable.

\( ^9 \)However, the dependency $\text{Var} \Delta \approx 1/n_l$ is only valid if $n < p^{-2} \sim 10^4$ to guarantee the first-order term to dominate, $p/n > r^3$. In case if this is not true, we can still render the network more robust by aggregating multiple copies with a mean, instead of adding more neurons. Our current guarantees thus work in case if $p^2 \leq n^{-1} \leq p$. In the supplementary material, we show that a more tight remainder, depending only on $p/n$, hence decreasing with $n$, is possible. However, it complicates the equation as it requires $D_5$. 

![Figure 1: Discrete (standard) neural network approximating a continuous network](image-url)
Under the assumptions of previous sections, the variance of the error decays as $\text{Var}\Delta \sim \sum C_i p_i / n_i$ as the error superposition is linear (see supplementary material for a proof), with $C_i$ not dependent on $n_i$. Given a fixed budget of neurons, the most fault-tolerant NN has its layers balanced: one layer with too few neurons becomes a single point of failure. Specifically, an optimal architecture with a fixed sum $N = \sum n_i$ has $n_i \sim \sqrt{p_i C_i}$.

Given the previous results, certifying $(\varepsilon, \delta)$-fault tolerance is trivial via a Chebyshev tail bound (proof in the supplementary material):

**Proposition 5.** A neural network under assumptions \[L, \lambda, \mu, \psi, \phi\] is $(\varepsilon, \delta)$-fault tolerant for $t = \varepsilon - \mathbb{E}\Delta_L > 0$ with $\delta = t^{-2}\text{Var}\Delta_L$ for $\mathbb{E}\Delta_L$ and $\text{Var}\Delta_L$ calculated by Theorem 7.\]

Evaluation of $\mathbb{E}\Delta$ or $\text{Var}\Delta$ using Theorem 1 would take the same amount of time as one forward pass. However, the exact assessment would need $O(2^n)$ forward passes by Proposition 1.

In order to make the networks more fault tolerant, we now want to solve the problem of loss minimization under fault tolerance rather than ERM (as previously formulated in (44)): \[
\inf_{(W, B) \in \mathcal{F}T} \mathcal{L}(w, B) \quad \text{where } \mathcal{F}T = \mathcal{F}T(L, \phi, p, \varepsilon, \delta) \text{ from Definition 5. Regularizing with Equation 1 can be seen as an approximate solution to the problem above. Indeed, } \text{Var}\Delta \approx p_i \sum_i \left( \frac{\partial L}{\partial y_i} \cdot y_i \right)^2 \text{ (from 11) is connected to the target probability (45). Moreover, the network is required to be continuous } \text{as it is a necessary condition for A3. See the supplementary material for complete details. Here } \hat{\mathcal{L}} \text{ is the regularized loss, } \mathcal{L} \text{ the original one, and } \lambda, \mu, \psi, \phi \text{ are the parameters:}
\]

\[
\hat{\mathcal{L}}(W) = \mathcal{L}(W) + \lambda \sum_{i=1}^{n_i} \left( \frac{\partial \mathcal{L}}{\partial y_i} \cdot y_i \right)^2 + \mu \left( \frac{\max_i W_l^i}{\min_i W_l^i} \right)^2 + \psi \cdot \text{smooth}(W_l) + \nu \| W \|_\infty
\]

We define the terms corresponding to $\lambda, \mu, \psi$ as $R_1 = \text{Var}\Delta / p_i$, $R_2 = q^2$, $R_3 = \text{smooth}(W_l)$. If we have achieved $\delta < 1/3$ by $R_3$, we can apply the well-known median trick technique (45), drastically increasing fault tolerance. We only use $R$ repetitions of the network with component-wise median aggregation to obtain $(\varepsilon, \delta \cdot \exp(-R))$-fault tolerance guarantee. See supplementary material for the calculations.

In addition, we show that after training, when $\mathbb{E}_x \nabla \mathcal{L} y_{\Delta+1}(x) = 0$, then $\mathbb{E}_x \mathbb{E}_y \Delta_{\Delta+1} = 0 + O(p^2)$ (proof in the supplementary material). This result sheds some light on why neural networks are inherently fault-tolerant in a sense that the mean $\Delta_{\Delta+1}$ is 0. Convolutional networks of architecture Conv-Activation-Pool can be seen as a sub-type of fully connected ones, as they just have locally-connected matrices $W_i$, and therefore our techniques still apply. Using large kernel sizes (see supplementary material for discussion), smooth pooling and activations lead to a better approximation.

We developed techniques to assess fault tolerance and to improve it. Now we combine all the results into a single algorithm to certify fault tolerance.

6 AN ALGORITHM FOR CERTIFYING FAULT TOLERANCE

We are now in the position to provide an algorithm (Algorithm 1) allowing to reach the desired $(\varepsilon, \delta)$-fault tolerance via training with our regularizer and then physically duplicating the network a logarithmic amount of times in hardware, assuming independent faults. We note that our algorithm works for a single input $x$ but is easily extensible if the expressions in Propositions are replaced with expectations over inputs (see supplementary material).

In order to estimate the required number of neurons, we use bounds from (41) which require $n \sim p_i / \varepsilon^2$. However, using the median approach allows for a fast exponential decrease in failure probability. Once the threshold of failing with probability $1/3$ is reached by $R_3$, it becomes easy to reach any required guarantee. The time complexity (compared to the one of training) of the algorithm is $O((D_2 + C_0) p_i / \varepsilon^2)$ and space complexity is equal to that of one training call. See supplementary material for the proofs of resource requirements and correctness.

\[\text{We note that the gradient of } \text{Var}\Delta \text{ is linear time-computable since it is a Hessian-vector product.}\]

\[\text{More neurons do not solve the problem, as } \mathbb{E}\Delta \text{ stays constant with the growth of } n \text{ by Theorem 1. Intuitively,}\]

\[\text{this is due to the fact that if a mean of a random variable is too high, more repetitions do not make the estimate lower.}\]
Data: Dataset $D$, input-output point $(x, y^*)$, failure probabilities $p_i$, depth $L$, activation function $\varphi \in C^\infty$, target $\varepsilon$ and $\delta'$, the error tolerance parameters from the Definition $5$ maximal complexity guess $C \approx \int |y'_l(t)| dt \approx R_{n}^{\max}$

Result: An architecture with $(\varepsilon, \delta')$-fault-tolerance on $x$

1. Select initial width $N = (n_1, ..., n_{L-1})$;
2. while true do
   3. Train a network to obtain $W, B$;
   4. Compute $q$ from Proposition $3$;
   5. If $q < 10^{-2}$, increase regularization parameter $\mu$ from Eq. $1$ continue; // go to line 3;
   6. Compute $\delta_0$ from Proposition $3$ using $q$;
   7. If $\delta_0 > 1/3$, increase $n$ by a constant amount, continue;
   8. Compute $R_3$ from Eq. $1$;
   9. If $R_3 > C$, increase regularization parameter $\psi$ from Eq. $1$ continue;
   10. Compute $E\Delta$ and $Var\Delta$ from Theorem $1$;
   11. If $E\Delta < \varepsilon$, output infeasible; // cannot do better than the mean $1$;
   12. Compute $\delta$ from Proposition $5$;
   13. If $\delta > 1/3$, increase $n$ by a constant amount and increase $\lambda$ in Eq. $1$ continue;
   14. Compute $R = O(\log \frac{1}{\delta})$;
   15. Output number of repetitions $R$, layer widths $N$, parameters $W, B$;
end

Algorithm 1: Achieving fault tolerance after training. The numbers $q_{\max} = 10^{-2}$ and $\delta_{\max} = 1/3$ are chosen for simplicity of proofs. The asymptotic behavior does not change with different numbers, as long as $\delta_{\max} < 1/2$ and the constraints on $q$ mentioned in the supplementary material are met.

Figure 2: The effect of the layer width $n_l$, $l = 1$, horizontal axis on the variance of the fault tolerance error $Var\Delta$, vertical axis

7 Experimental Evaluation

In this section, we test the theory developed in previous sections in proof-of-concept experiments. We first show that we can correctly estimate the first moments of the fault tolerance using $[1]$ for small (10-50 neurons) and larger networks (VGG). We test the predictions of our theory such as decay of $Var\Delta$, the effect of our regularizer and the guarantee from Algorithm $2$. See the supplementary material for the technical details where we validate the assumption of derivative decay ($\varphi'^2$) explicitly. Our code is provided at the anonymized repository [github.com/iclr-2020-fault-tolerance/code].

Increasing training dropout. We train sigmoid networks with $N \sim 100$ on MNIST (see ComparisonIncreasingDropoutMNIST.ipynb). We use probabilities of failure at inference and training stages $p_i = 0.05$ at the first layer and $10$ values of $p_i \in [0, 1.2p_i]$. The experiment is repeated $10$ times. When estimating the error experimentally, we choose $6$ repetitions of the training dataset to ensure that the variance of the estimate is low. The results are in the Table $2$. The experiments show that crashing MAE (Mean Absolute Error for the network with crashes at inference) is most dramatically affected by dropout. Specifically, training with $p_i \sim p_i$ makes network more robust at inference, which was well-established before. Moreover, the bound from $[1]$ can correctly order which network is trained with bigger dropout parameter with only $4\%$ rank loss, which is the fraction of incorrectly ordered pairs. All other quantities, including norms of the weights, are not able to order networks correctly. See supplementary material for a complete list of metrics in the experiment.
Quantity | Train rank loss | Test rank loss |
---|---|---|
Crashing, MAE | 5.6% | 5.6% |
Crashing, Accuracy | 19.8% | 17.7% |
Correct, MAE | 23.3% | 22.0% |
Correct, Accuracy | 31.7% | 39.9% |

(a) Experimental metrics

Quantity | Rank Loss |
---|---|
\(\text{Var}\Delta\) | 3.6% |
\(\text{Var}\Delta\) | 24.8% |
\(\text{Var}\Delta\) | 31.7% |
\(\text{Var}\Delta\) | 40.8% |

(b) Theoretical bounds

Table 2: Comparison of networks trained with increased dropout. Rank loss between \(p_{\text{train}}\) and various metrics. Experiment shows crashing and correct networks, MAE or accuracy and test/train datasets. Theoretical bounds include \(\text{I}^{\text{I}}\) and \(\text{I}^{\text{I}}\).

**Regularization for fault tolerance.** Previously, our bound is demonstrated to be able to correctly predict which network is more resilient. We therefore use it as a regularization technique suggested by Eq. [1], see Regularization.ipynb. We establish that the resilience of the network regularized with Dropout is similar to that of a network regularized with the bound

**Testing the bound on larger networks.** We test the bound on VGG16 and on a smaller convnet, see ConvNetTest-MNIST.ipynb and ConvNetTest-VGG16.ipynb and verify that they correctly predict the magnitude of the error

**Architecture and fault tolerance.** Comparing different architectures on a single image with \(p = 0.01\) (VGG16, VGG19, MobileNet) shows (and ConvNetTest-ft.ipynb) that the bigger the mean width of the layer (approximated by the number of parameters), the better is the fault tolerance, as predicted in Section 5. In addition, training networks on the MNIST dataset (see FaultTolerance-Continuity-FC-MNIST.ipynb) shows a decrease in variance with \(n_l\) as predicted by Theorem 1, see Figure 2: the variance decays as \(1/n_l\). We regularize with \(\psi = (10^{-4}, 10^{-2})\) for derivatives and smoothing respectively (see supplementary material for explanation of coefficients) and \(\lambda = 0.001\).

**Testing the algorithm.** We test the Algorithm 1 on the MNIST dataset for \(\varepsilon = 9 \cdot 10^{-3}, \delta = 10^{-5}\) and obtain \(R = 20, n_1 = 500, \lambda = 10^{-6}, \mu = 10^{-10}, \psi = (10^{-4}, 10^{-2})\). We evaluate the tail bound experimentally. Our experiment demonstrates the guarantee given by Proposition 5 and can be seen as an experimental confirmation of the algorithm’s correctness. See TheAlgorithm.ipynb.

We hence conclude that our proof-of-concept experiments show an overall validity of our assumptions and of our approach.

8 CONCLUSION

Fault tolerance is an important overlooked concrete AI safety issue [2]. This paper describes a probabilistic fault tolerance framework for NNs that allows to get around the NP-hardness of the problem. Since the crash probability in neuromorphic hardware is low, we can simplify the problem to allow for a polynomial computation time. We use the tail bounds to motivate the assumption that the weight perturbation is small. This allows us to use a Taylor expansion to compute the error. To bound the remainder, we require sufficient smoothness of the network, for which we use the continuous limit: nearby neurons compute similar things. After we transform the expansion into a tail bound to give a bound on the loss of the network. This gives a probabilistic guarantee of fault tolerance. Using the framework, we are able to guarantee sufficient fault tolerance of a neural network given parameters of the crash distribution. We then analyze the obtained expressions to compare fault tolerance between architectures and optimize for fault tolerance of one architecture. We test our findings experimentally on small networks (MNIST) as well as on larger ones (VGG-16, MobileNet). Using our framework, one is able to deploy safer networks into neuromorphic hardware.

Mathematically, the problem that we consider is connected to the problem of generalization [29, 27] since the latter also considers the expected loss change under a small random perturbation \(E_{W+U}[L(W+U) - L(W)]\), except that these papers consider Gaussian noise and we consider Bernoulli noise. Evidence [32], however, shows that sometimes networks that generalize well are not necessarily fault-tolerant. Since the tools we develop for the study of fault tolerance could as well be applied in the context of generalization, they could be used to clarify this matter.

12Variance for \(\text{I}^{\text{I}}\) is derived in the supplementary material
REFERENCES


Probabilistic Fault Tolerance of Neural Networks in the Continuous Limit
Supplementary material
Anonymous authors
September 2019

1 Introduction

First we prove all the propositions (labeled Proposition 1, 2, ... ) from the main paper. Their names and sections match the main paper. We also give additional results, they are labelled Additional Proposition 1, 2, etc. If they are formal statements of the results referred to in the main paper, they are in the same section as the reference.

We abbreviate Assumption 1 \( A1 \), Proposition 1 \( P1 \), Theorem 1 \( T1 \), Definition 1 \( D1 \).

Less precise statements with possible future research directions on fundamental questions required to make the guarantee even more strong are flushed right.

2 Definition of Probabilistic Fault Tolerance

Notations. For any two vectors \( x, y \in \mathbb{R}^n \) we use the notation \( (x, y) = \sum_{i=1}^{n} x_i y_i \) for the standard scalar product. Matrix \( \gamma \)-norm for \( \gamma = (0, +\infty) \) is defined as \( \| A \|_\gamma = \sup_{x \neq 0} \| A x \|_\gamma / \| x \|_\gamma \). We use the infinity norm \( \| x \|_\infty = \max |x_i| \) and the corresponding operator matrix norm. We call a vector \( 0 \neq x \in \mathbb{R}^n \) q-balanced if \( \min |x_i| \geq q \max |x_i| \). We denote \( |n| = \{1, 2, ..., n \} \). We define the Hessian \( H_{ij} = \partial^2 y(x)/\partial x^i \partial x^j \) as a matrix of second derivatives. We write layer indices down and element indices up: \( W_{ij} \). For the input, we write \( x_i \equiv x^i \). If the layer is fixed, we omit its index. We use the element-wise Hadamard product \( (x \odot y)_i = x_i y_i \).

Definition 1. (Neural network) A neural network with \( L \) layers is a function \( y_L : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} \) defined by a tuple \( (L, W, B, \varphi) \) with a tuple of weight matrices \( W = (W_1, ..., W_L) \) (or their distributions) of size \( W_l : n_l \times n_{l-1} \), biases \( B = (b_1, ..., b_L) \) (or their distributions) of size \( b_l \in \mathbb{R}^{n_l} \) by the expression \( y_l = \varphi(z_l) \) with pre-activations \( z_l = W_l y_{l-1} + b_l, l \in [L], y_0 = x \) and \( y_L = y_L \). Note that the last layer is linear. We additionally require \( \varphi \) to be 1-Lipschitz \( ^1 \). We assume that the network was trained using input-output pairs \( x, y^* \sim X \times Y \) using ERM \( ^2 \) for a loss \( \omega \). Loss layer for input \( x \) and the true label \( y^*(x) \) is defined as \( y_{L+1}(x) = \mathbb{E}_{y^* \sim Y|x} \omega(y_L(x), y^*) \) with \( \omega \in [-1, 1] \) \( ^3 \).

Definition 2. (Weight failure) Network \( (L, W, B, \varphi) \) with weight failures \( U \) of distribution \( U \sim D \) is the network \( (L, W + U, B, \varphi) \) for \( U \sim D \). We denote a (random) output of this network as \( y^W_U(x) = y_L(x) \) with activations \( \hat{y}_l \) and pre-activations \( \hat{z}_l \), as in \( D \).

\( ^1 \)1-Lipschitz \( \varphi \) s.t. \( |\varphi(x) − \varphi(y)| \leq |x − y| \). If \( \varphi \) is \( K \)-Lipschitz, we rescale the weights to make \( K = 1: W_l^{\prime j} \Rightarrow W_l^{\prime j}/K \). This is the general case. Indeed, if we rescale \( \varphi(x) \Rightarrow K\varphi(x) \), then \( y_{l-1} \Rightarrow K y_{l-1} \), and in the sum \( z_l' = \sum W_l^{\prime j}/K \cdot K y_{l-1} \equiv z_l \)

\( ^2 \)Empirical Risk Minimization – the standard task \( 1/k \sum y_{l-1} \omega(y_L(x_k), y^*_k) \to \min \)

\( ^3 \)The loss is bounded for the proof of Algorithm 1’s running time to work
Definition 3. (Bernoulli neuron failures) Bernoulli neuron crash distribution is the distribution with i.i.d. $\xi_i \sim \text{Be}(p_l)$, $U_{ij} = -\xi_i \cdot W_{ij}$. For each possible crashing neuron $i$ at layer $l$ we define $U_i = \sum_j |U_{ij}|$ and $W_i = \sum_j |W_{ij}|$, the crashed incoming weights and total incoming weights. We note that we see neuron failure as a sub-type of weight failure.

Definition 4. (Output error for a weight distribution) The error in case of weight failure with distribution $D((x,W)$ is $\Delta_l(x) = y_l^{W+U}(x) - y_l^W(x)$ for layers $l \in [L+1]

We extend the definition of $\varepsilon$-fault tolerance from [13] to the probabilistic case:

Definition 5. (Probabilistic fault tolerance) A network $(L,W,B,\varphi)$ is said to be $(\varepsilon,\delta)$-fault tolerant over an input distribution $(x,y^*) \sim X \times Y$ and a crash distribution $U \sim D((x,W)$ if $P((x,y^*) \sim X \times Y, U \sim D((x,W)) \{ \Delta_{L+1}(x) \geq \varepsilon \} \leq \delta$. For such network, we write $(W,B) \in FT(L,\varphi,p,\varepsilon,\delta)$.

3 The Hardness of The Fault Tolerance

3.1 NP-Hardness

Proposition 1. The task of evaluating $E\Delta^k$ for any $k = 1, 2, ...$ with constant additive or multiplicative error for a neural network with $\varphi \in C^\infty$, Bernoulli neuron crashes and a constant number of layers is NP-hard.

Proof. To prove that a problem is NP-hard, it suffices to take another NP-hard problem, and reduce any instance of that problem to our problem, meaning that solving our problem would solve the original one.

We take the NP-hard Subset Sum problem. It states: given a finite set of integers $x_i \in \mathbb{Z}$, $i \in [M]$, determine if there exist a non-empty subset $S \subseteq [M]$ such that $\sum_{i \in S} x_i = 0$.

We take a subset sum instance $x_i \in \mathbb{Z}$ and feed it as an input to a neural network with two first layer neurons with $\varphi$ being a piecewise-linear function from 0 to 1 at points $-\varepsilon$ and $\varepsilon$ for some fixed $\varepsilon \in (0,1)$. Note that in this proof we only compute $\varphi$ at integer points, and therefore it is possible re-define $\varphi \in C^\infty$ such that it has same values in natural points.

Note that inputs to it are integers, therefore the outputs are 1 if and only if (iff) the sum is greater than zero. First neuron has coefficients all 1s and second all $-1$s with no bias. The next neuron has coefficients 1 and 1 for both inputs, a bias $-1.5$ and threshold activation function, outputting 1 only if both inputs are 1. Again, since inputs to this neuron are integers, we can re-define $\varphi$ to be $C^\infty$. The final neuron is a (linear) identity function to satisfy Definition 4. It takes the previous neuron as its only input. Now, we see that $y(x) = 1$ if and only if the sum of inputs to a network is 0.

We now feed the entire set $S$ to the network as its input $x$. In case if $y(x) = 1$ (which is easy to check), we have arrived at a solution, as the whole set has zero sum. In the following we consider the case when $y(x) = 0$.

Suppose that there exist an algorithm calculating answer to the question $E\Delta^k > z$ for any finite-precision $z$.

Now, the expectation has terms inside $E\Delta^k = \sum_{s \in S} p^{|s|} (1 - p)^{|s| - |s'|} y_{s'}(x \circ s)$. Suppose that one of the terms is non-zero. Then $y(x \circ s) \neq 0$, which means that the threshold neuron outputs a value $> 0$ which means that sum of its inputs is greater than 1.5. This can only happen if they are both 1, which means that sum for a particular subset is both $\geq 0$ and $\leq 0$. By setting $z = 0$ we solve the subset sum problem using one call to an algorithm determining if $E\Delta^k > 0$. Indeed, in case if the original problem has no solution, the algorithm will output $E\Delta^k \leq 0$ as there are no non-zero terms inside the sum. In case if there is a solution, there will be a non-zero term in the sum and thus $E\Delta^k > 0$ (all terms are non-negative).

Now, suppose there exists an algorithm which always outputs an additive approximation to $E\Delta^k$ giving numbers $\mu$ and $\varepsilon$ such that $E\Delta^k \in (\mu - \varepsilon, \mu + \varepsilon)$ for some constant $\varepsilon$. Take a network from previous example with additional layer scaling output by $C = \frac{\mu}{p^2}$. Take a subset sum instance $x_i$. Output YES iff $0 \notin (\mu - \varepsilon, \mu + \varepsilon)$. This is correct because if $0 \notin (\mu - \varepsilon, \mu + \varepsilon)$, then it must be that $E\Delta^k > 0$ and vice versa. Multiplicative approximation has an analogous proof where we scale the outputs even more. 

\end{proof}
We note that computing the distribution of $\Delta$ for one neuron, binary input and weights and a threshold activation function is known as noise sensitivity in theoretical Computer Science. There exists an exact assessment for $W_k^{\dagger} \equiv w_k^\dagger$ however for the case $w_k^\dagger \neq 1$ the exact distribution is unknown [6].

### 3.2 Pessimistic Spectral Bounds

**Additional Proposition 1** (Norm bound or bound b1). For any norm $\| \cdot \|$ the error $\Delta_L$ at the last layer on input $x$ and failures in input can be upper-bounded as (for $\xi \odot x$ being the crashed input)

$$\|\Delta_L\| \leq \|W_L\| \cdot \ldots \cdot \|W_1\| \|\xi \odot x - x\|$$

**Proof.** We assume a faulty input (crashes at layer 0). By Definition 4 the error at the last layer is $\Delta_L = \hat{y}_L - y_L$. By definition 3 $y_L = W_L y_{L-1} + b_L$ and by Definition 3 $\hat{y}_L = W_L \hat{y}_{L-1} + b_L$. Thus, $\|\Delta_L\| = \|W_L \hat{y}_{L-1} + b_L - W_L y_{L-1} - b_L\| = \|W_L (\hat{y}_{L-1} - y_{L-1})\| \leq \|W_L\| \|\hat{y}_{L-1} - y_{L-1}\|$. Next, since $\hat{y}_{L-1} = \varphi(W_{L-1} \hat{y}_{L-2} + b_{L-1})$ and $y_{L-1} = \varphi(W_{L-1} y_{L-2} + b_{L-1})$, we use the 1-Lipschitzness of $\varphi$. In particular, we use that for two vectors $x$ and $y$ and $\varphi(x), \varphi(y)$ applied element-wise, we have $\|\varphi(x) - \varphi(y)\| \leq \|x - y\|$ because the absolute difference in each component is less on the left-hand side. We thus get $\|\hat{y}_{L-1} - y_{L-1}\| \leq \|W_{L-1} \hat{y}_{L-2} + b_{L-1} - W_{L-1} y_{L-2} - b_{L-1}\| = \|W_{L-1} (\hat{y}_{L-2} - y_{L-2})\| \leq \|W_{L-1}\| \|\hat{y}_{L-2} - y_{L-2}\|$. Plugging it in to the previous equation, we have $\|\Delta_L\| \leq \|W_L\| \cdot \ldots \cdot \|W_1\| \|\hat{x} - x\|$

Moreover we have failing input, thus $\hat{x} = \xi \odot x$ which completes the proof.

**Proposition 2** ($K$ using spectral properties [21]). $\|y_L(x_2) - y_L(x_1)\|_2 \leq \|x_2 - x_1\| \cdot \prod_{l=1}^L \|W_l\|_2$

**Proof.** Application of AF1 for $\| \cdot \|_2 = \| \cdot \|_2$ and $x_2 = \xi \odot x$

We see, there are bounds considering different norms of matrices $\|A\|_2$ (AF1) or using the triangle inequality and absolute values (AF2). They still lead to pessimistic estimates. All these bounds are known to be loose [19] due to $\|W\| = \sup_{x \neq 0} \|W x\|/\|x\|$ being larger than the input-specific $\|W x\|/\|x\|$. We will circumvent this issue by considering the average case instead of the worst case.

**Additional Corollary 1** (Infinity norm, connecting [9] to norm bounds). For an input $x$ with $C = \|x\|_{\infty}$ for failures at the input with $O(1)$ dependent only on layer size, but not on the weights,

$$\|\Delta_L\|_{\infty} \leq pC \|W_L\|_{\infty} \cdot \ldots \cdot \|W_1\|_{\infty} \cdot C(1) + O(p^2)$$

Here $\|x\|_{\infty} = \max |x_i|$

**Proof.** First we examine the expression (4) from the other paper [9] and show that it is equivalent to the result we are proving now:

$$\text{Erf} = \mathbb{E}\|\Delta\|_{\infty} \leq \sum_{i=1}^L C_i f_i K^{L-i} w_m^{(L)} \prod_{l=i+1}^L (N_l - f_l) w_m^{(l')}$$

here we have $C_i$ the maximal value at layer $l$, $K$ the Lipschitz constant, $w_m$ is the maximal over output neurons (rows of $W$) and mean absolute value over input neurons (columns of $W$) weight, $f_i$ is the number of crashed neurons.

Now we set $f_1 = pN_1$ and $f_i = 0$ for $i > 1$ and moreover we assume $K = 1$ as in the main paper.

---

4As we only have one neuron, the index is one-dimensional
Therefore the bound is rewritten as:
\[ E\| \Delta \|_\infty \leq C_1 N_1 p w_m^{(L)} \prod_{l=2}^{L} N_l w^{(l')} \]

Now we notice that the quantity \( N_l w^{(l')} = \| W^l \|_\infty \) and therefore
\[ E\| \Delta \|_\infty \leq p C_1 \frac{N_1}{N_L} \| W_2 \|_\infty \cdots \| W_L \|_\infty \]

Now we assume that the network has one more layer so that the bound from [9] works for a faulty input in the original network:
\[ E\| \Delta \|_\infty \leq p C_1 \frac{N_0}{N_L} \| W_1 \|_\infty \cdots \| W_L \|_\infty \]

Here \( C = \max \{|x_i|\} = \| x \|_\infty \). Next we prove that result independently using Additional Proposition 1 for \( \| \cdot \|_\infty \)
\[ \| \Delta_L \|_\infty \leq \| W_L \|_\infty \cdots \| W_1 \|_\infty \| \xi \odot x - x \|_\infty \]

Now we calculate \( E\| \xi \odot x - x \|_\infty \). We write the definition of the expectation with \( f(p, n, k) = p^k (1 - p)^{n-k} = p^k + O(p^{k+1}) \) being the probability that a binary string of length \( n \) has a particular configuration with \( k \) ones, if its entries are i.i.d. Bernoulli \( Be(p) \). Here \( S_l \) is the set of all possible network crash configurations at layer \( l \). Each configuration \( s_l \in S_l \) describes which neurons are crashed and which are working. We have \(|S_l| = 2^n\).
\[ E\| \xi \odot x - x \|_\infty = \sum_{s \in S} f(p, n, |s|) \max \{|s \odot x - x\} \]

Now since for \(|s| = 0\) the \( \max \{|s \odot x - x\} = \max \{|x - x\} = 0\), we consider cases \(|s| = 1\) and \(|s| > 1\). For \(|s| > 1\) the quantity \( f(p, n, |s|) = O(p^2) \) and therefore, in the first order:
\[ E\| \xi \odot x - x \|_\infty = p \sum |x_i| + O(p^2) \leq p N_0 \| x \|_\infty + O(p^2) \]

Next we plug that back into the expression for \( E\| \Delta \|_\infty \):
\[ E\| \Delta \|_\infty \leq \| W_L \|_\infty \cdots \| W_1 \|_\infty p N_0 \| x \|_\infty + O(p^2) \]

Now we note that this expression and the expression from [9] differ only in a numerical constant in front of the bound: \( \frac{N_0}{N_L} \) instead of \( N_0 \), but the bounds behave in the same way with respect to the weights.

\[ \square \]

**Additional Proposition 2** (Absolute value bound or bound b2). *The error on input \( x \) can be upper-bounded as:*
\[ E| \Delta_L | \leq p | W_L | \cdots | W_1 | | x | \]

*For \( | W | \) being the matrix of absolute values (\( | W |_{ij} = | W_{ij} | \)). \( | x | \) means component-wise absolute values of the vector.*

**Proof.** This expression involves absolute value of the matrices multiplied together as matrices and then multiplied by an absolute value of the column vector. The absolute value of a column vector is a vector of element-wise absolute values.

We assume a faulty input. By Definition 4, the error at the last layer is \( \Delta_L = \hat{y}_L - y_L \). By definition \( y_L = W_L y_{L-1} + b_L \) and by Definition 3 \( \hat{y}_L = W_L \hat{y}_{L-1} + b_L \). Thus for the \( i \)th component of the error,
\[ | \Delta_L | = | W_L \hat{y}_{L-1} + b_L - W_L^i y_{L-1} - b_L^i | = | W_L^i (\hat{y}_{L-1} - y_{L-1}) | = | \sum_j W_{ij}^i (\hat{y}_{L-1} - y_{L-1}) | \]
By the triangle inequality,

\[ |\Delta_L| \leq \sum_j |W_L^j| \cdot |\hat{y}_L^j - y_L^{j-1}| = (|W_L|\|\hat{y}_{L-1} - y_{L-1}|)^j \]

Next we go one level deeper according to Definition 1:

\[ |\hat{y}_L^j - y_L^{j-1}| = |\varphi(W_L^j\hat{y}_{L-2} + b_L^{j-1}) - \varphi(W_L^j y_{L-2} + b_L^{j-1})| \]

And then apply the 1-Lipschitzness property of \( \varphi \):

\[ |\hat{y}_L^j - y_L^{j-1}| \leq |W_L^j\hat{y}_{L-2} + b_L^{j-1} - W_L^j y_{L-2} - b_L^{j-1}| = |W_L^j (\hat{y}_{L-2} - y_{L-2})| \]

This brings us to the previous case and thus we analogously have

\[ |\Delta_L| \leq |W_L| \cdot |W_{L-1}| \cdot |\hat{y}_{L-2} - y_{L-2}| \]

Inductively repeating these steps, we obtain:

\[ |\Delta_L| \leq |W_L| \cdot \ldots \cdot |W_1| \cdot |\hat{x} - x| \]

Now we take the expectation and move it inside the matrix product by linearity of expectation:

\[ \mathbb{E}|\Delta_L| \leq \mathbb{E}|W_L| \cdot \ldots \cdot |W_1| |\hat{x} - x| = |W_L| \cdot \ldots \cdot |W_1| \mathbb{E}|\hat{x} - x| \]

The last expression involves \( \mathbb{E}|\hat{x} - x| \). This is component-wise expectation of a vector and we examine a single component. Since \( \xi^i \) is a Bernoulli random variable,

\[ \mathbb{E}|x_i \xi^i - x_i| = p \cdot |x_i| + (1 - p) \cdot 0 \]

Plugging it into the last expression for \( \mathbb{E}|\Delta_L| \) proves the proposition.

4 Realistic Simplifying Assumptions for Neuromorphic Hardware

4.1 When is The Weight Perturbation Small?

In this section we give sufficient conditions for which case the probability of large weight perturbation under a crash distribution is small. First, we define the properties of neuromorphic hardware.

Assumption 1. The probability of failure \( p = \max\{p_l|l \in [L]\} \) is small: \( p \lesssim 10^{-4} \ldots 10^{-3} \)

Assumption 2. The number of neurons at each layer \( n_l \) is sufficiently big, \( n_l \gtrsim 10^2 \)

Additional Assumption 1. Vectors \( W^j_l = \sum_j |W^{ij}_l| \) are \( q \)-balanced for \( q > 0 \) for each layer \( l \) \(^5\)

\(^5\)A similar assumption on an "even distribution" of weights was made in [20].
Toy examples. Naturally, we expect the error to decay with an increase in number of neurons $n$, because of redundancy. We show that this might not be always the case. First, consider a 1-layer NN with $n = n_0$ and $n_L = n_1 = 1$ at input $x_i = 1$: $y(x) = \sum x_i/n$. This is most robust network, as all neurons are interchangeable as they are computing the same function each. Essentially, this is an estimate of the mean of $\xi_i$ given $n$ samples. Here $\mathbb{E}\Delta = -\sum \mathbb{E}\xi_i x_i = -p$ and $\text{Var}\Delta = 1/n \sum \text{Var}\xi_i = p(1-p)/n \sim p/n$, variance decays with $n$. In contrast, the worst case $y(x) = x_1$ has all but one neuron unused. Therefore $\mathbb{E}\Delta = x_i \mathbb{E}\xi, p$ and $\text{Var}\Delta = x_i^2 \text{Var}\xi = p(1-p) \sim p$, variance does not decay with $n$.

This proposition gives sufficient conditions under which the weight perturbation is small and it is less and less as $n$ increases and $p$ decreases:

**Proposition 3.** Under $A_{\Xi}$ and $A_{\xi}$ for $\alpha > p$, the norm of the weight perturbation $U_l$ at layer $l$ is probabilistically bounded as: $\delta_0 = \mathbb{P}\{\|U_l\|_1 \geq \alpha\|W_l\|\} \leq \exp(-n_1 \cdot q \cdot d_{KL}(\alpha\|p\|))$ with $KL$-divergence between numbers $a, b \in (0,1)$, $d_{KL}(a, b) = a \log a/b + (1-a) \log (1-a)/(1-b)$

Proof. See [5] for the proof details as this is a standard technique based on Chernoff bounds for Binomial distribution with $p \to 0$. These are a quantified version of the Law of Large Numbers (which states that an average of identical and independent random variables tends to its mean). Specifically, Chernoff bounds show that the tail of a distribution is exponentially small: $\mathbb{P}\{X \geq \mathbb{E}X + \varepsilon \mathbb{E}X\} \leq \exp(-\varepsilon^2)$.

Specifically, if we consider a case $X = Bin(n, p)$ with $p \to 0$, for which $q = 1$, we have [5] for $\alpha = k/n > p$:

$$\mathbb{P}[X \geq k] \leq \exp(-nD_{KL}\left(\frac{k}{n}\|\alpha\|)\right)$$

In case if we rewrite $k = \alpha n$, this gives us the result.

Specifically, if we consider $\|U_l\|_1 = \sum \|U_i\|_1 \approx k\|W_l\|$. Therefore, this probability is $\mathbb{P}[X \geq k] = P[\|U_l\|_1 \geq \alpha\|W_l\|_1]$. This shows that the probability that a fraction of Bernoulli successes is more and more concentrated around its mean, $np$. Therefore, it is less and less probable that the this fraction is $\geq \alpha > p$.

Factor $q$ appears because in the analysis of the sum

$$\left(\sum_{i=1}^{n} \|W_i\|^2\right) \geq \frac{n^2 W_{\text{min}}^2}{nW_{\text{max}}^2} = nq$$

In order for the remainder to be small, we need additional assumptions. It is easy to come up with an example for which the first term $T_1$ is zero, but the error is still non-zero, illustrating that the remainder is important. Consider one neuron working and the rest having zero weights. Consider a summation of outputs of the neurons, each with a quadratic activation. Then $y = (x_i - \alpha)^2$, and $\nabla_x y = 0$ at $x_i = \alpha$. In addition, $\mathbb{E}\Delta = \Theta(p)$, however, the first term in Taylor expansion is $T_1 = 0$. The remainder here is $T_2 = O(p)$ and not $O(r^2)$, no matter how many neurons we have. The problem with this example is discontinuity: one neuron with non-zero weights is not at all like its neighbors. We thus show that discontinuity can lead to a lack of fault tolerance. Next, we generalize this sketch and show that some form of continuity is sufficient for the network to be fault-tolerant.

First, we reiterate on the toy motivating example from the previous section. Consider a 1-layer neural network

$$y(x) = \sum_{i=1}^{n} w_i x_i$$

4.2 Taylor Expansion for a Small Weight Perturbation

In this section, we develop a more precise expression for $\mathbb{E}\Delta$ and $\text{Var}\Delta$. Previously, we have seen that the perfect fault-tolerant network has $\text{Var}\Delta = \mathcal{O}(p/n)$. In this section, we give sufficient conditions when complex neural networks behave as the toy example from the previous section as well.

We would like to obtain a Taylor expansion of the expectation of random variable $\mathbb{E}\Delta = T_1 + T_2$ in terms of $p$ and $q = 1/n$ with $r = p + q$ where $T_1 = \mathcal{O}(p)$ and $T_2 = \mathcal{O}(r^2)$. For the variance, we want to have $\text{Var}\Delta = T_3 + T_4$ with $T_3 = \mathcal{O}(p/n)$ and $T_4 = \mathcal{O}(r^3)$. Our goal here is to make this expression decay as $n \to \infty$ as in the toy example.

We will show in Theorem 3 the first-order terms indeed behave as we expect. However, the expansion also contains a remainder (terms $T_2$ and $T_4$). In order for the remainder to be small, we need additional assumptions. It is easy to come up with an example for which the first term $T_1$ is zero, but the error is still non-zero, illustrating that the remainder is important. Consider one neuron working and the rest having zero weights. Consider a summation of outputs of the neurons, each with a quadratic activation. Then $y = (x_i - \alpha)^2$, and $\nabla_x y = 0$ at $x_i = \alpha$. In addition, $\mathbb{E}\Delta = \Theta(p)$, however, the first term in Taylor expansion is $T_1 = 0$. The remainder here is $T_2 = O(p)$ and not $O(r^2)$, no matter how many neurons we have. The problem with this example is discontinuity: one neuron with non-zero weights is not at all like its neighbors. We thus show that discontinuity can lead to a lack of fault tolerance. Next, we generalize this sketch and show that some form of continuity is sufficient for the network to be fault-tolerant.

First, we reiterate on the toy motivating example from the previous section. Consider a 1-layer neural network

$$y(x) = \sum_{i=1}^{n} w_i x_i$$
We assume that all neurons and inputs are used. Specifically, for the $q$-factor $q(x) = \max |x_i|/ \min |x_i|$, we have $q(x) \approx q(w) \sim 1$. We are interested in how the network behaves as $n \to \infty$ (the infinite width limit). For the input, we want the magnitude of individual entries to stay constant in this limit. Thus, $|x_i| = \mathcal{O}(1)$. Now we look at the function $y(x)$ that the network computes. Since the number of terms grows, each of them must decay as $1/n$: $w_i \sim 1/n$. The simplest example has $x = (1, \ldots, 1)$ and $w = (1/n, \ldots, 1/n)$, which results in $y(x) \equiv 1$ for all $n$. Now we consider fault tolerance of such a network. We take $\Delta = \sum_{i=1}^{n} w_i ((1 - \xi^i) - 1) x_i$ for $\xi^i \sim \text{Be}(p)$ being the indicator that the $i$th input neuron has failed. Therefore, $\mathbb{E} \Delta = -p/n \sum x_i w_i = -p$, and $\text{Var} \Delta = \sum x_i^2 w_i^2 p(1 - p) \sim p/n$. We see that the expectation does not change when $n$ grows, but variance decays as $1/n$. These are the values that we will try to obtain from real networks. Intuitively, we expect that the fault tolerance increases when width increases, because there are more neurons. This is the case for the simple example above. However, it is not the case if all but one neuron are unused. Then, the probability of failure is always $p$, no matter how many neurons there are. Thus, the variance does not decrease with $n$. We thus are interested in utilizing the neurons we have to their maximal capacity to increase the fault tolerance.

![Figure 1: Discrete (standard) neural network approximating a continuous network](image)

Functions, functionals and operators. We call maps from numbers to numbers as functions, maps from continuous limit the class of “good” networks for which fault tolerance is sufficient, via conditions for which the remainder terms are small. We use the first term explicitly when regularizing the network.

Additional Definition 2. (Continuous-discrete correspondence with $\mathcal{H}$ layers is an operator $y_L : \mathcal{H} \to \mathcal{H}$ defined by a tuple $(L, W, B, \varphi)$ with a tuple of weights functions $W = (W_1, \ldots, W_L)$ with $W_i \in \mathcal{H}(\mathbb{T}^2)$ (or their distributions), bias functions $B = (b_1, \ldots, b_L)$ (or their distributions) with $b_i \in \mathcal{F}$ by the expression $y_i(t) = \varphi(z_i(t))$, $z_i(t) = \int_{t' \in \mathbb{T}} W_i(t, t') y_{i-1}(t') dt' + b_i(t)$, $l \in \mathbb{N}, L$, $z_0 = y_0 = x$ and $y_L(t) = z_L(t)$. We note that a regular neural network has $T_l = [n_l]$. Continuous networks were re-introduced in $\mathcal{L}$. The classes of discrete and continuous networks can be generalized as Deep Function Machines $\mathcal{L}$, as a special case when $L_1$-norms of inputs and activations of discrete networks approximating continuous ones increase linearly with $n$. In $\mathcal{L}$, depth of a network is also considered to be continuous. Continuous networks are also called the integral representation $\mathcal{L}$. Previous work considers other limits rather than $L_1$-norm growing linearly, but here we only consider this case.

Next, we connect continuous and regular (discrete) networks. To do that, we define how each quantity transforms when $n$ grows.

Additional Definition 2. (Continuous-discrete correspondence with $\|x\|_1 = \mathcal{O}(n_0)$) Consider a continuous $\mathcal{NN}_c = (L, W_c, B_c, \varphi)$ and a discrete one $\mathcal{NN}_d = (L, W_d, B_d, \varphi)$. We consider a distribution of continuous inputs $\mathcal{F}$ $\exists x \sim P_c$ and a distribution of discrete inputs $\mathbb{R}^n \ni x \sim P_d$. These distributions are linked: each element from $P_c$ corresponds to exactly one from $P_d$. We define the approximation error $A$ as

$$A = \sup_{(x_c, x_d) \sim P_c \times P_d} \left[ \sup_{0 \leq t \leq L} \left[ \sup_{1 \leq i \leq n_l} \left| (y_i^d)^t - (y_i^c)^t \right| \left( \frac{i - 1}{n_l - 1} \right) \right] \right]$$
We define $n = \min\{n_0, \ldots, n_L\}$ the minimal width of the discrete network. If a series of discrete networks $NN_n$ has $A_n \to 0, n \to \infty$ for some $NN_c$, we say that $NN_n \to NN_c$.

The summary of correspondences between discrete and continuous quantities is shown in Table 1

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Discrete (AI[1])</th>
<th>Continuous (AI[1])</th>
<th>Relation (AI[2], AI[4])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>$x: [n_0] \to \mathbb{R}$</td>
<td>$x: [0, 1] \to \mathbb{R}$</td>
<td>$x_i \approx x \left( \frac{i-1}{n_0-1} \right)$</td>
</tr>
<tr>
<td>Norm</td>
<td>$|x_d|_1 = \sum_i</td>
<td>x_i</td>
<td>$</td>
</tr>
<tr>
<td>Bias</td>
<td>$b_i: [n_1] \to \mathbb{R}$</td>
<td>$b_i: [0, 1] \to \mathbb{R}$</td>
<td>$b_i \approx b_i \left( \frac{i-1}{n_1-1} \right)$</td>
</tr>
<tr>
<td>Weights</td>
<td>$W_i: [n_l] \times [n_{l-1}] \to \mathbb{R}$</td>
<td>$W_i: [0, 1]^2 \to \mathbb{R}$</td>
<td>$W_{i,j} \approx \frac{1}{n_{l-1}} \sum_{j=1}^{j-1} W_i\left( \frac{i-1}{n_l-1}, \frac{j-1}{n_{l-1}-1} \right)$</td>
</tr>
<tr>
<td>Pre-activations</td>
<td>$z_i = \sum_j W_{i,j} y_{j-1} + b_i$</td>
<td>$z_i(t) = \int^{t_1} W_i(t', t') y_{j-1}(t') dt' + b_i(t)$</td>
<td>$z_i \approx z_i \left( \frac{i-1}{n_l-1} \right)$</td>
</tr>
<tr>
<td>Number of changes</td>
<td>$C_d = \sum_{i,j}</td>
<td>W_{i+1,j} - W_{i,j}</td>
<td>$</td>
</tr>
</tbody>
</table>

Table 1: Correspondence between discrete and continuous quantities

**Additional Proposition 3.** Continuous networks with $\mathcal{H} = \mathcal{F}$ are universal approximators

**Proof.** Based on the proof of [15]. By the property of discrete networks, they are universal approximators [15].

Take a discrete network $y$ with a sufficiently low error, and define a continuous network $NN_c$ by using a piecewise-constant $W$ and $b$ from $y$. Then, $NN_c \equiv NN_d$. The weights and biases are bounded and piecewise-continuous with $n_l$ discontinuities at each layer. In addition, all functions are bounded since the weights are finite.

In the following we will always use $\mathcal{H} = \mathcal{F}$, as it is expressive enough (AI[3]), and it is useful for us. We give a sufficient condition for which $A_n \to 0$ as $n \to \infty$.

**Additional Proposition 4.** If a discrete network $NN_n$ is defined from a continuous network $NN_c$ in the following way, then $NN_n \to NN_c$ with error $A = \mathcal{O}(1/n)$. See (Figure 7)

$$x^i = x \left( \frac{i-1}{n_0-1} \right), \quad W_{i,j} = \frac{1}{n_{l-1}} W_l \left( \frac{i-1}{n_l-1}, \frac{j-1}{n_{l-1}-1} \right), \quad b_i = b_i \left( \frac{i-1}{n_l-1} \right)$$

In the following we write $i_l = \frac{i-1}{n_l-1}$ for an index $i = 1..n_l$, as the range for each of the indices is known.

**Proof.** For layer 0, the error is 0 by definition of $x^i$. Also, $x$ and its derivatives are bounded. Suppose we have shown that the error for layers $1..l-1$ is $\leq \varepsilon$, and that $y^{i_l}$ is bounded with globally bounded derivatives. Consider $|y^{i_l}(y) - y^{i_l}(y^{i+1}_{i+1})| = |b_i + b_{i+1} + \sum_j W_{i,j} y_{j-1} - W_l^j(y_i^j)(y_{i-1} + \varepsilon) - \int W_l^j(y_i^j)(y_{i-1} + \varepsilon) dt| \leq \varepsilon |W_l| |y_{i-1} + \varepsilon| + \|f(t)\| \leq \varepsilon |W_l| |y_{i-1} + \varepsilon| + \|f(t)\|$. The second term is a Riemann sum remainder and can be upper-bounded as $\frac{|f(t)|^2}{2r} \sup |f'(t)|$. We note that $||0, 1|| = 1$ and that $|f'(t)| = |W_l^j(y_i^j)(y_{i-1} + \varepsilon)| \leq \sup |W_l^j| \sup |y_i^j + \varepsilon| < \infty$ and does not depend on $n$ or the input, as the bound is global. Therefore, the error at layer $l$ is $\leq (\varepsilon + 1/n)C \leq D/n$ for the initial choice $\varepsilon = 1/n$. Thus, the full error is decaying as $1/n$: $A_n = \mathcal{O}(1/n)$.

Using that result, we conclude that for any sufficiently smooth operator $y$, there exist a sequence $NN_n$ approximating $y$. First, by [15], continuous networks are universal approximators. Taking one and creating discrete networks from it (as in AI[4]) gives the desired result as $\sup x \|y_c - y_d\|_{\infty} \leq A = \mathcal{O}(1/n) \leq \varepsilon$ and $\|y_c - y\| \leq \varepsilon$ by the approximation theorem for continuous nets.

**Assumption 3.** We assume that $NN_n \to NN_d$. We assume that there are global reasonable bounds on the derivatives of the function we want to approximate

$$\sup_{x \sim P} \left| \frac{\delta^k y_L(t_L)}{\delta y_i(t_i)} .. \frac{1}{\delta y_i(t_i)} \right| \leq D_k, \quad D_{a:b} = \max\{D_a, D_{a+1}, \ldots, D_b\}$$

Here for $i_s \in [n]$ and $q_j = |\{i_s = j\}|$ we define $(i_0, \ldots, i_k)! = \frac{n!}{q_1! \cdots q_{k+1}}$ a multinomial coefficient. In the paper, we only need $k \in \{1, 2\}$. 8
The limit means that close-by neurons inside a discrete network compute similar functions. This helps fault tolerance because neurons become redundant. There can be many of such sequences $N N_n$, but for the sake of clarity, one might consider that gradient descent always converges to the “same way” to represent a function, regardless of how many neurons the network has. This can be made more realistic if we instead consider the distribution of continuous networks to which the gradient descent converges to. In this limit, the distribution of activations at each layer (including inputs and outputs) stays the same as $n$ grows, and only the number of nodes changes. Note that this limit makes neurons ordered: their order is not random anymore. The limit works when $n_l$ are sufficiently large. After that threshold, the network stops learning new features and only refines the ones that it has created already.

The derivative bound part of $A_{3}$ can be enforced if there is a very good fit with the ground truth function $\tilde{y}(x)$. Indeed, if $\tilde{y}(x) \equiv y_L(x)$, then the derivative of the network depends on the function being approximated. For discrete-continuous quantities we write $X_\delta \approx X_c$ meaning that $|X_\delta - X_c| \leq \varepsilon$ for a sufficiently large $n$.

We note that our new assumptions extend the previous ones:

**Additional Proposition 5.** $A_{3}$ results in $A_{11}$ with

$$ q = \left( \int \int |W(t,t')| dt dt' \right)^2 + o(1), \ n_l \to \infty $$

Proof. A simple calculation:

$$ \sum |W^j| = n \cdot \frac{1}{n} \sum_i |W^{ij}| \to n \int |W(t,t')| dt dt', \ \sum |W^i|^2 \to n \int \left( \int |W(t,t')| dt \right)^2 dt $$

Dividing these two gives the result.

**Can we make the input norm bounded?** Sometimes, input data vectors are normalized: $\|x\| = 1$. In our analysis, it is better to do the opposite: $\|x\| = O(n)$. First, we want to preserve the magnitude of outputs to fit the same function as $n$ increases. In addition, the magnitude of pre-activations must stay in the same range to prevent vanishing or exploding gradients. We extend that condition to the input layer as well for convenience. Another approach would be to, for example, keep the norm $x$ constant as $n$ grows. In that case, to guarantee the same output magnitude, input weights must be larger. This is just a choice of scaling for the convenience of notation. We note that inputs can be still normalized component-wise.

**Less weight decay leads to no fault tolerance.** Consider the simplest case with Gaussian weights: $y = w^T x$, $w_i \sim N(0,\sigma^2)$. Then $y \sim N(0,\sigma^2 \|x\|_2^2)$. In order to avoid vanishing gradients, we set $\sigma^2 \|x\|_2^2 = \text{const}$. If $\|x\|_2 = O(n)$, then we must set $\sigma^2 \sim 1/n$ for the variance of the output to stay constant. This is a well-known initialization technique preserving the variance of pre-activations and thus eliminating vanishing or exploding gradients. In contrast to this, by $A_{3}$ we preserve the mean activation at each neuron instead, like for stochastic neural networks.

In the Gaussian case, the error $\Delta = -\sum \xi^i x_i w_i$. $E\Delta = 0$ and $\text{Var} \Delta = \sum x_i^2 E(\xi^i)^2 E w_i^2 = \sigma^2 p \|x\|_2^2 = p \cdot \text{const}$. This does not decay with $n$. A more formal statement can be found in $A_{11}$.

**The NTK limit.** In the NTK limit [14], there is a decay of the variance $\text{Var} \Delta$ since both $\|x\|_2 = 1$ and $\sigma^2 \sim 1/n$, but that happens because variance decreases in any case, not related to fault tolerance, since $\sigma^2 \|x\|_2^2 \sim 1/n$. The "lazy" regime [7] of the NTK limit implies that every hidden neuron is close to its initialization, which is random. Thus, there is no continuity in the network: close-by neurons do not compute similar functions. Thus, the NTK limit is incompatible with our continuous limit.

Below we formalize the claim that a network with constant pre-activations variance is not fault-tolerant. This means that an untrained network with standard initialization is not fault tolerant.

**Additional Proposition 6.** The fault tolerance error variance for a network $y = w^T x$ with $Ew_i = 0$ and $\text{Var}[y(x)] = \text{const}$ does not decay with $n$

Proof. Consider $y(x) = \sum w_i x_i$ and $\Delta = -\sum w_i x_i \xi^i$. Since $Ew_i = 0$, $Ey(x) = E \Delta = 0$. In addition, $\text{Var} y(x) = \sum \text{Var} w_i^2 x_i^2 E(\xi^i)^2 = p \text{Var} y(x) = p \cdot \text{const}$, not decaying with $n$.

Next, we want to harness the benefits of having a continuous limit by $A_{3}$. We want to bound the Taylor remainder, and for that, we bound higher-order derivatives. We bound them via the operator derivative of the continuous limit.
Operator derivatives. The derivatives for functionals and operators are more cumbersome to handle because of the many arguments that they have (the function the operator is taken at, the function the resulting derivative operator acts on, and the arguments for these functions), so we explicitly define all of them below. We consider operators of the following form, where \( x \in \mathcal{F} \) and \( Y[x] \in \mathcal{F} \) are functions:

\[
Y[x](t) = \sigma \left( \int W(t,t')x(t')dt' \right)
\]

This is one layer from AD\footnote{2}. We define the operator derivative of \( Y[x] \) point-wise with the RHS being a functional derivative:

\[
\frac{\delta Y[x]}{\delta x}(t) = \frac{\delta (Y[x](t))}{\delta x}
\]

For functionals \( F[x] \) we use the standard functional derivative:

\[
\frac{\delta F[x]}{\delta x}[\varphi] = \lim_{\varepsilon \to 0} \frac{F[x + \varepsilon \varphi] - F[x]}{\varepsilon}
\]

For our case \( F[x] = F_t[x] = Y[x](t) \), we have

\[
\frac{\delta F[x]}{\delta x}[\varphi] = \lim_{\varepsilon \to 0} \varepsilon^{-1} \left( \sigma \left( \int W(t,t')(x(t') + \varepsilon \varphi(t'))dt' \right) - \sigma \left( \int W(t,t')x(t')dt' \right) \right) = \sigma'(F[x]) \int W(t,t')\varphi(t')dt'
\]

Next, we consider the functional derivative at a point. This quantity is similar to just one component of the gradient of an ordinary function, with the complete gradient being similar to the functional derivative defined above. We define the derivative at a point \( \delta F[x]/\delta x(s) \) via the Euler-Lagrange equation, since we only consider functionals of the form \( F_t[x] = Y[x](t) \) for some fixed \( t \) which are then given by an integral expression

\[
F_t[x] = Y[x](t) = \sigma \left( \int \frac{W(t,t')x(t')}{L(t',x(t'),x'(t'))} dt' \right)
\]

We define the inner part \( G[x] = \int W(t,t')x(t')dt' \), thus, \( F[x] = \sigma(G[x]) \). In this case, since the integral only depends on the function \( x \) explicitly, but not on its derivatives, the functional derivative at point \( s \) is

\[
\frac{\delta F[x]}{\delta x}[s] = \sigma'(G[x]) \left( \frac{\partial L}{\partial x} - \frac{d}{ds} \frac{\partial L}{\partial x} \right) = \sigma'(G[x]) \frac{\partial L}{\partial x} = \sigma'(G[x])W(t,s)
\]

The definition of a functional derivative at a point \( \delta F[x]/\delta x(s) \) ("component of the gradient") can be reconciled with the definition of the functional derivative \( \delta F[x]/\delta x \) ("full gradient") if we consider the Dirac delta-function:

\[
\frac{\delta F[x]}{\delta x}[\varphi = \delta(t' - s)] = \sigma'(G[x]) \int W(t,t')\delta(t' - s) = \sigma'(G[x])W(t,s) \equiv \frac{\delta F[x]}{\delta x}(s)
\]

We define the operator derivative at a point in a point-wise manner via the functional derivative at a point:

\[
\left( \frac{\delta Y[x]}{\delta x(s)} \right)(t) = \frac{\delta Y[x](t)}{\delta x(s)}
\]

Having that definition, we compute for our case \( (\delta Y[x]/\delta x(s))(t) = \sigma'(G_t[x])W(t,s) \).

We now see that the rules for differentiating operators in our case are the same as the well-known rules for the derivatives of standard vector-functions. Indeed, if we consider \( y_i(x) = \sigma(\sum_j W_{ij}x_j) \) with the inner part \( g_i(x) = \sum_j W_{ij}x_j \) giving \( y_i(x) = \sigma(g_i(x)) \), then \( \partial y_i/\partial x_k = \sigma'(g(x))W_{ik} \). This looks exactly like the expression for \( \partial Y[x](t)/\partial x(s) \). By induction, this correspondence holds for higher derivatives as well. However, this does not imply that these quantities are equal. In fact, we will show that they differ by a factor of \( 1/\text{n}_t^k \) where \( k \) is the order of the derivative.

We characterize the derivatives of a discrete NN \( \partial^k y_L/\partial y_i^{t_l} \ldots \partial y_i^{t_1} \) in terms of operator derivatives of a continuous NN \( \delta^k y_L/\delta y_i^{(t_l)} \ldots \delta y_i^{(t_1)} \). We only assume the continuous limit:

**Proposition 4.** For a sequence \( NN_n \rightarrow NN_c \) with \( \varphi \in C^\infty \), the derivatives decay as:

\[
\frac{\partial^k y_L}{\partial y_i^{t_l} \ldots \partial y_i^{t_1}} = 1/\text{n}_0^k \cdot \frac{\delta^k y_L}{\delta y_i^{(t_l)} \ldots \delta y_i^{(t_1)}} + o(1), \text{n}_l \to \infty
\]
Intuitively, this means that the more neurons we have, the less is each of them important. First, consider a simple example $y = \sigma(\sum u_i x_i)$. Here the weight function is $u_i = 1/n$, $w(i) = 1$, $x_i = 1$, and $A = 0$. Then $\partial y/\partial x_i = \sigma'(\cdot)w_i \sim 1/n$ and $\partial^2 y/\partial x_i \partial x_j = \sigma''(\cdot)w_i w_j \sim 1/n^2$. We note that the expression inside the sigmoid has a limit and it’s close to the integral of continuity of $\sigma'$ and $\sigma''$.

**Proof.** Now we prove $\[1\]$. Consider the first and second derivatives. Note that the operator derivatives only depend on the number of layers and the dataset. It does not depend on any $n$ anymore.

$$ \frac{\partial y_L}{\partial x_{i_0}} = \sum_{i_{L-1}, \ldots, i_1} W_L^{i_L i_{L-1}} \ldots W_1^{i_1 i_0} \sigma'(z_L^{i_L}) \ldots \sigma'(z_1^{i_1}) $$

By definition of a neural net

$$ \approx \frac{1}{n_0} \int dt_1 W_1(t_1, t_0) \sigma'(z_1(t_1)) \ldots \int dt_{L-1} W_{L-1}(t_{L-1}, t_{L-2}) \cdot W_L(t_L, t_{L-1}) \cdot \sigma'(z_L(t_{L-1})) $$

continuous limit

$$ = \frac{1}{n_0} \frac{\partial y_L(i_L)}{\partial x(i_0)} $$

operator $\sim$ ordinary

Crucially, the factor $1/n_0$ appears because we **do not sum** over $i_0$, as it is fixed, but we have a weight vector $W_1 \sim 1/n_0$ nevertheless. For all other indices, we have a weight matrix $W_l \sim 1/n_{l-1}$ as well as a summation over $i_{l-1}$.

$$ \frac{\partial^2 y_L}{\partial x_{i_0} \partial x_{i_0'}} = \sum_{i_{L-1}, \ldots, i_1} W_L^{i_L i_{L-1}} \ldots W_1^{i_1 i_0} \sum_{s=1}^L \sigma''(z_s^{i_s}) \prod_{s' \neq s} \sigma'(z_s^{i_s'}) \sum_{i'_1, \ldots, i'_{l-1}} W_s^{i_s i'_s-1} \ldots W_1^{i_1 i'_0} $$

From previous

$$ \approx \frac{1}{n_0^2} \frac{\delta^2 y_L(i_L)}{\delta x(i_0) \delta x(i'_0)} $$

continuous+operator $\sim$ ordinary

Here, the factor $1/n_0^2$ appears because we never sum over $i_0$, but the weight matrix $W_1 \sim 1/n_0$ appears twice. $\Box$

**When does the limit hold?** Now, we have **assumed** that a network has a continuous limit in $A[3]$. However, this might not be the case: the NTK limit $[14]$ is an example of that, as weights there stay close to their random initialization $[7]$, thus, they are extremely dissimilar.

1. **Explicit duplication.** We copy each neuron multiple times and reduce the outgoing weights. If we set $W(t, t')$ to be piecewise-constant, then the approximation error is zero $A = 0$, and it does not depend on the degree of duplication. This is the obvious case where the network is becoming more fault-tolerant using duplication, and our framework confirms that. The problem with explicit duplication of non-regularized networks is that their fault-tolerance is suboptimal. Not all neurons are equally important to duplicate. Thus, it’s more efficient to utilize all the neurons in the best way for fault tolerance by duplicating only the important ones.

2. **Explicit regularization.** We make adjacent neurons compute similar functions, thus, allowing for redundancy. We first consider the local “number of changes” metric (Table 1). Specifically, for some function $z(t) \in [0, 1]$, $\int |z'(t)| dt$ represents how many times the function goes fully from 0 to 1 or vice versa. Our idea is to use that for the weights to quantify their discontinuity:

$$ C_1 = \sum_{ij} |W_t^{ij} - W_t^{i+1,j}| = \frac{1}{n_0} \sum_{ij} |W(i, j) - W(i + 1, j)| $$

$$ = \frac{1}{n_0} \sum_{ij} |W\left(\frac{i-1}{n_1-1}, \frac{j-1}{n_0-1}\right) - W\left(\frac{i}{n_1-1}, \frac{j}{n_0-1}\right)| $$

$$ \approx \frac{1}{n_0 n_1} \sum_{ij} |W_t^{ij}\left(\frac{i-1}{n_1-1}, \frac{j-1}{n_0-1}\right)| $$

$$ \approx \int |W_t^{ij}(t, t')| dt dt' $$
The term above, if small, guarantees that, for each input neuron, neighboring output neurons will use it in similar ways. The same is applied for $W^T$ as well: $C_2 = n_0/m_1 \sum_{ij} |W^{ij} - W^{ij+1}| \approx \int |W(t',t')| dt' dt''$ which guarantees that, for each output neuron, neighboring input neurons are used similarly by it.

In addition to making adjacent neurons computing similar functions, we add another term $C_3$ by Gaussian smoothing: a Gaussian kernel is convolved with the weights, and the weights are subtracted from the result. The difference shows how much the current value differs from an aggregated local "mean" value.

We explicitly enforce the continuous limit by adding a regularization term of $\text{smooth}(W) := C_1 + C_2 + C_3$. Here $\text{smooth}$ consists of three parts, see `FilterPlayground.ipynb`. The first part $C_1$, $C_2$ computes the numerical derivative with 2, 4, 10, 14 points in the array. The second part convolves the input with a Gaussian kernel and subtracts the original, resulting in a measure of discontinuity. All metrics are normalized to work with any size, so that scaling the network up does not change the magnitude of $C_i$ significantly. The implementation can be found in `continuity.py`.

We check the derivative decay prediction (Figure 1) which must follow from the continuity assumption $A_1$ experimentally. We run experiments for the MNIST dataset and architecture $(784, N, 100, 10) - 2$ hidden layers with sigmoid activations, unit Lipschitz coefficient, batch size of 1000, and $N$ from 50 to a few thousand. We measure $\|W\|_\infty = \max_i \sum_j |W^{ij}| \approx \max_i \int |W(t,t')| dt'$ which should stay constant (we measure the product of these over layers $l$, as only the total norm is expected to stay constant, whereas individual layers can change the magnitude of $W$). We also measure the components of the first derivative $\text{avg}((D^{ij}_l))$ and the components of the Hessian $\text{avg}((H^{ij}_l))$, $\text{avg}((H^{ij}_l))$ which we expect to decay with $n$ as by $\frac{1}{n}$. We show that without regularization, the weights $\|W\|_\infty$ increase, see `WeightDecay-FC-MNIST.ipynb`, and the derivatives and Hessians do decay but with a smaller rate, insufficient for $T_1$ to work, see `DerivativeDecay-FC-MNIST.ipynb`.

We repeat each experiment 10 times and report mean and standard deviation, see Figure 2. In contrast, our proposed regularization results in a smooth transition between neurons at each layer. They are grouped by their similarity. This can be visually seen in `WeightDecay-Continuity-FC-MNIST.ipynb` or in Figure 2. Contrary to networks without regularization, first layer weight profiles seem to have a meaningful image, and these images are similar for close-by neurons. There is also an increase in continuity of activations for a fixed input. For this regularization technique, we show that the product of the weights stabilizes, and the derivatives decay with proper slopes of $-1$ and $-2$. The accuracy, however, drops from 98% to 90%. We note that this does not demonstrate that our approach necessarily leads to a decrease in accuracy, as continuous networks are general by $A_3$.

We test the result on Fashion MNIST as well (Figure 3, notebook names are the same but with Fashion) and report similar behavior. On the Boston Housing dataset, non-regularized networks already have the desired characteristics (Figure 4).

Note: the condition above of $C_1 + C_2 + C_3$ being small is, strictly speaking, only a necessary condition for $A_1$ but not a sufficient one. Even if networks are smooth enough, they might not have a limit $NN_n \to NN_c$, as they could implement the function in drastically different ways: for example, the networks $NN_n$ can be all smooth, but approximate different continuous functions. However, this condition is sufficient for the derivatives $D_k$ to stay constant (second part of $A_3$), which is the only requirement for $T_1$ to work. Thus, our approach can give a formal guarantee of fault tolerance. An attempt to give a truly sufficient condition for $A_3$ would be to train the bigger network given duplicated weights of a smaller network, and penalizing a bigger network from having weights different from their initialization.

Different smoothing techniques. Currently, $C_2$ is unused in our implementation, as it was sufficient to use $C_1 + C_3$ to achieve the fault tolerance required in our experiments. Another method to make close-by neurons similar could be to use some clustering loss in the space of weights accounting for locality, like the Kohonen self-organizing map. One more idea is to regularize the weights with a likelihood of a Gaussian process, enforcing smoothness and a required range of dependencies. Another idea is to use the second derivative, which is connected to the curvature of a curve $(x, y(x))$ in the 2D space: $\kappa = y'' / (1 + (y')^2)^{3/2}$. The interpretation here is that $\kappa = 1/R$ for $R$ being the radius of curvature, a geometrical property of a curve showing how much it deviates from being a line.

3. Convolutional networks. For images, this limit naturally corresponds to scaling the images with intermediary pixels being just added between the original ones. Images are piecewise-continuous because...
adjacent pixels are likely to belong to the same object having the same color, and there are only finitely many objects on the image. Convolutional networks, due to their locality, fit our assumption on one condition. We need to have a large enough kernel size, as otherwise the non-smoothness is high. Specifically, for CNNs, \( C_1 \) is small, as neighboring neurons have similar receptive fields. In contrast, \( C_2 \) can be large in case if the kernel size is small: for example, the kernel \((-1, 1)\) in the 1D case will always result in a high discontinuity: the coefficients are vastly different and require more values in between them to allow for a continuous limit and redundancy.

The notebook ConvNetTest-VGG16-ManyImages.ipynb investigates into this. We note that we do not present this result in the main paper and make it a future research direction instead. In this paper, we give qualitative description of applicability of our techniques to convolutional networks (big kernel size for a small \( C_2 \) with respect to kernel size, smooth activation and pooling to allow for \( \Pi \) to work), as it would be out of the scope of a single paper to go into details.

4. Theoretical considerations for the general case.

How does the network behave when the number of neurons increases, and it is trained with gradient descent from scratch? First, there are permutations of neurons, which we ignore. Secondly, there could be many ways to represent the same function. One constraint is that the magnitude of outputs in the output layer is preserved. Intermediate layers need to have a non-vanishing pre-activation values to overcome vanishing/exploding gradients. In addition, input limit might be enforced such that \( x_i \approx x(\hat{i}) \). Now, gradient descent results in a discrete network \( NN \) which can be seen as a discretization of some continuous network \( \Pi \). Since NN’s derivatives are globally bounded, GD converges to a critical point. Each critical point determines the range of initializations which lead to it, partitioning the whole space into regions. Each fixed point with a sufficiently low loss thus corresponds to a set of continuous networks “passing through” a resulting discrete network. Each of the continuous networks can have different implementations in discrete networks of larger size. We choose a path in networks of different sizes \( n \) and denote the probability \( s_n \) over initializations to choose that particular continuous network. Therefore, on that path, derivatives decay as we want since \( NN_n \rightarrow NN_c \). The problem might arise if a particular continuous limit \( NN_c \) has an extremely small probability (over initializations) of gradient descent giving \( NN_n \): if \( s_n \rightarrow 0 \), this particular network \( NN_c \) is unlikely to appear. We leave the study of this as a future research direction.

Now we use the derivative decay from \( \Pi \) to show fault tolerance using a Taylor expansion. We write \( q = 1/n_d \) and \( r = p + q \). In the following we will use Assumption 3 only by its consequence – Proposition 4. We note that the conclusion of it can hold in other cases as well. We just give sufficient conditions for which it holds.

**Theorem 1.** For crashes at layer \( l \) and output of layer \( L \) under assumption 3 the mean and variance of the error can be approximated as

\[
E\Delta_L = p_L \sum_{i=1}^{n_1} \frac{\partial y_{\phantom{L}}}{} \left( \frac{\partial y_{\phantom{L}}}{} \right) + \Theta_\pm(1) D_2 r^2, \quad \text{Var}\Delta_L = p_L \sum_{i=1}^{n_1} \left( \frac{\partial y_{\phantom{L}}}{} \right)^2 + \Theta_\pm(1) D_2^2 r^3
\]

By \( \Theta_\pm(1) \) we denote any function taking values in \([-1, 1]\). The derivative \( \frac{\partial y_{\phantom{L}}}{\partial \xi_i^L}(\xi) \equiv -\frac{\partial y_{\phantom{L}}}{\partial \xi_i^L}(y_i - \xi \odot y_i)/\partial y_i \cdot y_i \) is interpreted as if \( \xi_i^L \) was a real variable.

**Proof.** We consider crashes at layer \( l \) as crashes in the input \( x \) to the rest of the layers of the network. Thus, without loss of generality, we set \( l = 0 \).

Consider \( \Delta(\xi) = y(1 - \xi \odot x) - y(x) \). Then we explicitly compute

\[
\frac{\partial \Delta(\xi)}{} = -\frac{\partial y(\hat{x}(\xi))}{\partial x_i} x_i, \quad \frac{\partial^2 \Delta(\xi)}{} = \frac{\partial^2 y(\hat{x}(\xi))}{\partial x_i \partial x_j} x_i x_j
\]

Now, consider \( \Delta(\xi) = \Delta(0) + (\Delta'(0), \xi) + \frac{1}{2}(\xi^T \Delta''(t(\xi)))\xi \) by the Taylor theorem with a Lagrange remainder.
We assume $\|x\| \leq 1$ (otherwise we rescale $W_1$). We group the terms into distinct cases $i = j$ and $i \neq j$:

$$E\Delta = -pD'(0) + \sum_{i=j} E(\xi_i^2) D''(t(\xi_i)) + \sum_{i \neq j} E(\xi_i^2) D''(t(\xi_i)) \xi_j^2$$

The second term is $|E_1| \leq p \cdot 2D_2/n_0^2 \cdot n_0 = O(pD_2/n_0) = (pqD_2)$

The third term is $|E_3| \leq n_0^2D_2p^2/n_0^3 = O(p^2D_2)$.

Therefore, we have an expansion

$$E\Delta = -\nabla_x \Delta(0) \cdot x + (q + p) \cdot pD_2 = -p \sum_i x_i \frac{\partial y(x)}{\partial x_i} + O(D_1p + D_2r^2)$$

The expectation just decays with $p$, but not with $n_0$. Now, consider the variance $\text{Var}\Delta = \text{Var}(\Delta(0)x) + 0.5 \xi_i^2 \Delta''(t(\xi_i)) = E\Delta^2 - (E\Delta)^2 = (E(V_1 + V_2)^2 - (E(V_1 + V_2))^2$.

$$\text{Var}V_1 + \text{Var}V_2 + 2Cov(V_1, V_2)$$

Consider $\text{Var}V_1 = p \sum (\frac{\partial y(x)}{\partial x_i})^2 x_i^2 = O(pqD_1)$. This is the leading term, the rest are smaller.

And the second term $\text{Var}V_2 \leq EV_2^2 = \sum_{ijkl} D_{ijkl}^2/n_0^4 E\xi_i \xi_j \xi_k \xi_l$. Here we consider various cases for indices $i, j, k, l$, based on the partition of $4 = 0 + 4 = 2 + 2 = 1 + 3$. If all indices are different, we get $O(p^4)$ from $E\xi$. If all are the same, we get $O(p/n^3)$. If we have 2 groups of 2, we get $O(p^3/n^4)$, if 2+1 we get $O(p^2/n^5)$. Thus, $\text{Var}V_2 = O(D_2^4r^4)$

Consider the final term $Cov(V_1, V_2) \leq EV_1V_2 = |EV_1| |EV_2|$. $|EV_1| \leq O(pD_2)$, $|EV_2| \leq D_2p/n$. $EV_1V_2 = \sum_{ijkl} D_{ijkl}^2/n_0^4 E\xi_i \xi_j \xi_k \xi_l$. All different indices give $p^4$, all same indices give $p/n_0^3$, and 2+1 give $p^2/n$. Thus, $\text{Cov}(V_1, V_2) = O(D_1^2D_2^2)$

Finally, $\text{Var}\Delta = p\nabla^2 \Delta(0)x + O(r^3D_1D_2^2) + O(r^4D_2^2) = \sum (\frac{\partial y(x)}{\partial x_i})^2 x_i^2 + O(r^3D_2^2) = O(D_1^2p/n + D_2^2r^3)$.

A better remainder. It is possible to obtain a remainder of $O(p^2/n) + O(p/n^2)$ for the variance instead of a generic $O(r^3)$. This makes the remainder decay with $n$ as well. A more fine-grained analysis of the remainder could
consider the expression for the variance and explicitly compute a correlation between $\xi(t(\xi))$ and $\xi(t(\xi))$. However, we were unsuccessful in doing so. Another approach is to take one more term in the expansion: $O(n^2)$ – it will make the previous term with $p^3$ go away, leaving only terms $p^2/n$ and $p/n^2$, as the difference expressions in the variance will cancel out. Another idea is to use a Taylor series $\text{Var}\Delta = \sum_{k=0}^{\infty} \frac{f_k}{2k!}x_k$ (existing in principle). Next, we explicitly bound the remainder since we have a global bound on $D_k \leq D$. Finally, if we expand around the mean $\mu = \mathbb{E}x$: $\mathbb{E}f(x) = f(\mu) + \frac{1}{2}\theta(x)(f''_{\mu}(\mathbb{E}x))$, we get a remainder $p/n$ right away, but need to compute $f(\mu)$ which is the network at a modified input $(x - px)$.

5  Probabilistic Guarantees on The Fault Tolerance Using Tail Bounds

Under our assumptions, $\text{Var}\Delta \sim \sum_{i} \frac{C_{i}p_{i}}{n_{i}}$. The constant $C_i$ comes from Theorem 1 (or its more computationally tractable form in Add. Coroll. 2). Since we know that the error superposition is linear (Additional Proposition 7), we sum the individual layer terms, hence $\sum_{i}$. Proposition 4 motivates the inverse dependency on $n_i$.

Median trick. Suppose that we have a random variable $X$ satisfying $\mathbb{P}\{X \geq \epsilon\} < 1/3$. Then we create $R$ independent copies of $X$ and calculate $X = \text{median}\{X_i\}_{i=1}^{R}$. Then $\mathbb{P}\{X \geq \epsilon\} < (1/3)^{R/2}$ because in order for the median to be larger than value $\epsilon$, at least half of its arguments must be larger than $\epsilon$, and all $X_i$ are independent. Thus $R = O(\log 1/\delta)$ in order to guarantee $(1/3)^{R/2} < \delta$. This is a standard technique.

Proposition 5. A neural network under assumptions 1, 3 is $(\epsilon, \delta)$-fault tolerant for $t = \epsilon - \mathbb{E}\Delta_L > 0$ with $\delta = \delta_0 + t^{-2}\text{Var}\Delta_L$ for $\mathbb{E}\Delta$ and $\text{Var}\Delta$ calculated by Theorem 1 and $\delta_0$ from Proposition 3.
Proof. We apply the Taylor expansion from Theorem 1. We directly apply the Chebyshev’s inequality for \( X = \Delta \):
\[
\delta = \mathbb{P}\{X \geq t + \mathbb{E}X\} \leq \frac{\text{Var}X}{t^2}
\]

\( \Box \)

5.1 Properties of The Fault Tolerance

**Proposition 6.** Suppose that a \( C^2 \) network is at a stationary point after training: \( \mathbb{E}_x \nabla_W L(x) = 0 \). Then in the first order on \( p, q, \mathbb{E}_x \mathbb{E}_q \mathbb{E}_L \mathbb{E}_{L+1} = 0 \)

**Proof.** Consider the quantity in question, use Additional Proposition 5 for it and apply \( \mathbb{E}_x \) on both sides:
\[
\mathbb{E}_\xi \mathbb{E}_{L+1} = \mathbb{E}_x \left( \frac{\partial L}{\partial y} \mathbb{E}_\xi \Delta \right) = -p \left( \mathbb{E}_x \frac{\partial L}{\partial W}, W \right)
\]

Now since we know that \( \mathbb{E}_x \frac{\partial L}{\partial W} = 0 \), the linear term is 0

**Additional Proposition 7.** (Linearity of error superposition for small \( p \) limit) Consider a network \((L, W, B, \varphi)\) with crashes at each layer with probability \( p_l, l \in \stackrel{n}{\overline{0}}, \mathcal{L} \). Then in the first order on \( p \), the total mean or variance of the error at the last layer is equal to a vector sum of errors in case crashes were at a single layer
\[
\mathbb{E}_\Delta^{p_0 \ldots p_L} = \mathbb{E}_\Delta^{p_0} + \ldots + \mathbb{E}_\Delta^{p_L}, \text{Var}\Delta^{p_0 \ldots p_L} = \text{Var}\Delta^{p_0} + \ldots + \text{Var}\Delta^{p_L}
\]

**Proof.** Consider each layer \( i \) having a probability of failure \( p_i \in [0, 1] \) for \( i \in \mathcal{L} \).

In this proof we utilize Assumption 1. We write the definition of the expectation with \( f(p,n,k) = p^k(1-p)^{n-k} \) being the probability that a binary string of length \( n \) has a particular configuration with \( k \) ones, if its entries are i.i.d. Bernoulli \( Be(p) \). Here \( S_l \) is the set of all possible network crash configurations at layer \( l \). Each configuration \( s_l \in S_l \) describes which neurons are crashed and which are working. We have \(|S_l| = 2^m_i\).

\[
\mathbb{E}_\Delta^{p_0 \ldots p_L} = \sum_{s_0 \in S_0} \ldots \sum_{s_L \in S_L} f(p_0, s_0, N_0) \ldots f(p_L, s_L, N_L)(\hat{y}_L - y_L)
\]

where
\[
\hat{y}_L = W_L \varphi(...) (\varphi(W_1(x \circ \xi_0) + b_1) \circ \xi_1) \ldots \circ \xi_{L-1} + b_L) \circ \xi_L
\]
\[
y_L = W_L \varphi(...) (\varphi(W_1 x + b_1) \ldots + b_L)
\]

We utilize the fact that the quantity \( f(p,n,k) = p^k(1-p)^{n-k} = p^k + O(p^k+1) \). Only those sets of crashing neurons \((s_0, \ldots, s_L)\) matter in the first order, which have one crash in total (in all layers). We denote it as \( s^*_k \in S_k \). Therefore we write
\[
\mathbb{E}_\Delta^{p_0 \ldots p_L} = \sum_{s^*_0 \in S_0} f(p_0, 1, N_0)(\hat{y}_L - y_L) + \ldots + \sum_{s^*_L \in S_L} f(p_L, 1, N_L)(\hat{y}_L - y_L)
\]

This is equivalent to a sum of individual layer crashes up to the first order on \( p \):
\[
\mathbb{E}_\Delta^{p_0 \ldots p_L} = \mathbb{E}_\Delta^{p_0} + \ldots + \mathbb{E}_\Delta^{p_L}
\]

The proof for the variance is analogous.

\( \Box \)

6 Algorithm for Certifying Fault Tolerance

In case if we consider the RHS quantities in Theorem 1 averaged over all data examples \((x, y^*)\), then the Algorithm 1 from the main paper would give a guarantee for every example: it will guarantee that \( \mathbb{P}[x, \xi][\Delta \geq \Delta + t] \) is small. Indeed, if we know that \( \mathbb{E}_x \text{Var}\Delta \) is small, we know that the total variance \( \text{Var}_x \mathbb{E}_\xi \Delta = \mathbb{E}_x \text{Var}_\xi \Delta + \text{Var}_x \mathbb{E}_\xi \Delta \) (by the law of total variance) is small as well. Indeed, the second term \( \text{Var}_x \mathbb{E}_\xi \Delta \) bounded by \( \sup_x \mathbb{E}_\xi \Delta \) which we assume to be small. In case if it is not small, it is unsafe to use the network, as even the expectation of the error is too high.
Given a small total variance \( \text{Var}_{x,\xi}\Delta \), we apply, as in Proposition 5, a Chebyshev’s inequality to the random variable \( \Delta \) over the joint probability distribution \( x, D|x \). This will give \( \mathbb{E}_{x,\xi}\Delta \geq \mathbb{E}\Delta + t \leq t^{-2}\text{Var}_{x,\xi}\Delta \). This probability indicates how likely it is that a network with crashes and random inputs will encounter a too high error.

Median aggregation of \( R \) copies works for the input distribution as well. Denote “\( y_i(x) \) is bad” as the event that the loss exceeds \( \varepsilon \) for the \( i \)th copy. We denote [True] = 1 and [False] = 0 (Iverson brackets). Now, \( \delta = P_{x,\xi}(\text{Med}(y_i(x)) \text{ bad}) = \mathbb{E}_x\mathbb{E}_\xi[\text{at least half of } y_i \text{ is bad}] \). Since the inner probability can be bounded as \( \leq t^{-2}\text{Var}\Delta(x)\exp(-R) \), taking an expectation over \( \mathbb{E}_x \) results in the quantity discussed before, \( \mathbb{E}_x\text{Var}\Delta \).

We note that it would not be possible to consider \( y_{L+1} \) to be the total loss, as it makes quantities such as \( \partial L/\partial W^ij(x,W) \) ill-defined, as they depend on \( x \) as well.

### 6.1 Analysis of the Algorithm 1

The algorithm consists of the main loop which is executed until we can guarantee the desired \((\varepsilon, \delta)\)-fault tolerance. It trains networks, and upon obtaining a good enough network with

\[
\text{Time and neurons requirements.}
\]

After each iteration of the main loop of the algorithm, the previous network can be deleted as it is no longer used. Therefore, each new iteration does not require any additional space. We only need to store the network itself and the computational graph for \( \text{Var}\Delta \), which depends on network’s gradients. The total space complexity is then \( \mathcal{O}\left(\sum_{i=1}^{L} n_i \times n_{i-1}\right) \) which is just the space to store the network’s weights.

1. \( q < 10^{-2} \). In that case the target value of \( \max|W_i|/\min|W_i| > 100 \) which is many times greater than the loss itself (by assumption \( \omega \in (-1,1) \)). In this case the algorithm increases the regularization parameter \( \mu \) to make \( q \) smaller. We assume that we double \( \mu \) each time. So by setting \( \mu^* = 1 \) the network will definitely achieve \( q > 10^{-2} \). Therefore the number of doublings of \( \mu \) is at most \( \log 1/\mu^0 \). Since \( \mu^0 \) is represented as a float, \( \log 1/\mu^0 = \mathcal{O}(1) \)

2. \( \delta_0 > 1/3 \). Since \( \delta_0 = \exp(-nqd_{KL}(\alpha|p_1)) \), it will decay exponentially with the decay of \( n_i \). In order to make \( \delta_0 < 1/3 \), we need to set \( n_qd_{KL}(\alpha|p_1) > 2 \) which leads to \( n \geq \frac{200}{d_{KL}(\alpha|p_1)} \). Since both \( \alpha \) and \( p \) are small, we take only the first-order term in \( d_{KL}(\alpha|p) = \alpha \log \alpha/p + (1 - \alpha) \log(1 - \alpha)/(1 - p) \). The last term here is \( \sim p - \alpha \geq -\alpha \). In case if \( \alpha \geq \varepsilon^2p \log\alpha/p \geq 2 \) and \( d_{KL}(\alpha|p) \geq \alpha \). Now, \( \alpha \) needs to be sufficiently small in order to guarantee the second-order term in the Taylor expansion \( \alpha^2 D_{12} \ll 1 \).

Therefore, \( \alpha^2 \sim 1/D_{12} \) and \( n \sim 200D_{12} \). If the algorithm increases \( n \) by a constant amount, it will need to take \( \mathcal{O}(D_{12}) \) operations.

3. \( R_3 > C \). In this case, we increase \( \psi \). Since this directly influences \( R_3 \) via regularization, doubling \( \psi \) every time results in a similar \( \mathcal{O}(1) \) performance as in the analysis for \( q \) and \( \mu \). Since \( R_3 \approx C \) is the number of changes that the weights make, which is \( > 1 \), and the loss is bounded by 1.

4. \( \delta = \varepsilon^{-2}\text{Var}\Delta \approx \varepsilon^{-2}\frac{C_{p1}}{n_{1}} > 1/3 \) (see Section 5) for \( C_l \) dependent on the function approximated by the NN and the continuous limit. Therefore, \( n \approx \mathcal{O}(C_{p1}/\varepsilon^2) \).

The total number of iterations is therefore \( \mathcal{O}(D_{12} + C_{p1}/\varepsilon^2) \) for \( C_l = n_{1}\text{Var}\Delta/p_1 \) for some \( n_{1} \) (this is now a property of the function being approximated and the continuous limit). We note that the constants 1/3 and \( 10^{-2} \) are chosen for the simplicity of the proof. The asymptotic behavior of the algorithm does not depend on them, as long as they are constant.

---

6 For \( \varepsilon: \log e = 1 \)

7 Technically, here we silently implied that \( \varepsilon^2p \ll 1/\sqrt{D_{12}} \) in order to make \( \alpha \geq \varepsilon^2p \), which means that we cannot implement a function with too high second derivatives in neuromorphic hardware with a constant \( p \), not matter how many neurons we take. Intuitively, this happens because for such a function, even a failure of \( \varepsilon^2p \) fraction of neurons, which is a reasonable expectation, is too large to begin with. We assume that we are fitting a function which is not like that. If we encounter this, we will see that by having a too high \( \mathbb{E}\Delta \) and the algorithm will output infeasible
Correctness: guarantee of robustness. Now we analyze correctness. We note that the algorithm is not guaranteed to find a good trade-off between accuracy and fault-tolerance. Up to this point, there is no complete theory explaining the generalization behavior of neural networks or their capacity. Therefore, we cannot give a proof for a sufficient trade-off without discovering first the complete properties of NNs capacity. We only show that the algorithm can achieve fault tolerance.

First, the condition on $E\Delta$ and $Var\Delta$ implies that the first-order terms in the expansion from $T_1$ are small enough. Now we argue that the remainder is small as well.

The condition $R_3 < C$ implies that discrete function is smooth enough to apply $\int [W_i(t,t')]dt dt' \approx C_1 < R_3 < C$ as well as $\int [W_i'(t,t')]dt dt' \approx C_2 < R_3 < C$. This means that the integral is small, which allows to bound the Riemann remainder $R_3/n_i$ from the proof of AP$_4$. This implies that there exist a continuous network $NN_c$ such that the approximation error $A$ from AP$_3$ is small. Here, the right metric is $R_3/n_i \approx C/n_i \ll 1$. The number of changes $C$ must be less than a number of neurons at a layer $n_i$.

Now, the remainder depends on the operator derivative bound $D_{12}$. By the Assumption 3 (part on $D_k$), they are small. Below we describe why this part holds.

Experimental evidence for $D_k$ being small. First, since we explicitly have a term $R_1$ for $Var\Delta \geq (\partial y/\partial y_i)^2 y_i$ (AC$_2$) in the regularizer, the first-order derivatives $D_1$ would be small for each layer by design of our algorithm. Next, second-order derivatives $\partial^2 y_i/\partial W_i^{ij} \partial W_i^{k,j'}$ are found to be small in experimental studies of the Hessian [11]. Since derivatives w.r.t. $y_i$ can be expressed via the derivatives w.r.t. weights by AP$_8$, and the continuous limit holds, $D_2$ is small as well.

General considerations for $D_k$ being small. Note, we can always renormalize $|x(t)| \leq 1$ and $|y(t)| \leq 1$ by rescaling the input and output weights. For the input layer, $D_{12}$ can be bounded via the properties of the ground truth function $y^*$, if $y$ approximates $y^*$ well. However, it is known that it can be false: the network can have a sufficient accuracy on the training dataset, but have much larger output-input derivatives $(d|out|/d|in|)$ than the ground truth function. Specifically, in the task of image recognition, we (humans) assume the problem to be quite smooth: a picture of a cat with a missing ear or whiskers is still a picture of a cat. However, it was shown [13] that modern CNNs use non-robust features. This implies that CNNs are much more sensitive to small changes in input $x$, in contrast to the smooth ground truth function $y^*$ that we want it to learn, making the bound for the derivative $D_1$ large.

For the hidden layers, it is possible that the continuous limit has large derivatives. For example, we can first apply an injective transformation $F$ with high derivatives $D_1$ would be small for each layer by design of our algorithm. Then, we have an overall smooth (identity) operator $y_L$ which, however, consists of two very non-smooth parts.

We list the following approaches that potentially could resolve this issue theoretically rather than experimentally. First, we can consider the infinite depth limit [23]. This would allow to have regularities throughout the network layers. Another approach is to study mathematically the ways that an operator can be decomposed into a hierarchical composition of operators. For example, for images, a natural decomposition of the image classification operator could first detect edges, then simple shapes, then groups them into elements pertinent to specific classes and then detects the most probable class [26]. At each stage, only robust features are used. Thus, for such a decomposition, output-hidden derivatives would be reasonable as well as the output-input ones: indeed, the decision to recognize a cat would not change significantly if some internal hidden layer features (ears or whiskers) are not present. Interestingly, just enforcing the continuous limit seems to make features more robust in our experiments, see hidden layer weights on Figures 2 and 3. Without regularization, the weights seem noisy, there are a lot of unused neurons, and even the used ones contain noisy patterns. In contrast, continuity-regularized networks seem to have first-layer weights similar to the input images. This could be an interesting research direction to continue.

We note, however, that such a study is not connected to the fault tolerance anymore, as it is a fundamental investigation into the properties of the hierarchical functions that we want to approximate, and into the properties of the neural networks which we can find by gradient descent.

The idea to construct $y = F \circ F^{-1}$ with $F^{-1}$ implementable by a neural network is taken from [8].
Figure 5: Boston-trained networks. Rank loss (worst value 0.5) of ordering networks using bounds (left 4) and relative error in predicting the value of $\Delta$ (right 4). Mean of the error $E\Delta$ (top) and standard deviation (bottom). Variable input on a fixed network (1st and 3rd columns) and variable network with fixed input (2nd and 4th columns).

7 Experimental Evaluation

Experiments were performed on a single machine with 12 cores, 60GB of RAM and 2 NVIDIA GTX 1080 cards running Ubuntu 16.04 LTS and Python/Tensorflow/Keras. We test the proposed bounds on two datasets as a proof of concept: the Boston Housing dataset (https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html) (regression) and the MNIST dataset (classification) (http://yann.lecun.com/exdb/mnist/). In addition, we use the Fashion MNIST dataset (classification, https://www.kaggle.com/zalando-research/fashionmnist). We use standard pre-trained networks from Keras (VGG, MobileNet, ..., https://keras.io/applications/).

For the Dropout experiment $p < 0.03$ is used as a threshold after which no visible change is happening. We use the unscaled version of dropout[8].

In the experiments, we use computationally tractable evaluations of the result given by $T_1$. The ”b1 bound” is the Spectral bound from $T_2$. The ”b2 bound” corresponds to $A_1T_2$ The ”b3 bound” corresponds to first-order terms from $T_1$ or the Additional Corollary [2] and the ”b4 bound” corresponds to an exact evaluation of single-neuron crashes (taking $O(n_{ij})$ forward passes).

7.1 Boston-trained networks with different initialization: rank loss with experimental error (additional)

We compare sigmoid networks with $N \sim 50$ trained on Boston Housing dataset (see ErrorComparisonBoston.ipynb). We use different inputs on a single network and single input on different networks. We compare the bounds using rank loss which is the average number of incorrectly ordered pairs. The motivation is that even if the bound does not predict the error exactly, it could be still useful if it is able to tell which network is more resilient to crashes. The second quantity is the relative error of $\Delta$ prediction, which is harder to obtain. Experimental error computed on a random subset $S'$ of all possible crashed configurations $S$ with $|S| = 2^{20} \sim 10^6$ is used as ground truth, with $|S'|$ big enough to make the expectation and variance results not change from one launch to another. The results are shown in Figure[9]. Even on this simple dataset we have only b3 and b4 giving meaningful results both for rank loss and relative error of error. The results show that b4 is always better than b3, as expected. The same is done for random networks.

---

[8] We note that it is the same as the scaled version up to the first order
### 7.2 Comparison on random networks (additional)

The same setup as for the Different Dropout Comparison experiment in the main paper is done for random networks with $N \sim 20$, ReLU activation function, see `ErrorComparisonRandom.ipynb`. The results are shown in Figure 6 and they are qualitatively the same as for Boston dataset.

![Random networks comparison](image)

**Figure 6**: Random networks. Rank loss (worst value 0.5) of ordering networks using bounds (left 4) and relative error in predicting the value of $\Delta$ (right 4). Mean of the error $E\Delta$ (top) and standard deviation (bottom). Variable input on a fixed network (1st and 3rd columns) and variable network with fixed input (2nd and 4th columns).

### 7.3 Additional charts

Table 2 presents the results of `ConvNetTest-ft.ipynb`. Figure 9 presents the results from `ComparisonIncreasingDropoutMNIST.ipynb`. Figure 10 shows results from `Regularization.ipynb`. Figure 8 shows how the algorithm’s (Main paper, Algorithm 1, `TheAlgorithm.ipynb`) state evolves over iterations. Figure 7 shows the resulting distribution of the error $\Delta$ at the output layer for the final network given by the algorithm.

![Additional charts](image)

**Figure 7**: The distribution $\Delta$ of the error in the output for the network obtained using Algorithm 1 from the main paper.
Figure 8: Evolution of the experimental error probability and the variance of the error over algorithm’s iterations. There are two stages. In the first stage (red, iterations \( \leq 4 \)) the algorithm increases continuity of the network via increasing \( \psi \). In the second stage (green, iterations \( \geq 5 \)) the algorithm increases the number of neurons \( n_1 \) and the regularization parameter \( \lambda \). First, it can be seen that at first stages the network is not continuous enough, as the algorithm makes it more continuous. This leads to an increase in the empirical probability \( \delta \) of the network outputting a loss \( > \varepsilon \) and in the increase in the gap between Theoretical \( \text{Var}\Delta \) and Experimental \( \text{Var}\Delta \). This happens because there are not enough neurons in the network. Later, as the number of neurons increases, the gap becomes smaller and the empirical probability decreases. Note that the first network (at iteration 0) empirically satisfies our fault tolerance guarantee. Nevertheless, we do not have a proof for such a network because it is not continuous enough. Therefore, in order to guarantee robustness, we need to proceed with the iterations.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \log \text{Var}\Delta )</th>
<th>( \log[\text{Parameters/Layers}] )</th>
<th>( \log[\text{Parameters}] )</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>-18.7</td>
<td>15.6</td>
<td>18.7</td>
<td>23</td>
</tr>
<tr>
<td>VGG19</td>
<td>-18.4</td>
<td>15.5</td>
<td>18.8</td>
<td>26</td>
</tr>
<tr>
<td>MobileNet</td>
<td>-9.1</td>
<td>10.7</td>
<td>15.3</td>
<td>93</td>
</tr>
</tbody>
</table>

Table 2: Comparison of bigger convolutional networks when there are faults at every layer with \( p = 10^{-2} \)

7.4 Error superposition (testing AP\(^7\) additional)

We test a random network with \( L = 4 \) on random input (see ErrorAdditivityRandom.ipynb). The error is computed on subsets of failing layers \( a, b \). Then all pairs of disjoint subsets are considered, and the relative error of \( \Delta \) estimation using linearity is computed as \( \| \Delta_{a,b} - \Delta_a - \Delta_b \| / \| \Delta_{a,b} \| \). The results (see Figure 11) show that this relative error is only few percent both for mean and variance, with better results for the mean.

7.5 Testing AP\(^9\) (additional)

The result predicts the expected error \( E_x E_\xi \frac{\partial L}{\partial y} \Delta \) to decay with the decay of the gradient of the loss. We tested that experimentally on the Boston dataset using sigmoid networks (see ErrorOnTraining.ipynb) and note that a similar result holds for ReLU\(^9\). The results are shown in Figure 12. The chart shows first that the experiment corresponds to b4, and b3 is close to b4. b3 is also equal to the result of AP\(^9\) both of which decay, as predicted.

7.6 Practical bounds

\(^9\)ReLU\( \notin C^1 \). However only a small percentage of neurons are near 0 thus approximation works for most of them.
Figure 9: Comparison of networks trained with different dropout. Shown: accuracy and $\text{Var}\Delta$ plots for different train dropout probability. Green curve shows accuracy of correct network, red shows accuracy for the crashing network. Dashed line shows train dataset and solid represent test dataset. Variance of $\Delta$ estimated by $b_4$ shown in orange, by $b_3$ in blue. Error bars are standard deviations in 10 repetitions of the experiment.

Figure 10: Horizontal axis represents increasing regularization parameter. Left vertical axis represents accuracy and has test crashing network accuracy (red) and correct network accuracy (green). Right axis shows variance of the error $\text{Var}\Delta$ estimated by bound $b_3$ used as a regularizer. Error bars are standard deviations in 5 repetitions of the experiment.

**Additional Proposition 8.** For neural network we have for any weight matrix $W_l$ and input to the $l$'th layer $y_{l-1}$:

$$\sum_{i,j} \frac{\partial y}{\partial W_{ij}^l} W_{ij}^l = \sum_j \frac{\partial y}{\partial y_{l-1}^j} y_{l-1}^j$$

The equality also holds for one particular $y_{l-1}^j$:

$$\sum_i \frac{\partial y}{\partial W_{ij}^l} W_{ij}^l = \frac{\partial y}{\partial y_{l-1}^j} y_{l-1}^j$$

**Proof.** Fix some layer $l$. The output of the network $y$ depends on the weight matrix $W_l$ and on the input to the $l$'th layer $y_{l-1}$. However we note that it only depends on their product and not on these quantities separately. Therefore we write $y = y_L(W_l, y_{l-1}, ...)$ and denote $x = y_{l-1}$, $W = W_l$ and $z = Wx$:

$$y = \eta(Wx)$$

Now we take one

$$\frac{\partial y}{\partial x_j} = \sum_i \frac{\partial \eta}{\partial z_i} \frac{\partial z_i}{\partial x_j}$$

Figure 11: Distribution of relative error in the predicted $\Delta$ in percent for 25 subset pairs. Distribution for the mean is shown in blue, for the variance in orange.
Additional Corollary 2. The error expressed in weights in the first order is as in \( \sum \delta W_{ij} \)

\[
E \Delta_L^i = -p \sum_{i,j} \frac{\partial y}{\partial W_{ij}} W_{ij}^k
\]

\[
\text{Var} \Delta_L^i = p \sum_{i} \sum_{j,k} \frac{\partial y}{\partial W_{ij}} \frac{\partial y}{\partial W_{ij}} W_{ij}^k W_{ij}^k
\]

Where \( \Delta_L^i \) is the error in case of neurons failing at layer \( L \). Layer 0 means failing input.

In other words, both mean and variance are defined by a tensor \( X_l = W_l \odot \frac{\partial y}{\partial W} \), \( z_l^i = - \sum X_l^{ij} \):

\[
E \Delta_L^i = p \sum_i z_l^i, \quad \text{Var} \Delta_L^i = p \sum_i (z_l^i)^2
\]

Proof. According to Theorem 1, the expression for mean and variance are:

\[
E \Delta \approx -p(\nabla y(x), x) = -p \sum_i \frac{\partial y}{\partial x_i} x_i
\]

\[
\text{Var} \Delta \approx ((\nabla y(x))^2, x^2) = p \sum_i \left( \frac{\partial y}{\partial x_i} x_i \right)^2
\]
Then by Additional Proposition 8, the expression used can be rewritten as (note the index swap $i \leftrightarrow j$):

$$\frac{\partial y}{\partial x_i} x_i = \sum_j \frac{\partial y}{\partial W_{ji}} W_{ji}$$

Then for the mean it is:

$$E\Delta = -p \sum_i \sum_j \frac{\partial y}{\partial W_{ji}} W_{ji}$$

and the indices can be swapped again for the sake of notation. For the variance we write the square as the inner sum repeated twice with different indices $k$ and $j$:

$$Var\Delta = p \sum_i \sum_j \sum_k \frac{\partial y}{\partial W_{ji}} W_{ji} \frac{\partial y}{\partial W_{ki}} W_{ki}$$

This gives the first statements. Then we notice that both expressions depend on the values of tensor $\frac{\partial y}{\partial W} \odot W$ but not on $W$ or $\frac{\partial y}{\partial W}$ alone. We therefore define $X = W \odot \frac{\partial y}{\partial W}$ and $z^i_j = -\sum_i X_{ij}$ (sum over the first index) and rewrite

$$E\Delta = -p \sum_i \sum_j X_{ij} = p \sum_j z^j$$

$$Var\Delta = p \sum_{ijk} X_{ji} X_{ki} = p \sum_i \left( \sum_j X_{ji} \right)^2 = p \sum_i (-z_i)^2$$

\[\square\]

**Additional Proposition 9.** For a neural network with $C^1$ activation function we have for a particular input $x$ in the first order as in $T1$:

$$E_\xi \Delta_{L+1} = -p(\nabla_W L, W)$$

**Proof.** Take the expression $A = E_\xi \Delta_{L+1} = \left( \frac{\partial L}{\partial y}, E_\xi \Delta \right)$ and consider $E_\xi \Delta$. By Additional Corollary $2$ we have

$$E_\xi \Delta = -p \sum_{ij} \frac{\partial y}{\partial W_{ij}} W_{ij}$$

Therefore,

$$A = -p \left( \frac{\partial L}{\partial y} \sum_{ij} \frac{\partial y}{\partial W_{ij}} W_{ij} \right) = -p \sum_{ij} W_{ij} \left( \frac{\partial L}{\partial y} \frac{\partial y}{\partial W_{ij}} \right)$$

By the chain rule for a single input $x$ we have

$$\frac{\partial L}{\partial W_{ij}} = \left( \frac{\partial L}{\partial y}, \frac{\partial y}{\partial W_{ij}} \right)$$

We plug that back in:

$$A = -p \sum_{ij} \frac{\partial L}{\partial W_{ij}} W_{ij}$$

This is the exact statement from the proposition.  

\[\square\]

8 Unused extra results
Skip-connections. If we consider a model with skip-connections (which do not fit Definition [1]) with faults at every node, we expect that an assumption similar to $A[2]$ would lead to a result similar to $[4]$. However, we did not test our theory on models with skip-connections.

Another idea for fault tolerance. In case if we know $p$ exactly, we can compensate for $E \Delta$ by multiplying each neuron’s output by $(1 - p)^{-1}$. In this way, the mean input would be preserved, and the output will be unbiased. However, this only works in case if we know exactly $p$ of the hardware.

Additional Proposition 10 (Variance for bound b1, not useful since the bound is not tight. Done in a similar manner as the mean bound in [9]). The variance of the error $\text{Var} \Delta$ is upper-bounded as

$$\text{Var} \Delta \leq \sum_{i=1}^{n} a_i \prod_{j=i+1}^{n} b_j$$

for

$$a_L = C_L^2 pL (\alpha_L + pL \beta_L) + 2C_L pL (1 - pL) \beta_L E \Delta_{L-1},$$

$$b_L = (1 - pL) (\alpha_L + (1 - pL) \beta_L),$$

$$C_L = \max\{\| y_i \| \},$$

$$\alpha_L = \max\{ w^{(L+1)}(1) || w^{(L+1)}(1) ||_1 - (w^{L+1}, w^{(L+1)}) \},$$

$$\beta_L = \max\{ || w^{(L+1)}(1)_i || \},$$

Proof. In the notation of [9], consider $| \Delta_L | = | F_{\text{new}} - F_{\text{fail}} | = | \sum_{i=1}^{N_L} w_i^{(L+1)} (y_j^{(1)}(L) - \hat{y}_j^{(1)}(L)) |, E \Delta_L \leq pL || w^{(L+1)} ||_1 C_L + K || w^{(L+1)} ||_1 (1 - pL) E \Delta_{L-1},$ where $\hat{y}_j^{(1)}$ is the corrupted output from layer $L$ and $K$ is the activation function Lipschitz constant. Thus, $E \Delta_L \leq \sum_{l=1}^{L} pL || w^{(L+1)} ||_1 C_l K^{L-l} \prod_{s=l+1}^{L} (1 - p_s) || w^{(s)} ||_1 = \sum_{l=1}^{L} pL || w^{(L+1)} ||_1 C_l \prod_{s=l+1}^{L} K (1 - p_s) || w^{(s)} ||_1$.

1. Variance: $\text{Var} \Delta = E(\Delta^2) - (E \Delta)^2 \leq E(\Delta^2)$, here we will calculate $E \Delta^2$:

2. $E \Delta^2 = E \sum_{i=1}^{N_L} \sum_{j=1}^{N_L} w_j^{(2)} j_j^{(1)} (1 - \xi_j^{(1)}) w_j^{(2)} j_j^{(1)} (1 - \xi_j^{(1)}) \leq f_1 C_L^2 w_j^2 + C_L^2 \frac{f_2 - f_1}{N_L - N_1} \sum_{j \neq j} w_j w_j$

$$\frac{f_2 - f_1}{N_L^2 - N_1} \leq \frac{f_2}{N_L^2} = p^2, \sum_{j \neq j} w_j w_j = \| w \|_2^2 = \| w \|_2^2.$$

Therefore, $E \Delta^2 \leq C_L^2 p \| w \|_2^2 + C_L^2 p^2 \| w \|_2^2 - \| w \|_2^2$.

3. $E \Delta_1$ $\Delta_1 = E \sum_{i=1}^{N_L} \sum_{j=1}^{N_L} w_j^{(2)} j_j^{(1)} (1 - \xi_j^{(1)}) w_j^{(2)} j_j^{(1)} (1 - \xi_j^{(1)}) \leq f_1 C_L^2 w_i^2 w_i^2 + C_L^2 \frac{f_2 - f_1}{N_L - N_1} \sum_{j \neq j} w_j w_j$

$$\leq pC_L^2 \| w \|_2^4 + p^2 C_L^2 (\| w \|_2^4 - \| w \|_2^2).$$

4. $E \Delta_2^2 = \sum_{i=1}^{N_L} \sum_{i \neq j} (w_i^{(L+1)})^2 E(y_i^{(L)} - y_i^{(L)} \xi_i^{(L)})^2 + \sum_{i \neq j} w_i^{(L+1)} w_j^{(L+1)} E \left[ (y_i^{(L)} - y_i^{(L)} \xi_i^{(L)}) (y_j^{(L)} - y_j^{(L)} \xi_j^{(L)}) \right]$.

5. $E \Delta_2^2 \leq \sum_{i=1}^{N_L} w_i^{(L+1)} w_i^{(L+1)} E(y_i^{(L)} - y_i^{(L)} \xi_i^{(L)}) + \sum_{i \neq j} w_i^{(L+1)} w_j^{(L+1)} E \left[ (y_i^{(L)} - y_i^{(L)} \xi_i^{(L)}) (y_j^{(L)} - y_j^{(L)} \xi_j^{(L)}) \right]$.

6. $P_1 = pC_L^2, Q_1 = p^2 C_L^2$.

7. $P_L = E(y - \hat{y} \xi)^2 = E_{\xi=0} \xi^2 \mathbb{P} \{ \xi = 0 \} + K^2 E \Delta_{L-1} \mathbb{P} \{ \xi = 1 \} \leq C_L^2 pL + K^2 E \Delta_{L-1} (1 - pL)$.

8. $Q_L = E(y - \hat{y} \xi)/(y - \hat{y} \xi') \leq C_L^2 E_{\xi=0} \xi^2 (1 - pL) K L E(\Delta_{L-1} + \Delta_{L-1}^2) + (1 - pL)^2 K^2 E \Delta_{L-1} \Delta_{L-1}^2.$
9. Consider a recurrence \( x_1 = a_1, \ x_n = a_n + b_n x_{n-1} \). Then \( x_n = \sum_{i=1}^{n} a_i \prod_{j=i+1}^{n} b_j \).

10. Define \( \alpha_L = \max\{w(L+1), w'(L+1)\} \) and \( \beta_L = \max\{\|w(L+1)\|, \|w'(L+1)\|_1 - (w(L+1), w'(L+1))\} \). Then

\[
\mathbb{E} \Delta'_L \Delta L \quad \mathbb{E} \Delta^2_L \leq P_L \alpha_L + Q_L \beta_L
\]

11. Thus, \( \mathbb{E} \Delta^2_L \leq a_L + b_L \mathbb{E} \Delta^2_{L-1} \), where

\[
a_L = C^2 L p_L (\alpha_L + p_L \beta_L) + 2K C L p_L (1 - p_L) \beta_L \mathbb{E} \Delta_{L-1},
\]

\[
b_L = K^2 (1 - p_L) (\alpha_L + (1 - p_L) \beta_L).
\]

Therefore,

\[
\mathbb{E} \Delta^2_L \leq \sum_{l=1}^{L} (C^2 L p_l (\alpha_l + p_l \beta_l) + 2K C L p_l (1 - p_l) \beta_l \mathbb{E} \Delta_{l-1}) \prod_{\nu=l+1}^{L} K^2 (1 - p_\nu) (\alpha_\nu + (1 - p_\nu) \beta_\nu)
\]

The goal of this proposition was to give an expression for the variance in a similar manner as it is done for the mean in [9]. However this proposition did not make it to the article because bound b1 was not showing any good experimental results.

9 Introduction into fault tolerance for neuromorphic hardware

Neuromorphic hardware (NH). Given the amount of processing power required for modern Machine Learning applications, emerging hardware technologies are nowadays reviving the neuromorphic project and its promise to cut off energy consumption of machine learning by several orders of magnitude [10]. Neuromorphic implementation of an NN is a physical device where each neuron corresponds to a piece of hardware, and neurons are physically linked forming weights. Thus, the computation is done (theoretically) at the speed of light, compared to many CPU/GPU cycles. The surge in performance could arguably even exceed the one that followed the switch from training on CPUs to training on GPUs and TPUs [3]. Recent results on neuromorphic computing report on concrete successes such as milliwatt image recognition [10] or basic vowel recognition using only four coupled nano-oscillators [22]. Since the components of a neuromorphic network are small [25] and unreliable [17], there are crashes in individual neurons or weights inside the network [25] [17]. They lead to a performance degradation. A failure within a neuromorphic architecture involved in a mission-critical application could be disastrous. Hence, fault tolerance in neural networks is an important concrete Artificial Intelligence (AI) safety problem [4]. In terms of fault tolerance, the unit of failure in these architectures is fine-grained, i.e., an individual neuron or a single synapse, with failure mode frequently being a complete crash. This is in contrast with the now classical case of a neural network as a software deployed on a single machine where the unit of failure is coarse-grained, i.e., the whole machine holding the entire neural network. For instance, in the popular distributed setting of ML, the so-called parameter-server scheme [16], the unit of failure is a worker or a server, but never a neuron or a synapse. Whilst very important, fine-grained fault tolerance in neural networks has been overlooked as a concrete AI safety problem.

References


