Supplement: How Sampling Impacts the Robustness of Stochastic Neural Networks

Sina Däubener and Asja Fischer
Department of Computer Science
Ruhr University Bochum, Germany
{sina.daeubener, asja.fischer}@rub.de

Contents

A Proofs

In this section we proof the theorems of the main paper and recap the needed definitions to do so.

Definition A.1 (Maximum magnitude attack). Given a multi-class classifier $f(\cdot)$, a loss function $\mathcal{L}(\cdot,\cdot)$, ℓ_p-norm $\|\cdot\|_p$ and a given perturbation strength η the optimization problem for a maximum magnitude attack can be written as

$$\textit{maximize } \ \mathcal{L}(f(x+\delta), y)\ ,\ w.r.t.\ \ \delta \ \ s.t.\ \|\delta\|_p \leq \eta \ . \tag{1}$$

Theorem A.1 (Sufficient and necessary robustness condition for linear classifiers). *Let $f: \mathbb{R}^d \times \Omega^h \rightarrow \mathbb{R}^k$ be a stochastic classifier with linear discriminant functions and $f^{\mathcal{A}}$ and $f^{\mathcal{I}}$ be two MC estimates of the classifier. Let $x \in \mathbb{R}^d$ be a data point with label $y \in \{1,\ldots,k\}$ and $\arg\max_c f_c^{\mathcal{A}}(x) = \arg\max_c f_c^{\mathcal{I}}(x) = y$, and let $x_{adv} = x + \delta^{\mathcal{A}}$ be an adversarial example computed for solving the minimization problem* (1) *for $f^{\mathcal{A}}$. It holds that $\arg\max_c f_c^{\mathcal{I}}(x+\delta^{\mathcal{A}}) = y$ if and only if*

$$\min_{c\neq y} \tilde{r}_c^{\mathcal{I}} > \|\delta^{\mathcal{A}}\|_2 \ \ , \textit{with} \tag{2}$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$$\tilde{r}_c^{\mathcal{I}} = \begin{cases} \infty \ , & \textit{if} \ \ \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) = \frac{\langle -\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)), \delta^{\mathcal{A}} \rangle}{\|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \|\delta^{\mathcal{A}}\|_2} \leq 0 \\ \frac{f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)}{\|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}})} \ , & \textit{otherwise} \ \ , \end{cases}$$

where $\alpha_c^{\mathcal{I},\mathcal{A}}$ is the angle between $-\nabla_x(f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))$ and $\delta^{\mathcal{A}}$.

Proof. An adversarial attack on $f^{\mathcal{I}}$ with the adversarial example $x + \delta^{\mathcal{A}}$ is not successful iff $\forall c \in \{1, 2, \ldots, k\}, c \neq y$:

$$f_y^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) - f_c^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) > 0 \ \ .$$

With Taylor expansion around x we can rewrite $f_y^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) - f_c^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right)$ as

$$\begin{aligned} & f_y^{\mathcal{I}}(x) + \langle \nabla_x f_y^{\mathcal{I}}(x), \delta^{\mathcal{A}} \rangle - f_c^{\mathcal{I}}(x) - \langle \nabla_x f_c^{\mathcal{I}}(x), \delta^{\mathcal{A}} \rangle \\ =& f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x) + \langle \nabla_x f_y^{\mathcal{I}}(x) - \nabla_x f_c^{\mathcal{I}}(x), \delta^{\mathcal{A}} \rangle \\ =& f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x) - \|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \|\delta^{\mathcal{A}}\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) \end{aligned} \tag{3}$$

where $\alpha_c^{\mathcal{I},\mathcal{A}} := \angle(-\nabla_x(f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)), \delta^{\mathcal{A}})$. We can distinguish two cases for each c.

Case 1: $\cos(\alpha_c^{\mathcal{I},\mathcal{A}}) \leq 0$. In this case last term of eq. (3) is negative or zero and thus

$$f_y^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) - f_c^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) \geq f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x) > 0 \ \ ,$$

where the second inequality holds since $\arg\max_c f_c^{\mathcal{I}}(x) = y$.

Case 2: $\cos(\alpha_c^{\mathcal{I},\mathcal{A}}) > 0$. In this case the last term of equation eq. (3) is positive and thus

$$f_y^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) - f_c^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) \leq f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x) \ \ .$$

As we see from rearranging eq. (3), it holds $f_y^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) - f_c^{\mathcal{I}}\left(x + \delta^{\mathcal{A}}\right) > 0$ if

$$\tilde{r}_c^{\mathcal{I}} := \frac{f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)}{\|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}})} > \|\delta^{\mathcal{A}}\|_2 \ \ . \tag{4}$$

For each class c either case 1 holds and we define $\tilde{r}_c^{\mathcal{I}} := \infty$, or condition (4) is fulfilled, which yields the condition stated in the theorem. ☐

In the main part of the paper we relaxed the linear classifier assumption by L-smoothness, which was defined as follows:

Definition A.2 (L-smoothness [Yang et al., 2022]). A differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is L-smooth, if for any $x_1, x_2 \in \mathbb{R}^d$ and any output dimension $c \in \{1, \ldots, k\}$:

$$\frac{\|\nabla_{x_1} f_c(x_1) - \nabla_{x_2} f_c(x_2)\|_2}{\|x_1 - x_2\|_2} \leq L \ \ .$$

Next we restate one property of L-smooth functions which we will use in our proof of theorem 2.

Proposition A.2 (Bubeck [2015]). *Let f be an L-smooth function on $\mathbb{R}^n$. For any $x, y \in \mathbb{R}^n$ it holds:*

$$|f(y) - f(x) - \langle \nabla_x f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|_2^2 \ \ .$$

Proof. From the fundamental theorem of calculus we know that for a differentiable function f it holds that $f(y) - f(x) = \int_x^y \nabla_t f(t) dt$. By substituting $x_t = x + t(y - x)$ we see that $x_0 = x$ and

$x_1 = y$ and thus we can write $f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^T \cdot (y-x)dt$. This allows the following approximations

$$\begin{aligned}
&|f(y) - f(x) - \langle \nabla_x f(x), y - x\rangle| \\
=&|\int_0^1 \nabla f(x + t(y-x))^T \cdot (y-x)dt - (\nabla_x f(x))^T (y-x)| \\
\leq& \int_0^1 |(\nabla f(x + t(y-x)) - \nabla_x f(x))^T) \cdot (y-x)|dt \\
\overset{\text{Cauchy-Schwarz}}{\leq}& \int_0^1 \|\nabla f(x + t(y-x)) - \nabla_x f(x)\|_2 \cdot \|y-x\|_2 dt \\
\overset{L\text{-smoothness}}{\leq}& L \cdot \|y-x\|_2^2 \cdot \int_0^1 t dt \\
=& \frac{L}{2} \cdot \|y-x\|_2^2 \ .
\end{aligned}$$

□

Theorem A.3 (Sufficient condition for the robustness of an L-smooth stochastic classifier)**.** *Let* $f : \mathbb{R}^d \times \Omega^h \to \mathbb{R}^k$ *be a stochastic classifier with L-smooth discriminant functions and* $f^{\mathcal{A}}$ *and* $f^{\mathcal{I}}$ *be two MC estimates of the prediction. Let* $x \in \mathbb{R}^d$ *be a data point with label* $y \in \{1, \dots, k\}$ *and* $\arg\max_c f_c^{\mathcal{A}}(x) = \arg\max_c f_c^{\mathcal{I}}(x) = y$*, and let* $x_{adv} = x + \delta^{\mathcal{A}}$ *be an adversarial example computed for solving the minimization problem* (1) *for* $f^{\mathcal{A}}$*. It holds that* $\arg\max_c f_c^{\mathcal{I}}(x + \delta^{\mathcal{A}}) = y$ *if*

$$\min_{c \neq y} r_c^{\mathcal{I}} > \|\delta^{\mathcal{A}}\|_2 \ ,$$

with

$$r_c^{\mathcal{I}} = \begin{cases} \infty \ , \textit{if } \|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 \leq 0, \\ \frac{f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)}{\|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2} \ , \textit{else} \end{cases}$$

and

$$\cos(\alpha_c^{\mathcal{I},\mathcal{A}}) = \frac{\langle -\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)), \delta^{\mathcal{A}} \rangle}{\|\nabla_x (f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x))\|_2 \cdot \|\delta^{\mathcal{A}}\|_2} \ .$$

Proof. For better readability we write $f_{y-c}^{\mathcal{I}}(x) := f_y^{\mathcal{I}}(x) - f_c^{\mathcal{I}}(x)$. Using the result from proposition A.2 and reordering the terms, we get the following lower bound:

$$\begin{aligned}
&f_{y-c}^{\mathcal{I}}(x + \delta^{\mathcal{A}}) \\
&\geq f_{y-c}^{\mathcal{I}}(x) + \langle \nabla_x (f_{y-c}^{\mathcal{I}}(x)), \delta^{\mathcal{A}} \rangle - \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2^2 \\
&= f_{y-c}^{\mathcal{I}}(x) - \langle -\nabla_x (f_{y-c}^{\mathcal{I}}(x)), \delta^{\mathcal{A}} \rangle - \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2^2 \\
&= f_{y-c}^{\mathcal{I}}(x) - \| - \nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \|\delta^{\mathcal{A}}\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) - \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2^2 \\
&= f_{y-c}^{\mathcal{I}}(x) - \left(\|\nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 \right) \cdot \|\delta^{\mathcal{A}}\|_2 \ . \qquad (5)
\end{aligned}$$

If eq. (5) is bigger than zero, the attack cannot be successful. Hence,

$$f_{y-c}^{\mathcal{I}}(x) - \left(\|\nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 \right) \cdot \|\delta^{\mathcal{A}}\|_2 \overset{!}{>} 0 \qquad (6)$$

$$f_{y-c}^{\mathcal{I}}(x) > \left(\|\nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 \right) \cdot \|\delta^{\mathcal{A}}\|_2 \ . \qquad (7)$$

Case 1: $\|\nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 > 0$**.** Transforming eq. (7) leads to

$$\frac{f_{y-c}^{\mathcal{I}}(x)}{\|\nabla_x f_{y-c}^{\mathcal{I}}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2} > \|\delta^{\mathcal{A}}\|_2 \ .$$

Case 2 : $\|\nabla_x f^{\mathcal{I}}_{y-c}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 = 0$**.** In this case eq. (6) is trivially fulfilled because $f^{\mathcal{I}}_{y-c}(x) > 0$ per definition.

Case 3: $\|\nabla_x f^{\mathcal{I}}_{y-c}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2 < 0$**.** For this case we get, that

$$-\frac{f^{\mathcal{I}}_{y-c}(x)}{|\|\nabla_x f^{\mathcal{I}}_{y-c}(x)\|_2 \cdot \cos(\alpha_c^{\mathcal{I},\mathcal{A}}) + \frac{L}{2} \cdot \|\delta^{\mathcal{A}}\|_2|} < \|\delta^{\mathcal{A}}\|_2 \ ,$$

which is always guaranteed based on the initial assumption that the benign input was classified correctly, which concludes the proof. ☐

The following proposition relates to footnote 5 from the main paper. We show, that the interval, in which the expectation of the gradient norm lies, decreases to the true length of μ with reducing the covariance.

Proposition A.4. *Let X be an n-dimensional random vector following a multivariate normal distribution with mean vector μ and diagonal covariance matrix Σ. Then the expectation of $\|X\|_2$ can be upper and lower bounded by*

$$\|\mu\|_2 \leq \mathbb{E}[\|X\|_2] \leq \sqrt{\|\mu\|_2^2 + tr\,(\Sigma)} \ .$$

Proof. We first look at the lower bound which by convexity of the norm and Jensen inequality can be derived via

$$\mathbb{E}[\|X\|_2] \leq \|\mathbb{E}[X]\|_2 = \|\mu\|_2 \ .$$

Let $X' \sim \mathcal{N}(0, \mathbf{1}_n)$, with $\mathbf{1}_n$ an $n \times n$-dimensional unit matrix. With Jensen inequality and concavity of the square-root function we derive the upper bound

$$\begin{aligned}
\mathbb{E}[\|X\|_2] &= \mathbb{E}\left[\|\mu + X' \cdot \Sigma^{1/2}\|_2\right] \\
&\leq \left(\mathbb{E}\left[\left\|\mu + X' \cdot \Sigma^{1/2}\right\|_2^2\right]\right)^{1/2} \\
&= \left(\|\mu\|_2^2 + 2\mu^T \Sigma^{1/2} \mathbb{E}[X'] + \mathbb{E}\left[X'^T {\Sigma^{1/2}}^T \Sigma^{1/2} X'\right]\right)^{1/2} \\
&= \sqrt{\|\mu\|_2^2 + tr\,(\Sigma)} \ .
\end{aligned}$$

☐

B Additional information on datasets, models and training

In the following we describe additional details on the datasets, models and training procedures which were not stated in the main paper due to space restrictions. Additionally, please find the code for the results in the main paper attached in the supplementary material.

B.1 Datasets

We used three well know datasets: FashionMNIST [Xiao et al., 2017], CIFAR10 and CIFAR100 [Krizhevsky et al.], which consist out of 60,000 training and 10,000 test images of dimension 28×28 or $32 \times 32 \times 3$ in case of CIFAR where each image is uniquely associated to one out of 10 or 100 possible labels. We took all datasets from the torchvision package with the predefined training and test split.

B.2 Models trained on FashionMNIST

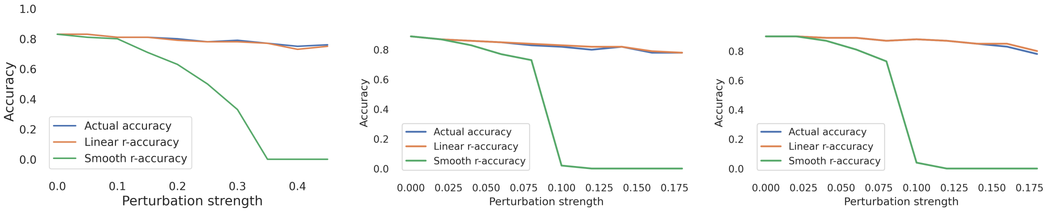
For training the BNN and IM we used the exact same hyperparameters. First, we assumed a standard normal prior decomposed as matrix variate normal distributions for the BNN and approximated the posterior distribution via maximizing the evidence lower bound (ELBO). For the IM we added a Kullback-Leibler distance from the trained parameter distribution to a standard matrix variate normal distribution as a regularization term. For both models we used a batch size of 100 and trained for 50 epochs with Adam [Kingma and Ba, 2015] and an initial learning rate of 0.001. To leverage the difference between IM and BNN we used 5 samples to approximate the expectation in the ELBO/IM-objective. We used the same learning rate, batch size, optimizer and amount of epochs for training the stochastic input networks. During a forward pass in training we created and used five noisy versions of each input, where the noise was drawn from a centered Gaussian distribution with variance 0.05 or 0.1 and the average prediction was fed into the cross-entropy loss.

B.3 Models trained on CIFAR10

As stated in the main part, we used the wide ResNet [Zagoruyko and Komodakis, 2017] of depth 28 and widening factor 10 provided by `https://github.com/meliketoy/wide-resnet.pytorch` with dropout probabilities 0.3 and 0.6 and also used the learning hyperparameters provided with the code which are: training for 200 epochs with batch size 100, stochastic gradient descent as optimizer with momentum 0.9, weight decay 5e-4 and a scheduled learning rate decreasing from an initial 0.1 for epoch 0-60 to 0.02 for 60-120 and lastly 0.004 for epochs 120-200.

C Additional experimental results

In this section we present the results which were not shown in the main part due to space restrictions.



(a) IM (b) dr 0.3 (c) dr 0.6

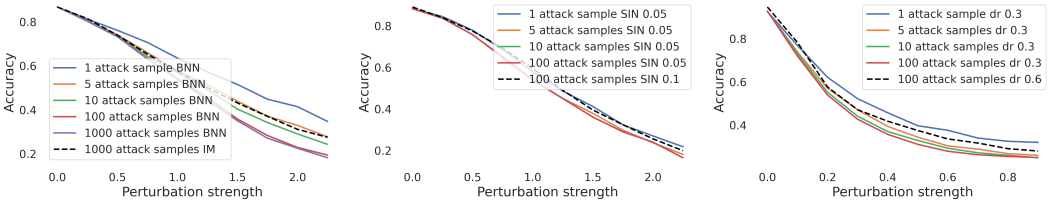
Figure 1: Adversarial accuracy of the a) smoothed IM on FashionMNIST and b),c) smoothed ResNet with dropout probability 0.3 and 0.6 on CIFAR10 vs percentage of images for which $\min_c r_c^{\mathcal{I}} > \|\delta^{\mathcal{A}}\|_2$ (smooth) and $\min_c \tilde{r}_c^{\mathcal{I}} > \|\delta^{\mathcal{A}}\|_2$ (linear) for 100 images from the respective test sets. Attacks were conducted on the smoothed classifier with using 10 attack samples.

C.1 Complementary experiments on accuracy of robustness conditions

For completeness we attached the results on the transferability of our derived sufficient conditions to: the IM on FashionMNIST and the two ResNet with dropout probability 0.3 and 0.6 on CIFAR10. For the BNN we used the same setting as described in the main paper and derive similar results (c.f. figure 1): while the percentage of samples fulfilling the condition $\min_c r_c^{\mathcal{I}} > \|\delta^{\mathcal{A}}\|_2$ approaches zero with growing perturbation strength the percentage of samples fulfilling the condition from theorem 1 in the main paper closely matches the real adversarial accuracy in a narrow environment. For models on CIFAR10 we had to adapt the noise added for the smooth classifier to 0.01 and reduce the amount of samples during inference to 50 such that it fits on one GPU.

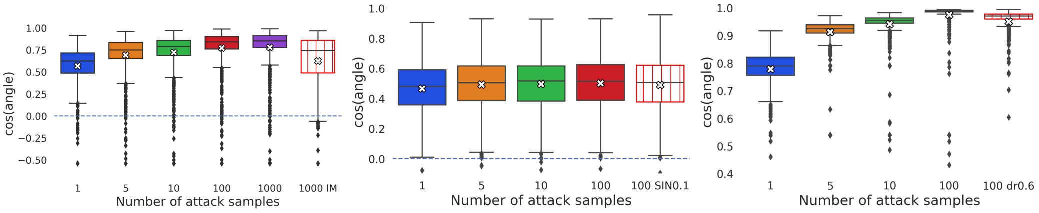
C.2 Complementary experiments on stronger attacks

We first present the results not shown in the main paper. That is, we investigate the accuracy under FGM attack with an increasing amount of samples used during the attack (c.f. figure 2). Similar to the observations in the main paper, the accuracy under attack is reduced by an increased amount of samples. However, we observe only a very small decrease in accuracy when increasing the amount of samples from 100 and 1,000 for the BNN, from 1 to 5 or above for the SIN 0.05 and when using 5 instead of 10 or 100 samples for the attack on the ResNet trained with dropout probability 0.3. This observation is mirrored by the reduction of $\cos(\alpha_c^{\mathcal{I},\mathcal{A}})$ displayed in figure 3 where we observe an increase of the cosine boxplots which matches the decrease of the adversarial accuracy when taking more samples during the attack.



(a) BNN (b) SIN 0.05 (c) dr 0.3

Figure 2: Accuracy under FGM attack for a) the BNN and b) the SIN 0.05 on FashionMNIST and c) ResNet with dropout probability 0.3 on CIFAR10 for different perturbation strengths and amount of samples used for calculating the attack. During inference we used 100 samples.

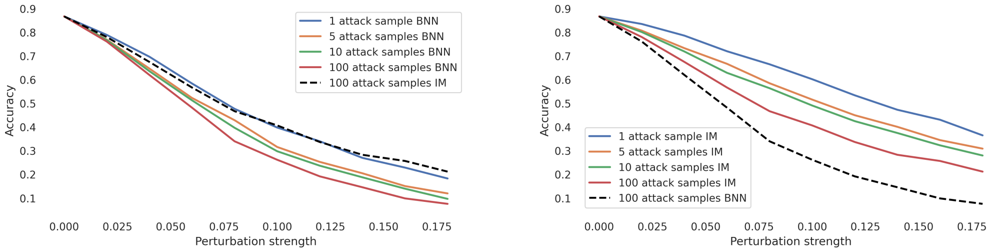


(a) BNN (b) SIN 0.05 (c) dr 0.3

Figure 3: Cosine of the angle for the first 1,000 test set images from the FashionMNIST and CIFAR10 test set for an a) BNN, b) SIN 0.05 and c) ResNet trained with dropout probability 0.3 when attacked with different amounts of samples and attack strength 1.5 and 0.3 with FGM respectively. White crosses indicate the mean value.

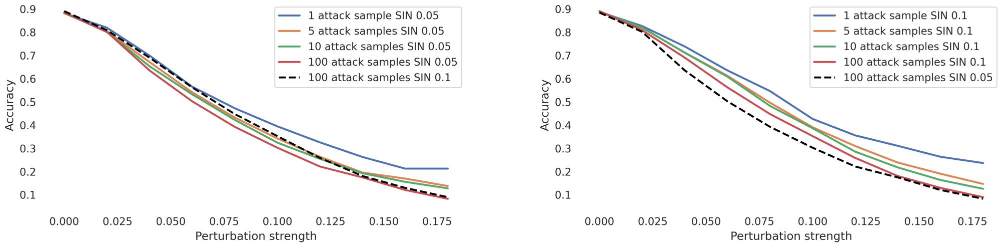
C.2.1 Attacks with FGSM (ℓ_∞- norm)

All adversarial examples in the main part of the paper were based on an ℓ_2-norm constraint, which we chose for the nice geometric distance interpretation. However, the first proposed attack scheme [Goodfellow et al., 2015] was based on ℓ_∞-norm, which we test in the following. In figure 4 and 5 we see the respective results on the different data sets. For all models the robustness is decreased with multiple samples and models with higher prediction variance also have a higher accuracy under this attack. Note, that the values of the perturbation strength are not comparable to the values under ℓ_2-norm constraint, since $\|x\|_\infty \leq \|x\|_2$.



(a) BNN

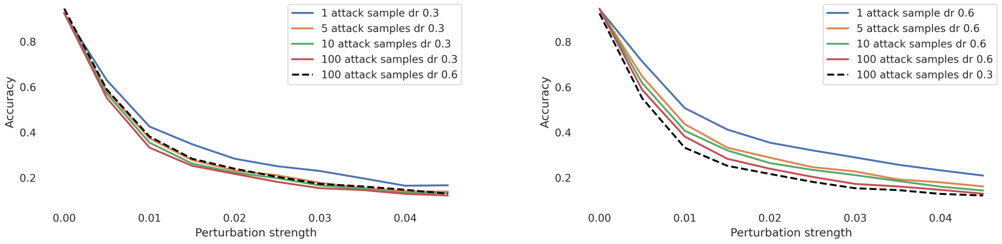
(b) IM



(c) SIN 0.05

(d) SIN 0.1

Figure 4: Accuracy under FGSM attack under ℓ_∞-norm constraint for a) the BNN, b) the IM, c) SIN 0.05 and d) SIN 0.1 on the first 1,000 test set images from FashionMNIST for different perturbation strengths and amount of samples used for calculating the attack. Predictions during inference are based on 100 samples.



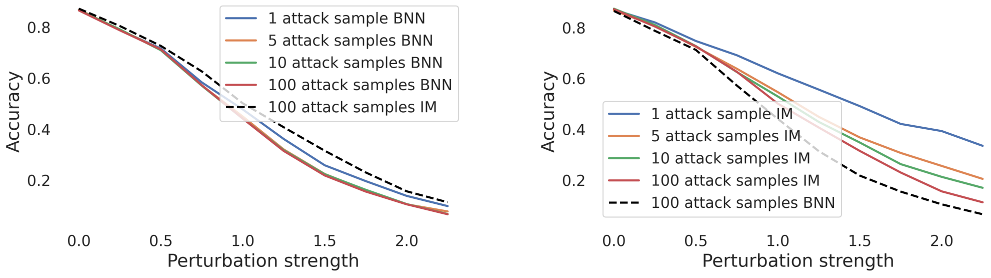
(a) droprate 0.3

(b) droprate 0.6

Figure 5: Accuracy under FGSM attack under ℓ_∞-norm constraint for the ResNet model with a dropout probability a) of 0.3 and b) of 0.6 on the first 1,000 test set images from CIFAR10 for different perturbation strengths and amount of samples used for calculating the attack. Predictions during inference are based on 100 samples.

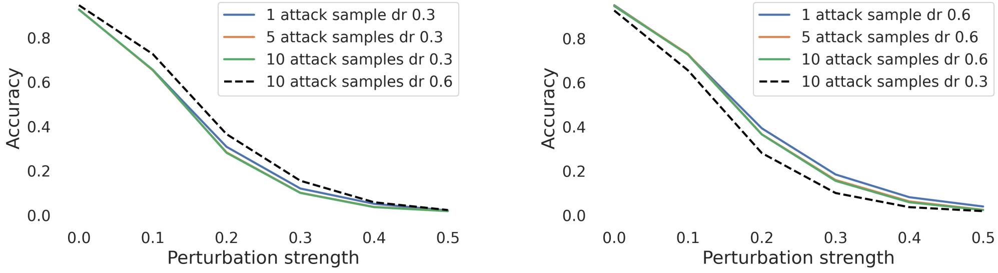
C.2.2 Attacks with PGD

Projected gradient descent [Madry et al., 2018] is a strong iterative attack, where multiple small steps of size ν of fast gradient method are applied. Specifically, we used the same ℓ_2-norm length constraint on η as in the experiments of the main part of the paper but chose step size $\nu = \eta/50$ and 100 iterations. Note that at each iteration a new network is sampled such that for an attack based on 1 samples, 100 different attack networks were seen, for an attack based on 5 samples 500 different networks and so on. In figure 6 and 7 we see that the overall accuracy is decreased compared to the results for FGM, but still, IM has a higher accuracy under attack than the BNN and so does the ResNet with a higher dropout probability. Further, the attacks get stronger with taking more samples, so the general observations made in the main paper also hold for strong attacks and are not due to a sub-optimal attack.



(a) BNN (b) IM

Figure 6: Accuracy under PGD attack with 100 iterations for a) the BNN and b) the IM on the first 1,000 test set images from FashionMNIST for different perturbation strengths and amount of samples used for calculating the attack.



(a) droprate 0.3 (b) droprate 0.6

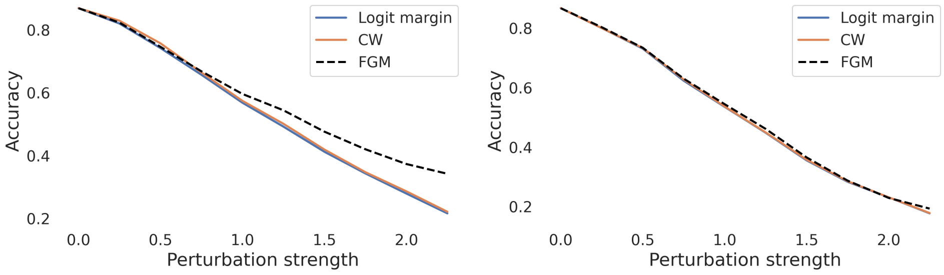
Figure 7: Accuracy under PGD attack with 100 iterations for the ResNet with a dropout probability of a) 0.3 and b) of 0.6 on CIFAR10 for different perturbation strengths and amount of samples used for calculating the attack.

C.3 Discussing the impact of extreme prediction values

Attacks based on softmax predictions can be unsuccessful for deterministic and stochastic neural networks alike when encountering overly confident predictions. That is, predictions where the softmax output is equal to 1, since these lead to zero gradients. A practical solution to circumvent this problem is to calculate the gradient based on the logits [Carlini and Wagner, 2017]. This approach is feasible for classifier whose predictions do not depend on the scaling of the output, that is, outputs which are equally expressive in both intervals $[0, 1]$ and $[-\infty, \infty]$. In Bayesian neural networks, where $f_c(x, \Theta)$ is per definition a probability the shortcut over taking the gradient over logits leads to distorted gradients. For completeness, we nevertheless look at the performance of an attack based on the logits for the two different models trained on FashionMNIST with softmax outputs: IM and BNN. We conducted an adversarial FGM attack based on 100 samples, but instead of using the cross-entropy loss we used the Carlini-Wagner (CW) loss [Carlini and Wagner, 2017] on the averaged logits, given by:

$$CW(x, \Theta^{\mathcal{A}}) = \max\left(\max_{i \neq t}(Z(x, \Theta^{\mathcal{A}})_i) - Z(x, \Theta^{\mathcal{A}})_t, 0\right) \quad ,$$

where $Z(x, \Theta^{\mathcal{A}})_t = \frac{1}{S^{\mathcal{A}}} \sum_{s=1}^{S^{\mathcal{A}}} Z(x, \theta_s)_t$ is an arbitrary averaged logit of an output node for input x. In figure 8 it is shown, that the attack with the CW loss on the infinite mixture model improves upon the original attack scheme (FGM), whereas it did not improve the attack's success for the BNN. We additionally conducted an attack based on the logit margin loss $\mathcal{L}(x + \delta^{\mathcal{A}}, y) = -(\min_{c \neq y} f_{y-c}(x))$ equivalent to the attack conducted on the SIN, but we found that it performs similar to the CW loss.

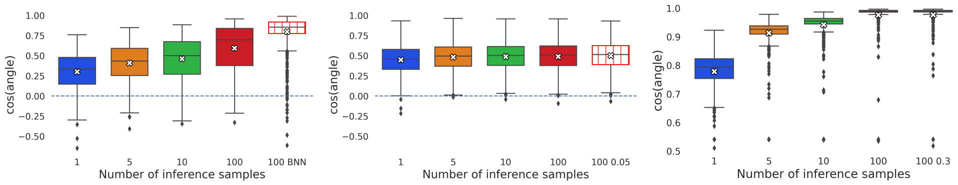


(a) IM (b) BNN

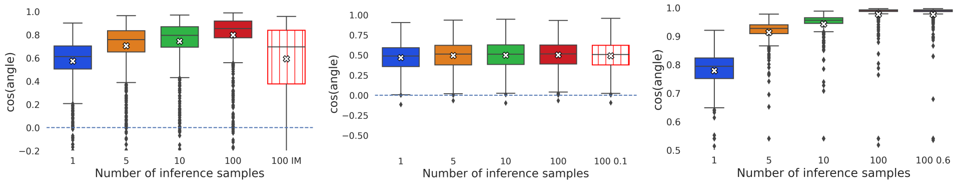
Figure 8: Accuracy under attack for the first 1,000 test set images from FashionMNIST with varying perturbation strengths and different attack objectives. Each attack was calculated based on 100 samples and ℓ_2-constraint for models trained on FashionMNIST.

C.4 Complementary experiments on robustness in dependence of the amount of samples used during inference

In the main part of the paper we argued why, surprisingly, the amount of samples during inference does not influence the robustness, even though we see in figure 9 that less sample lead to the smallest values for $\cos(\alpha_c^{\mathcal{I},\mathcal{A}})$. As stated in the main part, the increased gradient norm when using only few samples seems to compensate the assumed benefits with regard to the angle for using few samples (c.f. figure 10). This can also be seen in table 2 and figure 6 from the main paper, where a (negative) effect on the robustness can only be observed for one inference sample, which also leads to the worst test set accuracy. The benign prediction margins are also hardly effected by the increased number of samples during prediction (c.f. figure 11).

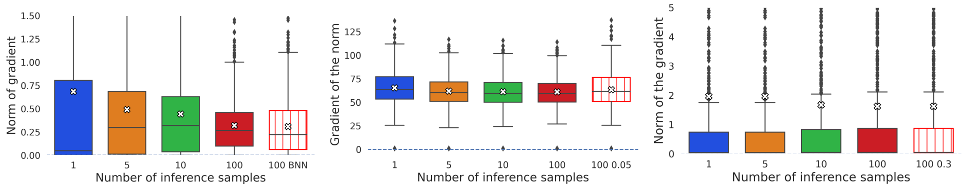


(a) IM (b) SIN 0.1 (c) droprate 0.6

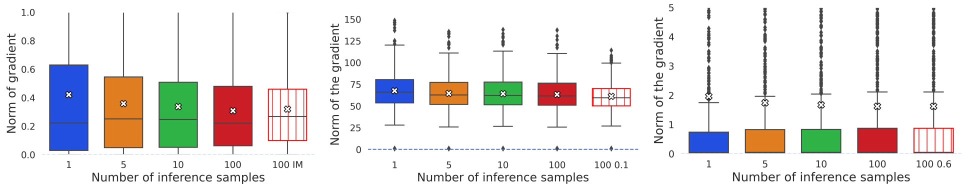


(d) BNN (e) SIN 0.05 (f) droprate 0.3

Figure 9: Cosine of the angle for models trained on FashionMNIST (a), b), d), e)) and trained on CIFAR10 (c) and f)) for different amounts of samples used during inference. Used attack direction δ was calculated based on 100 sample of FGM under ℓ_2-norm constraint with $\eta = 1.5$ and 0.3 respectively.

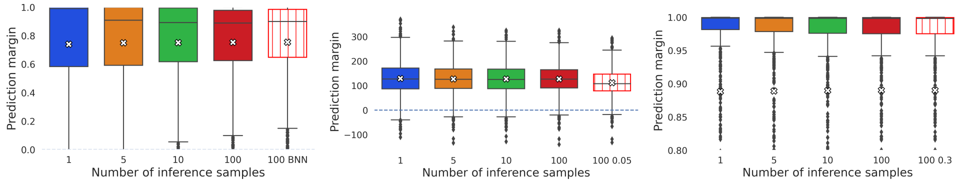


(a) IM (b) SIN 0.1 (c) droprate 0.6

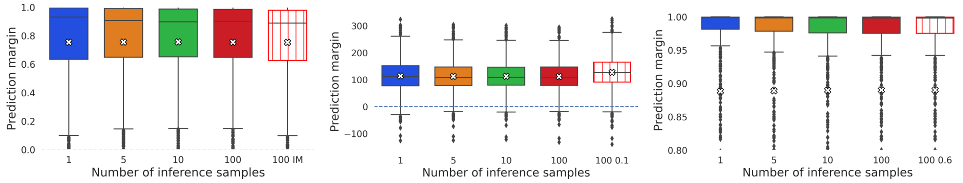


(d) BNN (e) SIN 0.05 (f) droprate 0.3

Figure 10: Norm of the gradient for models trained on FashionMNIST (a), b), d), e)) and trained on CIFAR10 (c) and f)) for different amounts of samples used during inference.



(a) IM (b) SIN 0.1 (c) droprate 0.6



(d) BNN (e) SIN 0.05 (f) droprate 0.3

Figure 11: Prediction margin for the benign inputs for models trained on FashionMNIST (a), b), d), e)) and trained on CIFAR10 (c) and f)) for different amounts of samples used during inference.

C.5 Experiments for CIFAR100

For the experiments on CIFAR100 we used yet another method to create stochastic neural networks, namely by applying a Laplace approximation [MacKay, 1992] to an already trained network. This is archived by adapting a Gaussian distribution over the network's parameters such that the mean is given by the maximum a posterior estimate and the covariance are calculated to match the local loss curvature. Specifically, we used the GitHub library from Daxberger et al. [2021] on top of the adversarial trained wide ResNet70-16 with clean accuracy 69.15 provided by Gowal et al. [2020]. Because of the high amount of parameters in this network we used a last-layer diagonal Gaussian approximation for fitting our posterior distribution from which we sample the θ_i's for deriving an approximate expected prediction. As in the previous experiments we observe, that the adversarial accuracy decreases with more samples during attack, while the angle between the attack and negative gradient during inference decreases which leads to a cosine increase.

Table 1: Adversarial accuracy decrease and $\cos(\alpha_c^{\mathcal{I},\mathcal{A}})$ increase on CIFAR100 with increased amount of samples during attack, where the attack was conducted with ℓ_∞ norm constraint and perturbation strength 8/225.

# SAMPLES	ADVERSARIAL ACCURACY	AVERAGE $\cos(\alpha_c^{\mathcal{I},\mathcal{A}})\pm$ STD
1	48.00	0.2028 ± 0.112
5	45.00	0.2550 ± 0.100
10	43.90	0.2680 ± 0.095

References

Sébastien Bubeck. Convex optimization: Algorithms and complexity. In *arXiv preprint `https://arxiv.org/pdf/1405.4980`*, 2015.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux–effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, (ICLR)*, 2015.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint*, 2020. URL `https://arxiv.org/pdf/2010.03593`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10, cifar-100 (canadian institute for advanced research). In *`http://www.cs.toronto.edu/~kriz/cifar.html`*.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, 1992. ISSN 0899-7667.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv preprint `https://arxiv.org/abs/1708.07747`*, 2017.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. In *International Conference on Learning Representations (ICLR)*, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *arXiv preprint `https://arxiv.org/abs/1605.07146`*, 2017.