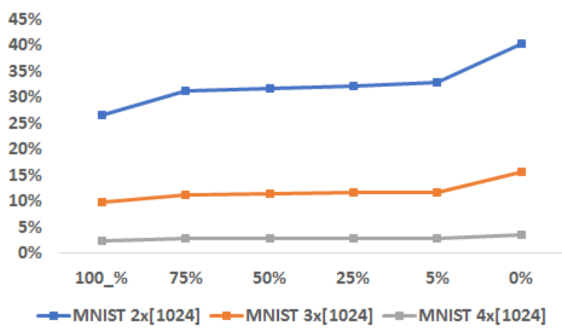
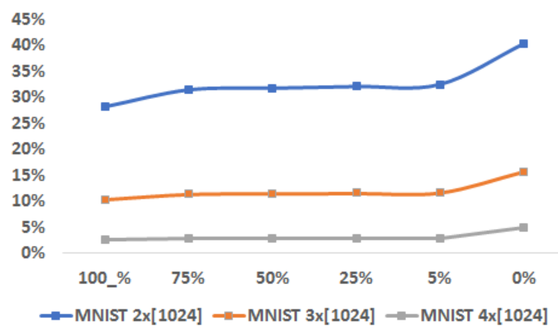
Table 4: **(Full table)** success rates with random attacks using Uniform noises and Bernoulli noises on 100 randomly chosen test images.

Perturbed ℓ_∞ magnitude	$\epsilon = 0.30$		$\epsilon = 0.25$		$\epsilon = 0.20$	
MNIST model	Uniform	Bernoulli	Uniform	Bernoulli	Uniform	Bernoulli
MNIST 2-layer CNN, ReLU	25%	67%	25%	72%	15%	65%
MNIST 2-layer CNN, tanh	35%	59%	91%	99%	83%	98%
MNIST 2-layer CNN, sigmoid	83%	100%	92%	100%	15%	44%
MNIST 2-layer CNN, arctan	18%	58%	7%	44%	22%	22%
MNIST 3-layer CNN, ReLU	72%	89%	69%	90%	53%	99%
MNIST 3-layer CNN, tanh	80%	90%	11%	25%	0%	41%
MNIST 3-layer CNN, sigmoid	7%	31%	14%	24%	30%	76%
MNIST 3-layer CNN, arctan	7%	79%	24%	83%	55%	73%
MNIST 2-layer (robust)-CNN, ReLU	12%	35%	8%	20%	4%	14%
MNIST 2-layer (robust)-CNN, tanh	16%	54%	14%	38%	10%	26%
MNIST 3-layer (robust)-CNN, ReLU	9%	48%	6%	18%	6%	10%
MNIST 3-layer (robust)-CNN, tanh	13%	27%	12%	21%	8%	15%
MNIST LeNet No Pool, ReLU	11%	55%	6%	26%	4%	12%
MNIST ResNet-3, ReLU	98%	98%	98%	98%	98%	100%
Perturbed ℓ_∞ magnitude	$\epsilon = 0.030$		$\epsilon = 0.025$		$\epsilon = 0.020$	
CIFAR model	Uniform	Bernoulli	Uniform	Bernoulli	Uniform	Bernoulli
CIFAR 5×[2048], ReLU	15%	18 %	15%	16%	13%	15%
CIFAR 6×[2048], ReLU	17%	- %	17%	20%	14 %	20 %
CIFAR 5-layer CNN, ReLU	23%	42 %	22 %	31%	17%	28%

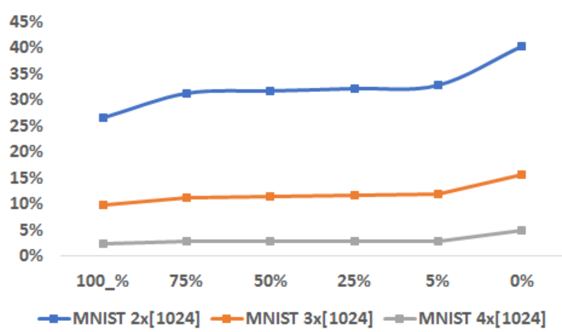
Figure 1: We plot the improvement of the largest ϵ certified by PROVEN with various confidence ($\gamma_L = \{99.99, 75, 50, 25, 5\}\%$) over the largest ϵ certified by worst-case robustness certification algorithms (Weng et al., 2018; Zhang et al., 2018). We consider both input perturbations being independent/correlated Gaussian random variables as in Case (ii) and indedepent random variables as in Case (i). The x-axis label in the figure: γ_L; y-axis label: Certification improvement of PROVEN over $\epsilon_{\text{worst-case}}$. The models are 2-4 layers MNIST networks with 1024 nodes per layer and ReLU actiavations.



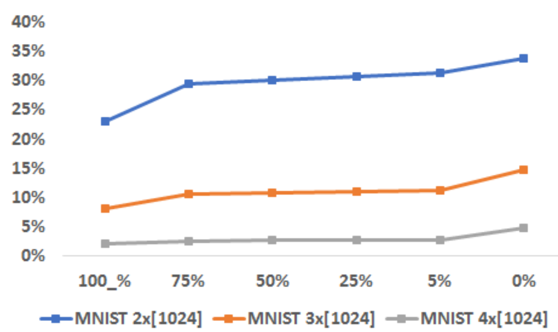
(a) Case (ii) Gaussian i.i.d.



(b) Case (ii) Positive correlated Gaussian



(c) Case (ii) General correlated Gaussian



(d) Case (i) Bounded independent inputs

Table 5: The largest ϵ that PROVEN can certify with confidence of at least $\gamma_L = \{99.99, 75, 50, 25, 5\}\%$ when X_i are independent random variables in Case (i). We compare the largest ϵ that PROVEN can certify with 99.99% with the largest ϵ from state-of-the-art worst-case robustness certification algorithms Fast-Lin (Weng et al., 2018) and show in the last column that PROVEN can certify more than the worst-case analysis by giving up 0.01% confidence. The results comparing PROVEN with CROWN (Zhang et al., 2018) are shown in the main paper in Table 3a with additional 'ada' after network names.

(a) Relu activation

Certification Method	Worst-case (Fast-Lin)	Our probabilistic approach: PROVEN					Certification
Guarantees γ_L	100%†	99.99%†	75%	50%	25%	5%	improvement†
MNIST 2×[20]	0.02722	**0.04394**	0.04782	0.04824	0.04859	0.04897	**61.4%**
MNIST 3×[20]	0.02127	**0.02694**	0.02831	0.02847	0.02860	0.02874	**26.7%**
MNIST 2×[1024]	0.02904	**0.03572**	0.03758	0.03778	0.03796	0.03814	**23.0%**
MNIST 3×[1024]	0.02082	**0.02253**	0.02303	0.02309	0.02313	0.02318	**8.2 %**
MNIST 4×[1024]	0.00796	**0.00813**	0.00817	0.00818	0.00818	0.00818	**2.1 %**
CIFAR 5×[2048]	0.00183	**0.00186**	0.00186	0.00186	0.00186	0.00186	**1.6 %**
CIFAR 7×[1024]	0.00189	**0.00192**	0.00192	0.00193	0.00193	0.00193	**1.6 %**

Table 6: The largest ϵ that PROVEN can certify with confidence of at least $\gamma_L = \{99.99, 75, 50, 25, 5\}\%$ when X_i are independent random variables in Case (i). We compare the largest ϵ that PROVEN can certify with 99.99% with the largest ϵ from state-of-the-art worst-case certification algorithms Fast-Lin and CROWN (Weng et al., 2018; Zhang et al., 2018) and show in the last column that PROVEN can certify more than the worst-case analysis by giving up 0.01% confidence.

(a) Sub-Gaussian noises, bounds

Certification Method	Worst-case	Our probabilistic approach: PROVEN					Certification
Guarantees γ_L	100%†	99.99%†	75%	50%	25%	5%	Improvement†
MNIST 2×[20], ReLU ada	0.02746	**0.04912**	0.05212	0.05246	0.05276	0.05307	**78.9 %**
MNIST 3×[20], ReLU ada	0.02236	**0.03828**	0.03966	0.03981	0.03995	0.04009	**71.2 %**
MNIST 2×[1024], ReLU ada	0.03158	**0.05560**	0.05756	0.05779	0.05798	0.05818	**76.1 %**
MNIST 3×[1024], ReLU ada	0.02397	**0.03524**	0.03583	0.03589	0.03595	0.03601	**47.1 %**
MNIST 4×[1024], ReLU ada	0.00962	**0.01288**	0.01293	0.01294	0.01295	0.01295	**33.9 %**
CIFAR 5×[2048], ReLU ada	0.00228	**0.00264**	0.00265	0.00265	0.00265	0.00265	**15.8 %**
CIFAR 7×[1024], ReLU ada	0.00189	**0.00209**	0.00210	0.00210	0.00210	0.00210	**10.6 %**
MNIST 2×[1024], tanh	0.02232	**0.02915**	0.03005	0.03013	0.03022	0.03033	**30.6%**
MNIST 3×[1024], tanh	0.01121	**0.01360**	0.01376	0.01378	0.01380	0.01381	**21.3 %**
MNIST 4×[1024], tanh	0.00682	**0.00745**	0.00750	0.00750	0.00751	0.00751	**9.2 %**
CIFAR 5×[2048], tanh	0.00081	**0.00085**	0.00085	0.00085	0.00085	0.00085	**4.9 %**
MNIST 2×[1024], sigmoid	0.02785	**0.03285**	0.03404	0.03419	0.03426	0.03441	**18.0%**
MNIST 3×[1024], sigmoid	0.01856	**0.02296**	0.02342	0.02348	0.02353	0.02358	**23.7 %**
MNIST 4×[1024], sigmoid	0.01778	**0.02170**	0.02224	0.02229	0.02232	0.02237	**22.1 %**
MNIST 2×[1024], arctan	0.02105	**0.02796**	0.02907	0.02915	0.02924	0.02936	**32.8%**
MNIST 3×[1024], arctan	0.01250	**0.01462**	0.01486	0.01488	0.01490	0.01493	**17.0 %**
MNIST 4×[1024], arctan	0.00726	**0.00829**	0.00836	0.00837	0.00838	0.00838	**14.2 %**
MNIST 2-layer CNN, ReLU	0.04565	**0.06367**	0.06884	0.06989	0.07082	0.07181	**1.4X**
MNIST 2-layer CNN, tanh	0.0331	**0.09987**	0.13538	0.1437	0.15135	0.15981	**3.0X**
MNIST 2-layer CNN, sigmoid	0.09242	**0.18777**	0.2218	0.22906	0.23553	0.24243	**2.0X**
MNIST 2-layer CNN, arctan	0.03747	**0.13114**	0.18872	0.20279	0.21577	0.23028	**3.5X**
MNIST 3-layer CNN, ReLU	0.04609	**0.06301**	0.0674	0.06828	0.06904	0.06986	**1.4X**
MNIST 3-layer CNN, tanh	0.03348	**0.05917**	0.06676	0.06828	0.06962	0.07108	**1.8X**
MNIST 3-layer CNN, sigmoid	0.07477	**0.13204**	0.14844	0.15186	0.15471	0.15781	**1.8X**
MNIST 3-layer CNN, arctan	0.02868	**0.05514**	0.06272	0.06425	0.06559	0.06702	**1.9X**
MNIST ResNet-3, ReLU	0.01751	**0.01827**	0.01864	0.01869	0.01876	0.01881	**1.0X**
CIFAR 5-layer CNN, ReLU	0.00402	**0.00465**	0.00471	0.00472	0.00473	0.00473	**1.2X**
TinyImagenet, 7-layer CNN, ReLU	0.07245	**0.07367**	0.07367	0.07368	0.07369	0.0737	**1.0X**
MNIST 2-layer (robust)-CNN, ReLU	0.09304	**0.11424**	0.12224	0.1238	0.12515	0.12658	**1.2X**
MNIST 2-layer (robust)-CNN, tanh	0.12795	**0.37451**	0.76167	0.90881	1.06778	1.2689	**2.9X**
MNIST 3-layer (robust)-CNN, ReLU	0.10494	**0.11984**	0.1253	0.12631	0.12717	0.12809	**1.1X**
MNIST 3-layer (robust)-CNN, tanh	0.20596	**0.24122**	0.27452	0.28091	0.28649	0.29239	**1.2X**

Table 7: **Subgaussian noises**: With input perturbations being independent random variables in case (i), we randomly choose $\{10, 50, 100\}$ input samples (images) in each trial and then compute the average of the largest ϵ that can be certified by worst-case framework CNN-Cert (Boopathy et al., 2019) (denoted as $\epsilon_{\text{worst-case}}$) and by PROVEN with 99.99% confidence (denoted as ϵ_{PROVEN}) together with the improved certification of ϵ_{PROVEN} over $\epsilon_{\text{worst-case}}$ (denoted as Improv.). We present the mean and std of the average ϵ and the improvements for $\{10, 50, 100\}$ samples in a total of 100 random trials, showing that the mean and std converge as the number of samples increases.

Models	bound	10 samples			50 samples			100 samples		
		$\epsilon_{\text{worst-case}}$	ϵ_{PROVEN}	Improv.	$\epsilon_{\text{worst-case}}$	ϵ_{PROVEN}	Improv.	$\epsilon_{\text{worst-case}}$	ϵ_{PROVEN}	Improv.
MNIST 3×[1024], ReLU,ada	Mean	0.02559	0.03703	44.75%	0.02581	0.03734	44.70%	0.02579	0.03733	44.74%
	std	0.00165	0.00222	1.12%	0.00076	0.00102	0.57%	0.00054	0.00071	0.43%
MNIST 3×[1024], tanh	Mean	0.01195	0.01375	15.17%	0.01193	0.01374	15.22%	0.01192	0.01374	15.25%
	std	0.00065	0.00068	2.66%	0.00030	0.00030	1.27%	0.00020	0.00021	0.77%
MNIST 4×[1024], ReLU,ada	Mean	0.00998	0.01329	33.18%	0.00994	0.01325	33.24%	0.00997	0.01328	33.21%
	std	0.00051	0.00066	0.57%	0.00021	0.00027	0.27%	0.00014	0.00018	0.15%
CIFAR 5×[2048], ReLU,ada	Mean	0.00224	0.00264	18.07%	0.00222	0.00262	17.93%	0.00222	0.00263	18.06%
	std	0.00020	0.00025	2.39%	0.00009	0.00011	1.12%	0.00005	0.00006	0.55%
CIFAR 5×[2048], arctan	Mean	0.00091	0.00100	9.28%	0.00091	0.00100	9.32%	0.00092	0.00100	9.32%
	std	0.00008	0.00009	3.17%	0.00003	0.00003	1.15%	0.00001	0.00002	0.56%
CIFAR 7×[1024], ReLU,ada	Mean	0.00176	0.00195	10.68%	0.00174	0.00192	10.73%	0.00174	0.00193	10.70%
	std	0.00018	0.00020	1.87%	0.00007	0.00008	0.75%	0.00003	0.00004	0.37%

Table 8: **Gaussian correlated noises**: compare PROVEN with worst-case certification CNN-Cert (Boopathy et al., 2019)

Certification Method	Worst-case	Our probabilistic approach: PROVEN					Certification
Guarantees γ_L	100%†	99.99%†	75%	50%	25%	5%	Improvement†
MNIST 2-layer CNN, ReLU	0.04565	**0.06975**	0.07203	0.07256	0.0731	0.07388	**1.5X**
MNIST 2-layer CNN, tanh	0.0331	**0.14265**	0.1617	0.16626	0.17091	0.17782	**4.3X**
MNIST 2-layer CNN, sigmoid	0.09242	**0.22809**	0.24401	0.24769	0.25141	0.25684	**2.5X**
MNIST 2-layer CNN, arctan	0.03747	**0.20091**	0.23355	0.24136	0.24946	0.2616	**5.4X**
MNIST 3-layer CNN, ReLU	0.04609	**0.06816**	0.07004	0.07046	0.07089	0.07152	**1.5X**
MNIST 3-layer CNN, tanh	0.03348	**0.06809**	0.0714	0.07216	0.07293	0.07405	**2.0X**
MNIST 3-layer CNN, sigmoid	0.07477	**0.15139**	0.15852	0.16012	0.16171	0.16403	**2.0X**
MNIST 3-layer CNN, arctan	0.02868	**0.06406**	0.06734	0.06811	0.06888	0.07	**2.2X**
MNIST ResNet-3, ReLU	0.01751	**0.01868**	0.01883	0.01884	0.01887	0.0189	**1.1X**
CIFAR 5-layer CNN, ReLU	0.00402	**0.00465**	0.00471	0.00472	0.00473	0.00473	**1.2X**
Tiny Imagenet, 7-layer CNN, ReLU	0.07245	**0.07368**	0.0737	0.07371	0.07372	0.07372	**1.0X**
MNIST 2-layer (robust)-CNN, ReLU	0.09304	**0.12361**	0.1269	0.12764	0.12838	0.12946	**1.3X**
MNIST 2-layer (robust)-CNN, tanh	0.12795	**0.88968**	1.3172	1.43648	1.56153	1.74872	**7.0X**
MNIST 3-layer (robust)-CNN, ReLU	0.10494	**0.12618**	0.12829	0.12875	0.12921	0.12989	**1.2X**
MNIST 3-layer (robust)-CNN, tanh	0.20596	**0.28015**	0.29364	0.29681	0.29994	0.30452	**1.4X**

Table 9: **Gaussian iid noises**: compare PROVEN with worst-case certification CNN-Cert (Boopathy et al., 2019)

Certification Method	Worst-case	Our probabilistic approach: PROVEN					Certification
Guarantees γ_L	100%†	99.99%†	75%	50%	25%	5%	Improvement†
MNIST 2-layer CNN, ReLU	0.04565	**0.06975**	0.07204	0.07256	0.0731	0.07388	**1.5X**
MNIST 2-layer CNN, tanh	0.0331	**0.14261**	0.16169	0.16626	0.1709	0.17781	**4.3X**
MNIST 2-layer CNN, sigmoid	0.09242	**0.22811**	0.24399	0.24769	0.25141	0.25682	**2.5X**
MNIST 2-layer CNN, arctan	0.03747	**0.20094**	0.23356	0.24136	0.24949	0.26153	**5.4X**
MNIST 3-layer CNN, ReLU	0.04609	**0.06816**	0.07004	0.07046	0.07089	0.07152	**1.5X**
MNIST 3-layer CNN, tanh	0.03348	**0.06808**	0.07139	0.07216	0.07292	0.07405	**2.0X**
MNIST 3-layer CNN, sigmoid	0.07477	**0.1514**	0.15852	0.16012	0.16171	0.16403	**2.0X**
MNIST 3-layer CNN, arctan	0.02868	**0.06405**	0.06734	0.06811	0.06888	0.07001	**2.2X**
MNIST ResNet-3, ReLU	0.01751	**0.01868**	0.01883	0.01884	0.01887	0.0189	**1.1X**
TinyImageNet, 7-layer CNN, ReLU	0.07245	**0.07368**	0.0737	0.07371	0.07372	0.07372	**1.0X**
MNIST 2-layer (robust)-CNN, ReLU	0.09304	**0.1236**	0.1269	0.12764	0.12838	0.12946	**1.3X**
MNIST 2-layer (robust)-CNN, tanh	0.12795	**0.88787**	1.31724	1.43648	1.56164	1.74802	**6.9X**
MNIST 3-layer (robust)-CNN, ReLU	0.10494	**0.12618**	0.12829	0.12875	0.12921	0.12989	**1.2X**
MNIST 3-layer (robust)-CNN, tanh	0.20596	**0.28014**	0.29365	0.29681	0.29995	0.30454	**1.4X**