ROBUST ENSEMBLES OF NEURAL NETWORKS USING ITÔ PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual neural networks (ResNets) can be modeled as dynamical systems where the evolution of dynamical systems represents the inference in ResNets. We exploit this connection and the theory of stochastic dynamical systems to construct a novel ensemble of Itô processes as a new deep learning representation that is more robust than classical residual networks. An Itô process obtained by solving a suitably-formulated stochastic differential equation derived from a residual network has a probability density function that is not readily perturbed by small changes in the neural network's inputs. Our robust stochastic Itô ensemble of neural networks achieve an accuracy of 73.91% on the CIFAR-10 dataset against the PGD attack with $\epsilon = 2.0$ under the L_2 norm, while the accuracy of Madry's robustness toolbox on the same attack is 18.59%. Similarly, our stochastic Itô ensemble of neural networks achieves an accuracy of 79.66% on PGD attack with $\epsilon = 16/255$ under the L_∞ norm, while the accuracy of Madry's robustness toolbox on the same attack is 18.13%. The Itô ensemble trained on ImageNet achieves an accuracy of 28.53% against PGD attacks under the L_∞ norm with $\epsilon = 16/255$ and accuracy of 65.74% under the L_2 norm with $\epsilon = 3.0$, respectively. This significantly improves state-of-the-art accuracy of 5% and 35.16% for Madry's robustness tool against the same PGD attacks under the L_∞ and L_2 norms, respectively. Further, our approach achieves these high robustness values without any explicit adversarial training or a significant loss of accuracy on benign inputs.

1 INTRODUCTION

Deep neural networks (DNNs) have emerged as a very effective learning representation achieving near human-level performance in many domains such as computer vision (Gkioxari et al., 2015), natural language processing (Majumder et al., 2017), and speech recognition (Hannun et al., 2014). Despite this success, the use of deep learning models in high-assurance systems with safety and security requirements such as autonomous vehicles (Bojarski et al., 2016) and medical diagnoses (De Fauw et al., 2018) faces a trust deficit. The lack of robustness of these models and their susceptibility to adversarial attacks (Kurakin et al., 2016; Szegedy et al., 2013) that can change the prediction of a deep neural network via small imperceptible perturbations make deep learning models less trustworthy. This limitation is further aggravated by deep neural networks generally exhibiting very high confidence on incorrect predictions (Guo et al., 2017a; Hendrycks & Gimpel, 2016). Consequently, this lack of robustness hinders their deployment in safety-critical applications. There is a pressing need for a principled approach to learning robust deep learning models that are resilient to adversarial attacks and can abstain from making decisions on inputs for which they are likely to make a wrong prediction.

A number of approaches have been recently proposed to increase the robustness of deep learning models. Adversarial training (Tramèr et al., 2017; Engstrom et al., 2020) uses adversarial samples in the training phase to make the models more robust. Another set of alternative approaches use the projection of inputs to data manifold (Lamb et al., 2018; Ilyas et al., 2017; Jang et al., 2020) or other preprocessing methods (Xie et al., 2019; Guo et al., 2017b). These approaches are robust to existing attack methods but their use of adversarial samples or predefined transformations (often achieved via another deep neural network such as autonecoders) makes these approaches susceptible to newer attack strategies. Certifiable-defense approaches (Wong et al., 2018; Dvijotham et al., 2018;

Raghunathan et al., 2018; Dutta et al., 2018) have also been recently proposed to make deep learning models robust against worst-case input over a defined range of perturbations. These theoretical guarantees on worst-case inputs hold only for small perturbations; consequently, their use is limited in practice and their performance is typically inferior to approaches based on adversarial training, particularly for high-dimensional inputs.

In this paper, we address this challenge of robust and trustworthy deep learning using a new representation that exploits the connection between dynamical systems and residual neural networks (ResNets), and uses the theory of stochastic dynamical systems. Dynamical systems can model ResNets where the inference in the network is represented by the evolution of the dynamical system (Chen et al., 2015; Chang et al., 2017; Sonoda & Murata, 2017; Chen et al., 2018; Lu et al., 2018). We construct a novel deep ensemble using a special class of stochastic dynamical systems, namely the Itô drift-diffusion process with suitably bounded diffusion term. Itô process is the sum of the integral of a process over time and of another process over a Brownian motion. The drift over time models the typical inference in a ResNet and the diffusion Brownian motion models the added stochastic noise that makes the model robust to adversarial perturbations. We form an ensemble of these Ito processes by considering multiple such models and multiple inferences over the same model. If a majority of the ensemble agrees on a particular prediction, Itô ensemble makes that prediction; otherwise, it abstains from making a decision.

Robustness Approach	Accuracy (%)
Itô Ensemble	**84.60**
Engstrom et al. (2020)	53.49
Balunovic & Vechev (2020)	46.2
Zhang et al. (2019)	40.5
Pang et al. (2019) (ϵ=0.01)	48.4

Benchmark	Norm	Accuracy (%)	
		Itô Ensemble	Madry toolbox
CIFAR-10	L_2	**73.91**	18.59
ImageNet	L_2	**69.51**	43.04
CIFAR-10	L_∞	**79.66**	18.13
ImageNet	L_∞	**28.53**	5.00

Table 1: (**left**) Our Itô ensemble approach outperforms SOTA defenses for the PGD attack on CIFAR-10 with $\epsilon = 8/255$ unless specified otherwise. (**right**) Itô ensemble outperforms Madry toolbox (Engstrom et al., 2020) under PGD attack with L_2 norm, $\epsilon = 2.0$, and L_∞ norm $\epsilon = 16/255$.

We highlight a few results demonstrating the robustness of Itô process ensembles in Table 1. Our Itô ensemble approach has higher accuracy compared to several state-of-the-art robustness approaches. The accuracy of our approach is 84.60% on CIFAR-10 against the PGD attack in L_∞ norm with $\epsilon = 8/255$ and the next best approach is Engstrom et al. (2020) (Madry toolbox) with an accuracy of 53.49%. On CIFAR-10 and ImageNet benchmarks, our Itô ensemble approach is significantly more robust than Engstrom et al. (2020) (Madry toolbox) against PGD attacks in both L_2 and L_∞ norms for different values of attack strength ϵ. Thus, our Itô ensembles exhibit remarkable robustness against adversarial attacks without any explicit adversarial training.

Figure 1: Examples of benign images on which our approach using Itô processes abstains from making a decision while the original ResNet model makes a decision despite high uncertainty. The first image is found by Itô process ensemble to be confusing between *binoculars* and *cannon*, the second between a *radiator* and a *projector*, the third between a *trench-coat* and *bicycle*, and the last one between *stove* and *coffee-pot*. This uncertainty in Itô ensemble resembles human judgement.

Our approach using Itô ensembles can abstain from making decisions on a confusing input. In Section 4, we demonstrate that abstentions further improve the robustness of the Itô ensembles to adversarial examples compared to the state-of-the-art approaches. Further, we notice that Itô ensembles abstain even on benign data inputs where manual inspection demonstrates high aleatoric or epistemic uncertainty as shown in Figure 1. Our experiments show that this new approach of using Itô ensembles achieves high robustness without significant loss in accuracy on benign inputs.

2 RELATED WORK

Residual neural networks are a common neural network architecture that learn only the residuals not learned by the previous layers. ResNets (He et al., 2016) are residual neural networks where residual learning is adopted for every few stacked neural network layers and such building blocks are used to design the complete residual neural network. The dynamics of ResNets and other similar neural networks can be described using ordinary and partial differential equations (Chen et al., 2015; Chang et al., 2017; Sonoda & Murata, 2017; Weinan, 2017; Chen et al., 2018; Lu et al., 2018). One time-step of the dynamics models each building block of the ResNets. Such a dynamical model enables memory efficiency in training and adaptive inference. In contrast, we use stochastic differential equations (Itô processes) and demonstrate their robustness to adversarial attacks.

A number of adversarial attacks on deep neural networks have been proposed in literature and shown to be effective across different architectures. Attacks such as the fast gradient sign method (FGSM) (Szegedy et al., 2013), the projected gradient decent (PGD) (Madry et al., 2017) and other approaches (Nicolae et al., 2018) have demonstrated the fragility of deep neural networks to small perturbations in their inputs. The most effective state-of-art defenses use adversarial training (Tramèr et al., 2017; Engstrom et al., 2020) or some projection or transformation of inputs (Lamb et al., 2018; Ilyas et al., 2017; Jang et al., 2020; Xie et al., 2019; Guo et al., 2017b). The use of adversarial examples or predefined transformations makes these approaches vulnerable to new attacks. In contrast, our approach using Itô process does not need adversarial examples.

Our use of stochastic dynamical systems is inspired by their presence in biological systems (Kitano, 2004; Bressloff, 2014; Allen, 2010) where they impart robustness to external perturbations. As an example, (Arkin et al., 1998) study gene expression using Gillespie's stochastic formulation of chemical kinetics and show that protein numbers can vary markedly from one cell to another with important consequences for biological robustness (Gonze et al., 2002).

3 ROBUST LEARNING USING ITÔ ENSEMBLES

Residual networks (ResNets) can be modeled as dynamical systems where the evolution of the dynamical system represents the inference in ResNets (Chen et al., 2015; Chang et al., 2017; Sonoda & Murata, 2017; Chen et al., 2018; Lu et al., 2018). We connect this view to the theory of stochastic differential equations and construct an ensemble using a class of Itô processes with suitably bounded diffusion term. This Itô process ensemble exhibits remarkable robustness against adversarial attacks without any explicit adversarial training.

FROM RESNETS TO STOCHASTIC ITÔ PROCESSES

A building block of a residual neural network (He et al., 2016) with the residual mapping $\mathcal{F}(\mathbf{x}(i), \mathbf{W}(i))$ can be described using the following equation: $\mathbf{x}(i+1) = \mathcal{F}(\mathbf{x}(i), \mathbf{W}(i)) + \mathbf{x}(i)$. Here, $\mathbf{x}(i)$ is the input to the i^{th} residual network building block and $\mathbf{x}(i+1)$ is the corresponding output that serves as an input to the next building block. The weights of the neural network layers in this ResNet building block are denoted by $\mathbf{W}(i)$.

After taking suitable limits, the evolution of the ResNet can be described by the ResNet ordinary differential equation (ODE): $\frac{d\mathbf{x}(t)}{dt} = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$. Here, $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t)) = \lim_{\delta t \to 0} \frac{\mathcal{F}(\mathbf{x}(t), \mathbf{W}(t))}{\delta t}$ and $\mathbf{x}(0)$ is the input to the neural network. The ResNet ODE can be naturally generalized into an Itô process by using a Brownian motion term with diffusion coefficient $\Sigma(t) = (\sigma_{ij}(t))$: $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))\, dt + \Sigma(t)\, dB(t)$. There are two competing objectives here:

- Very large values of the diffusion term $\Sigma(t)$ can completely overshadow the drift term $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ leading to a poor accuracy even on benign inputs. When the diffusion term is very large, the paths of the Itô process can completely diverge from the solution of the original ResNet from which the Itô process was obtained.
- Very small values of $\Sigma(t)$ make the model closer to the original ResNet and equally non-robust. $\Sigma(t) = 0$ reproduces the original non-stochastic ResNet with no additional robustness. As we increase the diffusion term, the robustness of the neural network increases; this is experimentally demonstrated in Section 4.

So, a natural question to ask is: *How do we select the diffusion term $\Sigma(t)$ such that the Itô process satisfies these two competing objectives of accuracy and robustness?*

At one hand, the generated Itô process must retain similar accuracy on benign models as the original ResNet, that is, its solutions are determined mainly by the term $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ and Brownian motion noise does not make it diverge significantly. On the other hand, the choice of added diffusion term $\Sigma(t)$ must make the model robust enough to be resilient to adversarial perturbations on the inputs.

Robustness of Stochastic Itô ResNet Ensembles

Given the Itô process $\mathbf{x}(t)$ satisfying the stochastic differential equation $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))\ dt + \Sigma(t)\ dB(t)$, it is known (Oksendal, 1992) that the probability density $p(\mathbf{x}, t)$ can be mathematically characterized by the following equation:

$$\frac{\partial p(\mathbf{x},t)}{\partial t} + \mathcal{G}(\mathbf{x}, \mathbf{W}) \nabla p(\mathbf{x}, t) = -p(\mathbf{x}, t) \sum_i \frac{\partial \mathcal{G}}{\partial \mathbf{x}_i} + \frac{1}{2} \sum_i \sum_j \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j} \left((\sum_k \sigma_{ik}(t) \sigma_{jk}(t)) p(\mathbf{x}, t) \right)$$

For robust networks, the rate of change of the residual learning map is much smaller than the residual map itself, that is $\sum_i \frac{\partial \mathcal{G}}{\partial \mathbf{x}_i} < \eta_1\ \mathcal{G}(\mathbf{x}, \mathbf{W}) \frac{\nabla p(\mathbf{x},t)}{p(\mathbf{x},t)}$ for some small $\eta_1 \to 0$. Hence, the probability density $p(\mathbf{x}, t)$ can be simplified to the following equation:

$$\frac{\partial p(\mathbf{x},t)}{\partial t} + (1 + \eta_1) \mathcal{G}(\mathbf{x}, \mathbf{W}) \nabla p(\mathbf{x}, t) = \frac{1}{2} \sum_i \sum_j \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j} \left(\left(\sum_k \sigma_{ik}(t) \sigma_{jk}(t) \right) p(\mathbf{x}, t) \right)$$

Using the fact that the double derivative of the probability density for robust neural networks is much smaller than the residual map itself, that is $\frac{1}{2} \sum_i \sum_j \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j} (p(\mathbf{x}, t)) < \eta_2\ \mathcal{G}(\mathbf{x}, \mathbf{W}) \frac{\nabla p(\mathbf{x},t)}{p(\mathbf{x},t)}$ for some small $\eta_2 \to 0$, we choose $\sigma_{ij}(t) \leq \frac{\omega}{1+t}$ for a constant ω. Then, the equation for the probability density function of a ResNet with n-dimensional inputs can be further simplified as

$$\frac{\partial p(\mathbf{x},t)}{\partial t} + (1 + \eta_1 - n\omega^2 \eta_2) \mathcal{G}(\mathbf{x}, \mathbf{W}) \nabla p(\mathbf{x}, t) = 0$$

As $\eta_1 \to 0$ and $\eta_2 \to 0$ for robust neural networks, our choice of $\sigma_{ij}(t) \leq \frac{\omega}{1+t}$ for a constant ω reduces the equation describing the probability density function to $\frac{\partial p(\mathbf{x},t)}{\partial t} + \mathcal{G}(\mathbf{x}, \mathbf{W}) \nabla p(\mathbf{x}, t) = 0$. Interpreting $p(\mathbf{x}, t)$ as a function that is constant along the trajectories of a ordinary differential equation i.e. $\frac{dp(\mathbf{x},t)}{dt} = 0$, $p(\mathbf{x}, t)$ corresponds to the following differential equation: $\frac{d\mathbf{x}(t)}{dt} = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$. Hence, under our choice of $\sigma_{ij}(t) \leq \frac{\omega}{1+t}$ for a constant ω, the solution to the stochastic differential equation agrees with the ResNet ODE for robust neural networks. Our implementation of the stochastic robust Itô ensemble of residual neural network is formed by discretizing the stochastic differential equation $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))\ dt + \Sigma(t)\ dB(t)$ with the constraint that $\sigma_{ij}(t) = \frac{\omega}{1+t}$.

4 Results

We train stochastic Itô ensembles of residual neural networks using both CIFAR-10 (Krizhevsky et al., 2014) and ImageNet (Deng et al., 2009) benchmarks. We evaluate the robustness of our stochastic Itô ensembles against two popular adversarial attacks: the fast gradient sign method (FGSM) (Szegedy et al., 2013) and the projected gradient descent (PGD) (Kurakin et al., 2016) under both L_2 and L_∞ norms. We use the conformance in prediction of our Itô ensemble to exploit their robustness. If a majority of residual network models in our ensemble predict the same label for a given data item, the ensemble makes a prediction as this majority label. Otherwise, the stochastic Itô ensemble assigns no label and abstains from making any decision on the given input data. Our experiments indicate that this capability of the Itô ensemble to abstain from making decisions on non-conforming inputs not only helps defend the model against adversarial attacks, it also decreases incorrect predictions on benign data.

CIFAR-10 RESULTS

Our experiments are performed on a 40-core 256GB RAM server with 4 NVIDIA V100 GPUs.

Question 1: Does the Itô ensemble achieve competitive accuracy on benign inputs?

We train the standard ResNet models on benign data and compare their accuracy with the accuracy of our stochastic Itô ensembles on benign data. We obtain the stochastic ResNet models in the stochastic Itô ensemble by starting with the weights of a standard ResNet model and training them for 40 epochs with a learning rate of 0.0001 using the Adam optimizer. Table 2 compares the accuracy of the standard model with Itô ensembles.

Architecture	Accuracy (%)		Incorrect Prediction (%)		Correct + Abstention (%)
	Original	Itô Ensemble	Original	Itô Ensemble	Itô Ensemble (%)
ResNet-18	93.33	91.51	6.67	5.72	94.28
ResNet-34	92.92	91.33	7.08	5.80	94.20
ResNet-50	93.86	91.59	6.14	4.64	95.29

Table 2: Our stochastic Itô ensembles and the standard ResNet neural network architectures have similar accuracy on CIFAR-10 test data. Because of its ability to abstain from assigning a label when majority of predictions do not conform, the fraction of data where the Itô ensemble predicts an incorrect label is lower that the fraction of data where the original ResNet model is incorrect.

Our stochastic Itô ensembles of 20 models in Table 2 are trained using a diffusion term corresponding to $\omega = 0.2$. 20 inferences are obtained from each stochastic model in our Itô ensemble. The accuracy of the stochastic Itô ensembles on CIFAR-10 test data is comparable to that of the standard models on three ResNet architectures: ResNet-18, ResNet-34, and ResNet-50. Our stochastic Itô ensemble abstains when a majority of the ensemble models do not agree on a single prediction. This lowers the incorrect predictions of Itô ensemble compared to the original model. As shown in Figure 1, some of the correct predictions by original ResNet are on images with high aleatoric uncertainty on which Itô ensemble correctly abstains.

Question 2: Is the stochastic Itô ensemble robust against adversarial attacks?

We train stochastic Itô ensembles of 20 ResNet-50 models on CIFAR-10 data with diffusion terms corresponding to $\omega = 0.2$ and $\omega = 0.4$. 20 independent inferences are drawn from each stochastic model in the Itô ensemble. We evaluate the robustness of our stochastic Itô ensemble against FGSM and PGD under both L_2 and L_∞ norms. We compare the accuracy of predictions from our stochastic Itô ensembles with that of Madry's robustness toolbox (Engstrom et al., 2020).

ϵ	Accuracy for PGD (%)	Correct + Abstention for PGD (%)	Accuracy for FGSM (%)	Correct + Abstention for FGSM (%)
0.2	91.34	94.83	91.41	94.93
0.5	90.37	94.53	90.78	94.87
1.0	86.39	91.41	89.22	93.82
2.0	73.91	79.80	83.31	89.41

Table 3: The accuracy of our stochastic Itô ensemble with diffusion term corresponding to $\omega = 0.2$ on the PGD attack for different values of ϵ under the L_2 norm.

Robustness under the L_2 norm. The accuracy of our stochastic Itô ensembles on the fast gradient sign method (FGSM) (Szegedy et al., 2013) and the projected gradient descent (PGD) (Kurakin et al., 2016) under the L_2 norm is shown in Table 3. Our stochastic Itô ensembles use the ResNet-50 architecture with the diffusion term corresponding to $\omega = 0.2$.

Our stochastic Itô ensemble approach with a diffusion terms corresponding to $\omega = 0.2$ shows an accuracy of 73.91% against the PGD attack with $\epsilon = 2.0$ under the L_2 norm. This compare favorably with the 18.59% accuracy of Madry's robustness toolbox on the same PGD attack. The accuracy of our stochastic Itô ensemble approach improves to 79.07% when the diffusion term corresponds

to $\omega = 0.4$. Further, the sum of correct labels and abstentions from our Itô ensemble approach is 88.43% against the PGD attack with $\epsilon = 2.0$ under the L_2 norm.

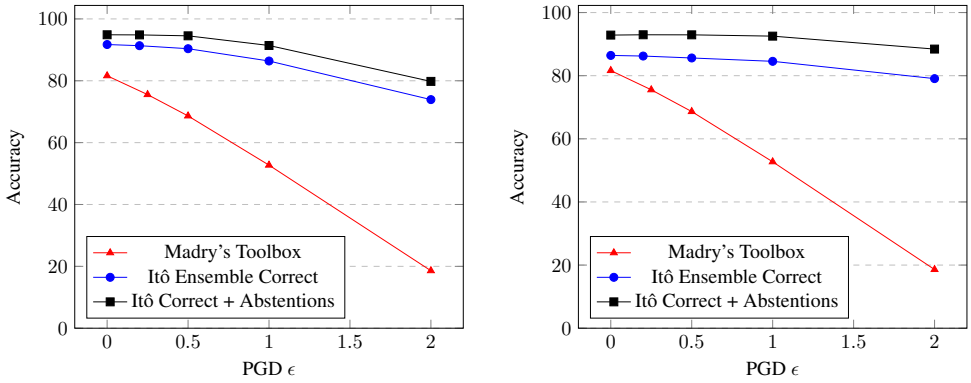

Figure 2: The accuracy of our stochastic Itô ensemble with a diffusion term corresponding to $\omega = 0.2$ (left) and $\omega = 0.4$ (right) on CIFAR-10 compares favorably with Madry's Robustness Toolbox using the L_2 norm for different values of ϵ.

Figure 2 shows the accuracy of our Itô ensemble approach and compares it with the accuracy of Madry's robustness toolbox (Engstrom et al., 2020) on CIFAR-10 test data. Both our stochastic Itô ensembles with $\omega = 0.2$ and $\omega = 0.4$ have higher accuracy on benign data and their accuracy remains higher than Madry's robustness toolbox for all values of ϵ under the L_2 norm. The accuracy of the Itô ensemble degrades more gracefully as the value of ϵ increases under the L_2 norm.

Robustness under the L_∞ norm. We investigate the accuracy of our stochastic Itô ensemble approach under the L_∞ norm and compare it to the accuracy of Madry's robustness toolbox. Table 4 shows the accuracy of our Itô ensemble with diffusion term corresponding to $\omega = 0.2$.

ϵ	Accuracy for PGD (%)	Correct + Abstention for PGD (%)	Accuracy for FGSM (%)	Correct + Abstention for FGSM (%)
$\frac{4}{255}$	85.82	92.91	86.02	93.00
$\frac{8}{255}$	84.60	92.48	85.24	92.84
$\frac{16}{255}$	79.66	89.06	82.74	91.20
$\frac{32}{255}$	62.97	74.05	72.49	84.08

Table 4: The accuracy of our stochastic Itô ensemble with diffusion term corresponding to $\omega = 0.2$ on the PGD attack for different values of ϵ under the L_∞ norm.

We study the performance of our stochastic Itô ensemble on PGD and FGSM and attacks of varying magnitudes under the L_∞ norm, and determine that our stochastic Itô ensemble is robust against adversarial noise. Figure 3 shows the accuracy of Madry's robustness toolbox and our Itô ensemble on adversarial images under the PGD attack. The accuracy of our stochastic Itô ensemble with diffusion term corresponding to $\omega = 0.4$ is higher than that of the model from Madry's toolbox for both the original unperturbed images and PGD adversarial images with $\epsilon = 8/255$ and $\epsilon = 16/255$.

Question 3: How does the robustness of the stochastic Itô ensemble approach change with the number of models in the ensemble and the number of inferences?

The Itô ensemble has two sources of diversity - different stochastic models and multiple inferences on the same model. We investigate the accuracy of our Itô ensemble with different number of stochastic models and different number of inferences from each stochastic model. Figure 4 (left) illustrates the results of our investigations. A significant increase of 5.1% is observed for the sum of correct outcomes and abstentions by increasing the number of stochastic models from 3 to 20 and the number of sampled independent inferences for each stochastic model from 3 to 20. Increasing the number of stochastic models in the ensemble has more significant influence on the performance of the Itô ensemble than increasing the number of independent inferences from each stochastic model.

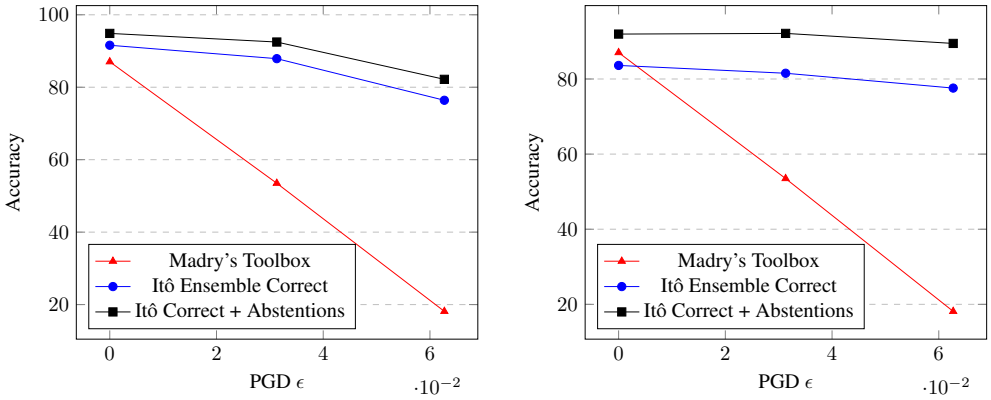

Figure 3: The accuracy of our stochastic Itô ensemble on CIFAR-10 compares favorably with Madry's Robustness Toolbox using the L_∞ norm. (left) $\omega = 0.2$ (right) $\omega = 0.4$.

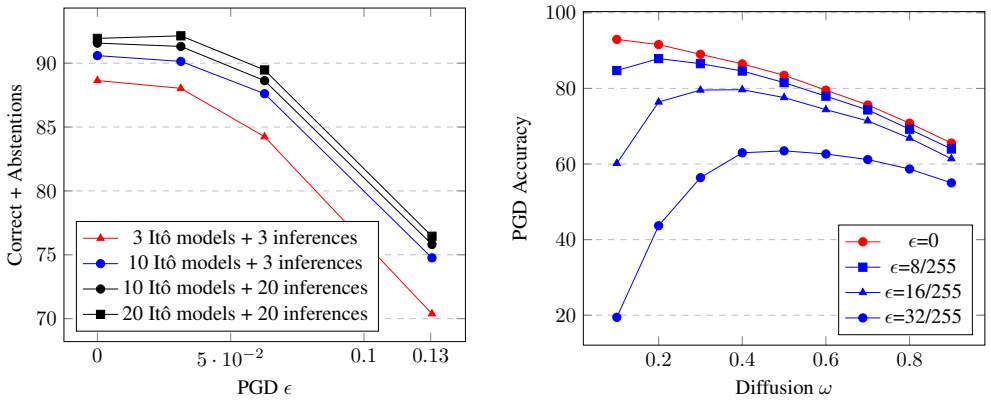

Figure 4: (left) The impact of the number of models and inferences on the sum of correct outcomes and abstentions for our stochastic Itô ensemble with $\omega = 0.4$. (right) Diffusion with different values of ω in stochastic Itô ensemble vs. PGD accuracy under the L_∞ norm.

Question 4: How can we control the diffusion term ω to trade-off robustness and accuracy?

Figure 4 (right) shows the effect of varying diffusion terms with different ω on the accuracy of the stochastic Itô ensemble under the L_∞ norm for PGD attacks with $\epsilon = 8/255, 16/255$ and $32/255$. The accuracy of the stochastic Itô ensemble first increases as the value of ω increases and then starts decreasing for any given value of ω. This shows a tradeoff between the accuracy of the neural network on benign data and its ability to be robust to large adversarial perturbations.

IMAGENET RESULTS

These experiments are performed on a 92-core 480GB RAM server with 8 NVIDIA V100 GPUs.

Question 1: Does the Itô ensemble achieve competitive accuracy on benign inputs?

We study the accuracy of our Itô ensemble of 5 ResNet-50 models with 20 independent inferences per model for different values of ω and associated diffusion terms. As shown in Table 5, small values of ω do not significantly reduce the accuracy of the Itô ensemble on benign data.

Question 2: Is the Itô ensemble approach robust against adversarial attacks?

Figure 5 shows the accuracy of our stochastic Itô ensemble approach on ImageNet against the PGD attack with various values of ϵ under L_2 as well as L_∞ norms. The accuracy of our Itô ensemble compares favorably with the results from Madry's robustness toolbox Engstrom et al. (2020).

Diffusion Term ω	Original Accuracy	Itô Ensemble Accuracy	% Decrease in Accuracy	Correct + Abstentions (%)
0.1	76.13	76.04	0.09	77.87
0.2	76.13	73.59	2.54	78.29
0.3	76.13	67.79	8.34	76.40
0.4	76.13	61.66	14.47	76.42

Table 5: Accuracy of Itô ensembles on benign data with diffusion corresponding to different ω.

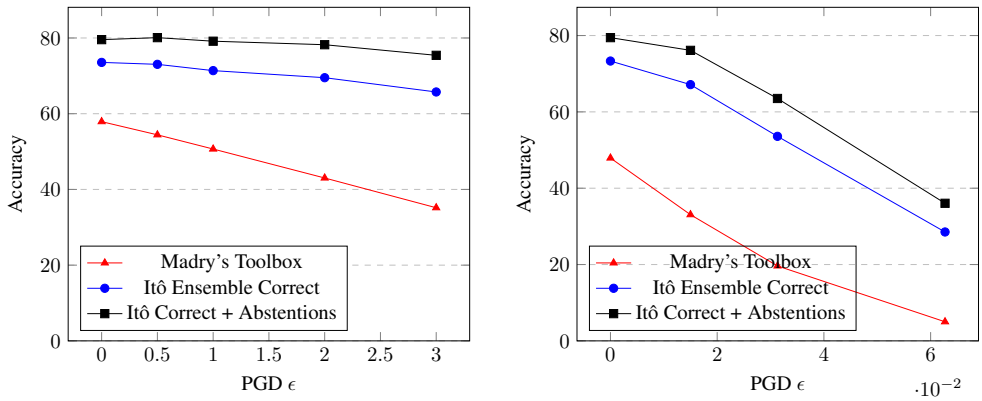

Figure 5: The accuracy of our stochastic Itô ensemble with ω=0.2 on ImageNet compares favorably with Robustness Toolbox using the L_2 norm (left) and the L_∞ norm (right) for different values of ϵ.

Question 3: How can we control the diffusion term ω to trade-off robustness and accuracy?

Table 5 and Table 6 show that the diffusion parameter ω can be used to establish a desired trade-off between robustness and benign accuracy for the ImageNet data set.

Diffusion Term ω	Itô Ensemble Benign Accuracy (%)	Itô Ensemble PGD Accuracy (%)	Correct + Abstentions (%)
0.1	76.04	26.23	27.89
0.2	73.59	53.42	61.32
0.3	67.79	60.11	70.01
0.4	61.66	58.57	73.55

Table 6: Accuracy of our Itô ensemble approach on PGD attack ($\epsilon = 8/255$) with different diffusion parameters ω. Higher values of ω lead to more robust models.

5 Conclusion

We have shown that ensembles of neural networks corresponding to a class of Itô processes are more robust than classical residual networks. An Itô process obtained by solving a suitably-formulated stochastic differential equation derived from a residual network has a probability density function that is robust to adversarial input perturbations. Further, the achieved robustness does not require any explicit adversarial training; hence, it is likely to generalize to unforeseen attacks. We empirically evaluated the robustness of our Itô ensembles and demonstrated that they achieve higher accuracy under FGSM/PGD attacks over the L_2/L_∞ norm compared to state-of-the-art methods. This robustness is attained without significantly sacrificing accuracy on benign data. Further, Itô ensemble abstains on benign inputs with high uncertainty reflecting uncertainty-aware learning. Our paper is a step towards the use of Itô processes and stochastic differential equation models to build robust ensembles in deep learning. This will aid the adoption of deep learning in safety-critical applications.

REFERENCES

Linda JS Allen. *An introduction to stochastic processes with applications to biology*. CRC Press, 2010.

Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2020.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Paul C Bressloff. *Stochastic processes in cell biology*, volume 41. Springer, 2014.

Bo Chang, Lili Meng, Eldad Haber, Frederick Tung, and David Begert. Multi-level residual networks from dynamical systems view. *arXiv preprint arXiv:1710.10348*, 2017.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Neural information processing systems*, pp. 6571–6583, 2018.

Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5261–5269, 2015.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Souradeep Dutta, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. Output range analysis for deep feedforward neural networks. In *NASA Formal Methods Symposium*, pp. 121–138. Springer, 2018.

Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *UAI*, volume 1, pp. 2, 2018.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness: A library for experimenting with, training and evaluating neural networks, with a focus on adversarial robustness., 2020. URL `https://github.com/MadryLab/robustness`.

Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1080–1088, 2015.

Didier Gonze, José Halloy, and Albert Goldbeter. Deterministic versus stochastic models for circadian rhythms. *Journal of biological physics*, 28(4):637–653, 2002.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017a.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017b.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Uyeong Jang, Susmit Jha, and Somesh Jha. On the need for topology-aware generative models for manifold-based defenses. In *International Conference on Learning Representations*, 2020.

Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55, 2014.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Alex Lamb, Jonathan Binas, Anirudh Goyal, Dmitriy Serdyuk, Sandeep Subramanian, Ioannis Mitliagkas, and Yoshua Bengio. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. *arXiv preprint arXiv:1804.02485*, 2018.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, pp. 3276–3285. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

Bernt Oksendal. *Stochastic Differential Equations (3rd Ed.): An Introduction with Applications*. Springer-Verlag, Berlin, Heidelberg, 1992. ISBN 3387533354.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. volume 97 of *Proceedings of Machine Learning Research*, pp. 4970–4979, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/pang19a.html`.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Sho Sonoda and Noboru Murata. Double continuum limit of deep neural networks. In *ICML Workshop Principled Approaches to Deep Learning*, volume 1740, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pp. 8400–8409, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.