Defense Against Adversarial Attacks by Langevin Dynamics

Vignesh Srinivasan [1] **Arturo Marban** [1 2 3] **Klaus-Robert Müller** [2 3 4 5] **Wojciech Samek** [1 3] **Shinichi Nakajima** [2 3 6]

Abstract

In the midst of an ensuing arms race between adversarial examples and defense methods, the problem of robustness against adversarial attacks remains unsolved even on the toy MNIST dataset. This paper proposes a novel, simple yet effective defense strategy. Given an off-manifold adversarial sample, our algorithm drives the adversarial samples towards high density regions of the data generating distribution of the target class by Metroplis-adjusted Langevin algorithm (MALA) with *perceptual boundary taken into account*. Although the motivation is similar to projection methods, e.g., Defense-GAN, our method, called MALA for DEfense (MALADE) is equipped with significant dispersion—projection is distributed broadly, and therefore any attack cannot accurately align the input so that the MALADE moves it to a targeted untrained spot where the model predicts a wrong label. In our experiment, MALADE exhibited state-of-the-art performance against various elaborate attacking strategies.

1. Introduction

Deep neural networks (DNNs) (Krizhevsky et al., 2012; LeCun et al., 1998; Szegedy et al., 2015; Simonyan & Zisserman, 2014; He et al., 2016) have shown excellent performance in many applications, while they are known to be susceptible to adversarial attacks, i.e., examples crafted intentionally by adding slight noise to the input (Goodfellow et al., 2014; Papernot et al., 2017; Bruna et al., 2013; Nguyen et al., 2015; Evtimov et al., 2017; Athalye & Sutskever, 2017). These two aspects are considered to be two sides of the same coin: deep structure induces complex interactions between weights of different layers, which provides flexibility in expressing complex input-output relation with relatively small degrees of freedom, while it can make the output function unpredictable in *spots* where training samples exist sparsely. If adversarial attackers would manage to find such spots in the input space close to a *real* sample, they can manipulate the behavior of classification, which can lead to a critical risk of security in applications, e.g., self-driving cars, for which high reliability is required.

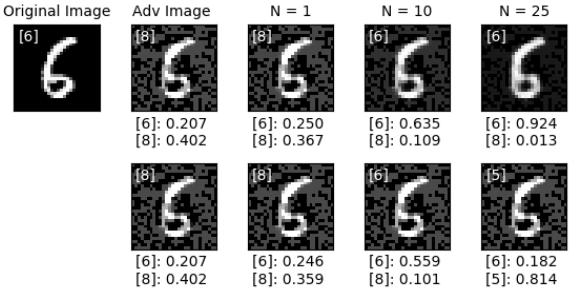


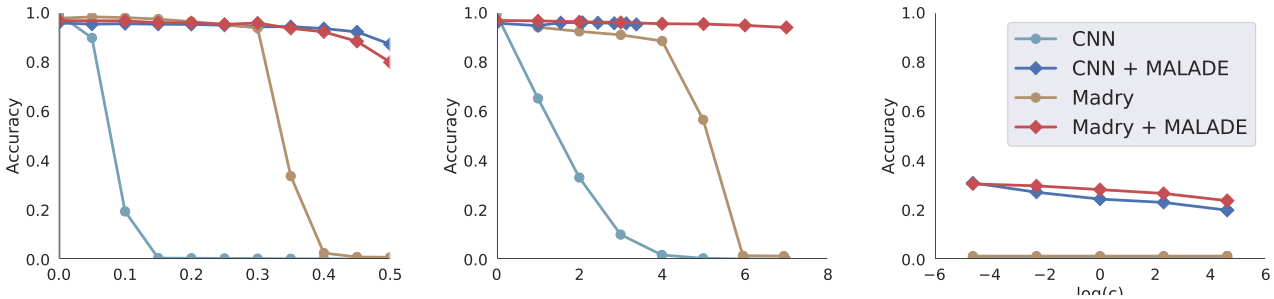
Figure 1. Top-left image is the original image while the next image is the adversarial image crafted for it. Top-row shows the steps N = 1, 10, and 25 of MALADE and the bottom-row shows the sampling using MALA based on the *marginal* training distribution $p(\boldsymbol{x})$. Since MALADE has prior knowledge of the target labels, the gradient flow drives it towards the right class. With MALA, the flow of the gradients can lead it to a cluster of neighboring classes.

Existing defense strategies can be roughly classified into three categories: (i) incorporation of adversarial samples in the training phase (Tramèr et al., 2017; Papernot et al., 2016; Strauss et al., 2017; Gu & Rigazio, 2014; Lamb et al., 2018; Madry et al., 2017; Kannan et al., 2018; Liu et al., 2018; Xie et al., 2018), (ii) projection of adverserial samples onto the estimated data manifold (Schott et al., 2018; Song et al., 2017; Samangouei et al., 2018; Ilyas et al., 2017; Lamb et al., 2018; Shen et al., 2017), and (iii) preprocessing to destroy elaborate spatial coherence hidden in adversarial samples (Guo et al., 2017; Liao et al., 2017; Meng & Chen, 2017; Xie et al., 2017). Since the methods in category (iii) have been rendered ineffective against elaborate attacking strategies (Athalye et al., 2018; Uesato et al., 2018; Engstrom et al., 2018), this paper focuses on the first two categories.

Based on previous successes and failures, this paper proposes a novel defense strategy, where adversarial samples are driven towards high density areas of the data manifold. Our approach is to *relax* the adversarial sample by Metropolis-adjusted Langevin algorithm (MALA) (Roberts

[1]Fraunhofer HHI [2]TU Berlin [3]Berlin Big Data Center [4]Korea University [5]MPI for Informatics [6]RIKEN AIP. Correspondence to: VS <vignesh.srinivasan@hhi.fraunhofer.de>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning.



(a) PGD-L_∞ (b) PGD-L_2 (c) EAD-$(L_1 + L_2)$

Figure 2. Classification accuracy on MNIST against (a) PGD-L_∞, (b) PGD-L_2, and (c) EAD attacks. The curves correspond to the original convolutional neural network (CNN) classifier, the CNN classifier protected by MALADE (CNN+MALADE), the Madry classifier (Madry et al., 2017), and the Madry classifier protected by MALADE (Madry+MALADE). In each plot, the horizontal axis indicates the intensity of the adversarial perturbations, i.e., ε for PGD and $\log c$ for EAD. All plots show that MALADE significantly boosts the robustness of both classifiers.

& Rosenthal, 1998; Roberts et al., 1996), an efficient Markov chain Monte Carlo (MCMC) sampling method.

MALA requires the gradient $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ of the energy function, which corresponds to the gradient of the log probability or the score function of the training data. However, just naively applying MALA would have an apparent drawback: if there exist high density regions (clusters) close to each other but not sharing the same label, MALA could drive a sample into another cluster, which degrades the classification accuracy. To overcome this drawback, we replace the (marginal) training distribution $p(\boldsymbol{x})$ with the conditional training distribution $p(\boldsymbol{x}|\boldsymbol{y})$ given label $\boldsymbol{y}$. More specifically, our novel defense method, called MALA for defense (MALADE), relaxes the adversarial sample based on the conditional gradient $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|\boldsymbol{y})$ by using a novel estimator for the conditional gradient *without knowing the label $\boldsymbol{y}$ of the test sample*. Thus, MALADE drives the adversarial sample towards high density regions of the data generating distribution for the original class, where the classifier is well trained to predict the correct label.

In our experiment, we show that MALADE performs comparably to the state-of-the-art methods for the known attacks with low perturbation intensity, while it significantly outperforms the state-of-the-art methods for unknown attacks or for high perturbation intensity. We also show an additional advantage of our method as a projection method—MALADE can be combined with the state-of-the-art method in category (i), which further boosts the performance.

2. Proposed Method

Metropolis-adjusted Langevin Algorithm (MALA MALA is an efficient Markov chain Monte Carlo (MCMC) sampling method which uses the gradient of the energy (negative log-probability $E(\boldsymbol{x}) = -\log p(\boldsymbol{x})$). Sampling is performed sequentially by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{\alpha} \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t) + \boldsymbol{\kappa}, \tag{1}$$

where $\boldsymbol{\alpha}$ is the step size, and $\boldsymbol{\kappa}$ is random perturbation subject to $\mathcal{N}(\boldsymbol{0}, \delta^2 \boldsymbol{I}_L)$. By appropriately controlling the step size $\boldsymbol{\alpha}$ and the noise variance δ^2, the sequence is known to converge to the distribution $p(\boldsymbol{x})$.[1] Nguyen et al. (2016) successfully generated realistic artificial images that follow the natural image distribution with the gradient estimated by denoising autoencoders (DAE). They relied on the following proposition:

Proposition 1 *(Alain & Bengio, 2014) The function $\boldsymbol{r}(\boldsymbol{x})$ minimizing the DAE objective $\mathbb{E}_{p(\boldsymbol{x})\mathcal{N}(\boldsymbol{\nu};\boldsymbol{0},\sigma^2\boldsymbol{I}_L)}\left[\|\boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\nu}) - \boldsymbol{x}\|^2\right]$ satisfies*

$$\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{x} = \sigma^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) + o(\sigma^2) \ \text{ as } \ \sigma^2 \to 0. \tag{2}$$

MALA for DEfense (MALADE) MALADE performs sampling on the *joint distribution* $p(\boldsymbol{x}, \boldsymbol{y})$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{\alpha} \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x}_t)} \left[\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t, \boldsymbol{y})\right] + \boldsymbol{\kappa}. \tag{3}$$

with the gradient estimated by the *supervised* DAE (sDAE).

Theorem 1 *Let $J(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{k=1}^{K} y_k \log \widehat{y}_k(\boldsymbol{x})$ be the cross entropy loss of the classifier output $\widehat{\boldsymbol{y}} \in [0,1]^K$ for the true label $\boldsymbol{y}$, and assume that the classifier output accurately reflects the conditional probability of the training data, i.e., $\widehat{\boldsymbol{y}}(\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})$, then the function $\boldsymbol{r}(\boldsymbol{x})$ minimizing the sDAE objective*

[1] For convergence, a rejection step after Eq.(1) is required. However, it was observed that a variant, called Mala-approx (Nguyen et al., 2016), without the rejection step gives reasonable sequence for moderate step sizes. We use Mala-approx in our proposed method.

Table 1. Classification performance of defense methods against whitebox and blackbox attacks on MNIST. The intensity of perturbation is set to $\epsilon = 0.3, 0.4$ for FGSM, PGD-L_∞ and MIM, and to $\epsilon = 4$ for PGD-L_2. For EAD, we set $\beta = 0.01$ and $c = 0.01$.

Setting	Norm	Attack	CNN	Madry	CNN + MALADE	Madry + MALADE
whitebox	L_∞ $\epsilon = 0.3$	FGSM	11.77	97.52	93.54	95.59
		PGD	0.00	93.71	94.22	95.76
		R+PGD	-	-	92.65	93.51
		BPDA	-	-	84.74	94.70
		BPDAwEOT	-	-	82.48	91.54
		MIM	0.00	97.66	94.32	94.53
		worst case	0.00	**93.71**	82.48	91.54
	L_∞ $\epsilon = 0.4$	PGD	0.00	0.02	93.51	92.18
		BPDA	-	-	66.16	83.90
		BPDAwEOT	-	-	62.15	80.65
		worst case	0.00	0.02	62.15	**80.65**
	L_2	FGM	30.79	97.68	94.68	96.05
		PGD	0.01	92.68	95.91	96.76
		CW	0.00	85.53	90.07	91.14
		worst case	0.0	85.53	90.07	**91.14**
	L_2-L_1	EAD	0.00	0.01	**31.00**	30.59
blackbox	$\rho = 0.25$ $T = 5,000$	SaltnPepper	36.49	41.61	80.41	80.72
		Boundary Attack	32.39	1.10	93.79	95.80
		worst case	32.39	1.10	80.41	**80.72**

$\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})\mathcal{N}(\boldsymbol{\nu};\mathbf{0},\sigma^2\boldsymbol{I}_L)}\left[\|\boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\nu})-\boldsymbol{x}\|^2 - 2\sigma^2 J(\boldsymbol{x}+\boldsymbol{\nu},\boldsymbol{y})\right]$ *satisfies*

$$\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{x} = \sigma^2 \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x})}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x},\boldsymbol{y})\right] + O(\sigma^3). \quad (4)$$

Since $p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$, if the label distribution is flat (or equivalently the number of training samples for all classes is the same), i.e., $p(\boldsymbol{y}) = 1/K$, the residual of sDAE gives

$$\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{x} = \sigma^2 \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x})}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}|\boldsymbol{y})\right] + O(\sigma^3).$$

The first term is the gradient of the log-*conditional*-distribution on the label, where the label is estimated from the prior knowledge (the expectation is taken over the training distribution of the label, given $\boldsymbol{x}$). If the number of training samples are non-uniform over the classes, the weight (or the step size) should be adjusted so that all classes contribute equally to the training.

3. Experiments

3.1. MNIST

We perform elaborate experiments on MNIST applying MALADE on CNN as well as state-of-the-art model Madry (Madry et al., 2017). Fig. 2 shows classification accuracy of defense methods against PGD-L_∞ and PGD-L_2 and EAD. The attacking strategies tested in Fig. 2 are not optimized for stochastic defense methods. Hence, we chose BPDA and EOT as reasonable whitebox attacks against MALADE. Fig. 4 shows the performance of "Mardy + MALADE" against PGD-L_∞, BPDA, EOT, and their reasonable combinations. Table 1 summarizes classification performance of defense strategies against various attacks. We see that, although BPDA and EOT reduce the accuracy of Madry + MALADE to some extent, their effect is limited and even in the worst case, i.e., against "BPDA with EOT", Madry + MALADE performs better than the baseline methods, i.e., ABS and Madry against PGD-L_∞ attack. Although MALADE is outperformed by Madry up to $\epsilon = 0.3$, it outperforms Madry in the other cases, i.e., against PGD-L_∞ with BPDA and EOT for larger intensity $\epsilon = 0.4$, PGD-L_2, and EAD. Table 1 also shows results under the blackbox scenario with the state-of-the-art attacking strategies, SaltnPepper and Boundary attack. The table clearly shows high robustness of MALADE against those attacks, while Madry exhibits its vulnerability against them.

3.2. CIFAR10 and TinyImagenet

Here we apply our defense strategy for larger datasets. Tables 2 report on classification accuracy on CIFAR10 and on

Table 2. Summary of classification performance on CIFAR10 and TinyImagenet. For PGD attack, the number of steps was fixed at 100. For blackbox attacking scenario, the lowest row gives the *worst case* result over the considered attacking strategies.

Dataset	Setting	Condition	Attack	CNN	Madry	CNN + MALADE	Madry + MALADE
CIFAR10	*whitebox*	L_∞ $\varepsilon = 8$	PGD	0.00	32.86	0.48	**33.89**
		L_∞ $\varepsilon = 16$	PGD	0.00	10.28	0.16	**11.46**
	blackbox	$\rho = 0.05$	SaltnPepper	13.54	11.09	15.19	11.12
		$T = 10,000$	Boundary Attack	12.76	34.97	17.72	73.57
			worst case	12.76	11.09	**15.19**	11.12
Dataset	Setting	Condition	Attack	CNN	ALP	CNN + MALADE	ALP + MALADE
Tiny-Imagenet	*whitebox*	L_∞ $\varepsilon = 1$	PGD	9.49	27.05	10.07	**31.80**
		L_∞ $\varepsilon = 4$	PGD	0.20	12.78	0.26	**13.23**
	blackbox	$\rho = 0.05$	SaltnPepper	2.45	9.91	6.23	9.94
		$T = 10,000$	Boundary Attack	9.12	7.38	63.54	44.35
			worst case	2.45	7.38	6.23	**9.94**

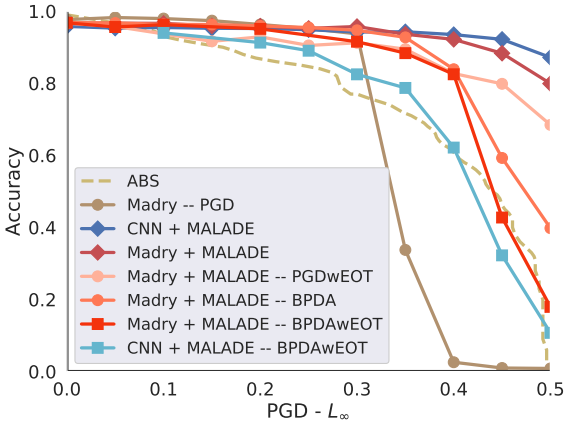


Figure 3. PGD-L_∞

Figure 4. Classification accuracy on MNIST dataset against PGD L_∞ attacks adapted for MALADE.

TinyImagenet, respectively, against whitebox (PGD-L_∞) and blackbox (SaltnPepper and Boundary) attacks. On TinyImagenet, we show the performance of ALP (Kannan et al., 2018) as the baseline defense method. Similarly to Madry+MALADE, it is straightforward to combine MALADE and ALP to form ALP+MALADE.

Overall, classification accuracy is not high, which reflects the fact that the state-of-the-art has not achieved satisfactory level of robustness against adversarial attacks in the scale of CIFAR10 and TinyImagenet datasets. However, the observation that MALADE consistently improves the robustness of classifiers implies that our approach is a hopeful direction for solving the issue of defense against adversarial attacks in large scale problems.

4. Conclusion

In this work we have proposed to use MALA, which is guided through a sDAE - MALADE. This framework allows to drive the adversarial samples towards the underlying data manifold and thus towards the high density regions of the data generating distribution which were originally used for training the nonlinear learning machine. We have empirically showed that MALADE is fairly robust—it compares favorably or significantly outperforms the state-of-the-art defense methods under different adversarial scenarios and attacking strategies.

Acknowledgments

This work was supported by the Fraunhofer Society under the MPI-FhG collaboration project "Theory & Practice for Reduced Learning Machines". This work was also supported by the German Ministry for Education and Research (BMBF) as Berlin Big Data Center (01IS18025A) and Berlin Center for Machine Learning (01IS18037I), the German Research Foundation (DFG) as Math+: Berlin Mathematics Research Center (EXC 2046/1, project-ID: 390685689), and the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451, No. 2017-0-01779).

References

Alain, G. and Bengio, Y. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Athalye, A. and Sutskever, I. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Bruna, J., Szegedy, C., Sutskever, I., Goodfellow, I., Zaremba, W., Fergus, R., and Erhan, D. Intriguing properties of neural networks. 2013.

Engstrom, L., Ilyas, A., and Athalye, A. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Guo, C., Rana, M., Cissé, M., and van der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ilyas, A., Jalal, A., Asteri, E., Daskalakis, C., and Dimakis, A. G. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lamb, A., Binas, J., Goyal, A., Serdyuk, D., Subramanian, S., Mitliagkas, I., and Bengio, Y. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. *arXiv preprint arXiv:1804.02485*, 2018.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liao, F., Liang, M., Dong, Y., Pang, T., Zhu, J., and Hu, X. Defense against adversarial attacks using high-level representation guided denoiser. *arXiv preprint arXiv:1712.02976*, 2017.

Liu, X., Li, Y., Wu, C., and Hsieh, C.-J. Adv-bnn: Improved adversarial defense through robust bayesian neural network. *arXiv preprint arXiv:1810.01279*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Meng, D. and Chen, H. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147. ACM, 2017.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, June 2015. doi: 10.1109/CVPR.2015.7298640.

Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., and Clune, J. Plug & play generative networks: Conditional iterative generation of images in latent space. *arXiv preprint arXiv:1612.00005*, 2016.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pp. 582–597. IEEE, 2016.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Roberts, G. O. and Rosenthal, J. S. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Roberts, G. O., Tweedie, R. L., et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, volume 9, 2018.

Schott, L., Rauber, J., Brendel, W., and Bethge, M. Robust perception through analysis by synthesis. *arXiv preprint arXiv:1805.09190*, 2018.

Shen, S., Jin, G., Gao, K., and Zhang, Y. Ape-gan: Adversarial perturbation elimination with gan. *ICLR Submission, available on OpenReview*, 4, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Strauss, T., Hanselmann, M., Junginger, A., and Ulmer, H. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1709.03423*, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015. doi: 10.1109/CVPR.2015.7298594.

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Uesato, J., O'Donoghue, B., Oord, A. v. d., and Kohli, P. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Xie, C., Wu, Y., van der Maaten, L., Yuille, A., and He, K. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.