STOCHASTIC SECURITY: ADVERSARIAL DEFENSE USING LONG-RUN DYNAMICS OF ENERGY-BASED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The vulnerability of deep networks to adversarial attacks is a central problem for deep learning from the perspective of both cognition and security. The current most successful defense method is to train a classifier using adversarial images created during learning. Another defense approach involves transformation or purification of the original input to remove adversarial signals before the image is classified. We focus on defending naturally-trained classifiers using Markov Chain Monte Carlo (MCMC) sampling with an Energy-Based Model (EBM) for adversarial purification. In contrast to adversarial training, our approach is intended to secure highly vulnerable pre-existing classifiers. To our knowledge, no prior defense method is capable of securing naturally-trained classifiers, and our method is the first to validate a post-training defense approach that is distinct from current successful defenses which modify classifier training.

The memoryless behavior of long-run MCMC sampling will eventually remove adversarial signals, while metastable behavior preserves consistent appearance of MCMC samples after many steps to allow accurate long-run prediction. Balancing these factors can lead to effective purification and robust classification. We evaluate adversarial defense with an EBM using the strongest known attacks against purification. Our contributions are 1) an improved method for training EBM's with realistic long-run MCMC samples for effective purification, 2) an Expectation-Over-Transformation (EOT) defense that resolves ambiguities for evaluating stochastic defenses and from which the EOT attack naturally follows, and 3) state-of-the-art adversarial defense for naturally-trained classifiers and competitive defense compared to adversarial training on CIFAR-10, SVHN, and CIFAR-100.

1 MOTIVATION AND CONTRIBUTIONS

Deep neural networks are highly sensitive to small input perturbations. This sensitivity can be exploited to create adversarial examples that undermine robustness by causing trained networks to produce defective results from input changes that are imperceptible to the human eye (Goodfellow et al., 2015). The adversarial scenarios studied in this paper are primarily untargeted white-box attacks on image classification networks. White-box attacks have full access to the classifier (in particular, to classifier gradients) and are the strongest attacks against the majority of defenses.

Many whitebox methods have been introduced to create adversarial examples. Strong iterative attacks, such as Projected Gradient Descent (PGD) (Madry et al., 2018), can reduce the accuracy of a naturally-trained classifier to virtually 0. Currently the most robust form of adversarial defense is to train a classifier on adversarial samples in a procedure known as *adversarial training* (AT) (Madry et al., 2018). Another defense strategy, which we will refer to as *adversarial preprocessing* (AP), uses defensive transformations to purify an image and remove adversarial signals before classification (Song et al., 2018; Srinivasan et al., 2019; Yang et al., 2019). AP is an attractive strategy compared to AT because it has the potential to secure vulnerable pre-existing classifiers. Defending naturally-trained classifiers is the central focus of this work.

The work (Athalye et al., 2018) revealed that many preprocessing defenses can be overcome with minor adjustments to the standard PGD attack. Both stochastic behavior from preprocessing and the

computational difficulty of end-to-end backpropagation can be circumvented to attack the classifier through the defensive transformation In this paper we carefully address (Athalye et al., 2018) to evaluate AP with an EBM using attacks with the greatest known efficiency and effectiveness.

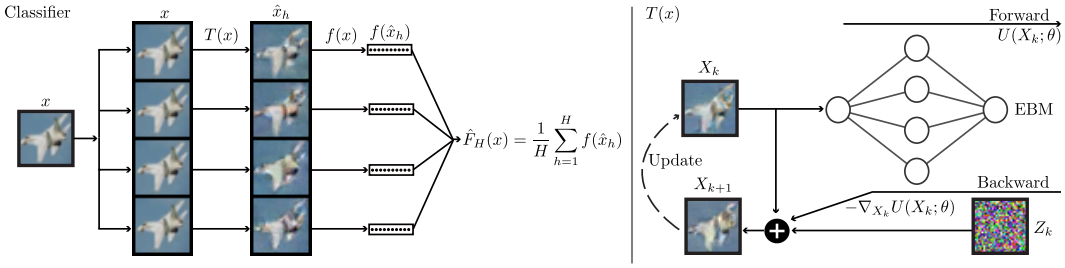

Figure 1: *Left:* Visualization of calculating our stochastic logits $\hat{F}_H(x)$ from (5). The input image x is replicated H times and parallel Langevin updates with a ConvNet EBM are performed on each replicate to generate $\{\hat{x}_h\}_{h=1}^{H}$. Purified samples are sent in parallel to our naturally-trained classifier network $f(x)$ and the resulting logits $\{f(\hat{x}_h)\}_{h=1}^{H}$ are averaged to produce $\hat{F}_H(x)$. The logits $\hat{F}_H(x)$ give an approximation of our true classifier logits $F(x)$ in (4) that can be made arbitrarily precise by increasing H. *Right:* Graphical diagram of the Langevin dynamics (3) that we use for $T(x)$. Images are iteratively updated with a gradient from a naturally-trained EBM (1) and Gaussian noise Z_k.

Langevin sampling using an EBM with a ConvNet potential (Xie et al., 2016) has recently emerged as a method for AP (Du & Mordatch, 2019; Grathwohl et al., 2020). However, the proposed defenses are not competitive with AT (see Table 1 and (Croce & Hein, 2020)). In the present work we demonstrate that EBM defense of a naturally-trained classifier can be stronger than standard AT in (Madry et al., 2018) and competitive compared to modified AT such as (Zhang et al., 2019; Carmon et al., 2019).

Our defense tools are a classifier trained with labeled natural images and an EBM trained with unlabeled natural images. For prediction, we perform Langevin sampling with the EBM and send the sampled images to the natural classifier. An intuitive visualization of our defense method is shown in Figure 1. Langevin chains constitute a memoryless trajectory that removes adversarial signals, while metastable sampling behaviors preserve image classes over long-run trajectories. Balancing these two factors leads toward effective adversarial defense. Our main contributions are:

- A simple but effective adjustment to improve the convergent learning procedure in (Nijkamp et al., 2020). Our adjustment enables stable long-run sampling for complex datasets such as CIFAR-10.
- An Expectation-Over-Transformation (EOT) defense that prevents the possibility of a stochastic defense breaking due to random variation in prediction instead of an adversarial signal. The EOT attack (Athalye et al., 2018) naturally follows from the EOT defense.
- Experiments showing state-of-the-art defense for naturally-trained classifiers and competitive defense compared to state-of-the-art AT.

2 Improved Convergent Learning of Energy-Based Models

The Energy-Based Model introduced in (Xie et al., 2016) is a Gibbs-Boltzmann density

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\{-U(x;\theta)\} \tag{1}$$

where $x \in \mathbb{R}^D$ is an image signal, $U(x;\theta)$ is a ConvNet with weights θ and scalar output, and $Z(\theta) = \int_{\mathcal{X}} \exp\{-U(x;\theta)\}dx$ is the intractable normalizing constant. Given i.i.d. samples from a data distribution $q(x)$, one can learn a parameter θ^* such that $p(x;\theta^*) \approx q(x)$ by minimizing the expected negative log-likelihood $\mathcal{L}(\theta) = E_q[-\log p(X;\theta)]$ of the data samples. Network weights θ

are updated using the loss gradient

$$\nabla \mathcal{L}(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta U(X_i^+; \theta) - \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta U(X_i^-; \theta) \tag{2}$$

where $\{X_i^+\}_{i=1}^n$ are a batch of training images and $\{X_i^-\}_{i=1}^m$ are i.i.d. samples from $p(x; \theta)$ obtained via MCMC. Iterative application of the Langevin update

$$X_{k+1} = X_k - \frac{\tau^2}{2} \nabla_{X_k} U(X_k; \theta) + \tau Z_k, \tag{3}$$

where $Z_k \sim \mathrm{N}(0, I_D)$ and $\tau > 0$, is often used to obtain the samples $\{X_i^-\}_{i=1}^m$.

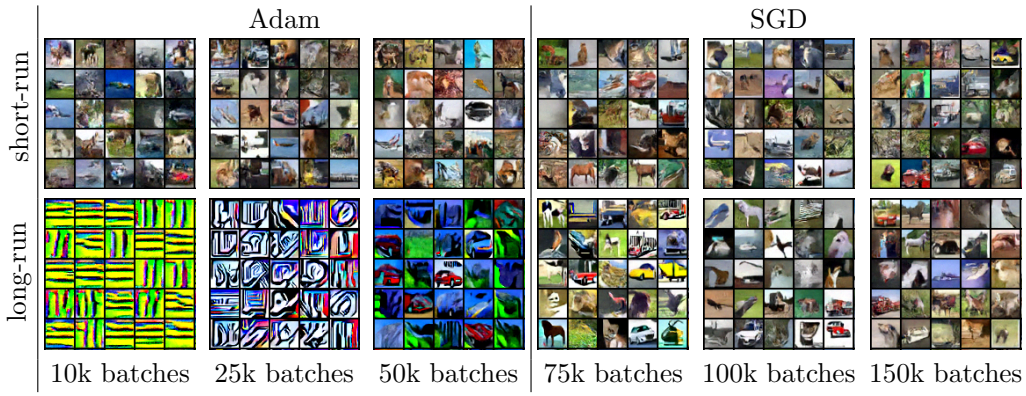

Figure 2: Comparison of long-run and short-run samples over model updates for our improved method of convergent learning. The model is updated in a non-convergent learning phase with the Adam optimizer for the first 50,000 batches. The majority of short-run synthesis realism is learned during this phase, but the long-run samples are very unrealistic. The second learning phase uses SGD with a low learning rate. Short-run synthesis changes very little, but the long-run distribution gradually aligns with the short-run distribution.

The recent work (Nijkamp et al., 2020) reveals that EBM learning heavily gravitates towards an unexpected outcome where short-run MCMC samples have a realistic appearance and long-run MCMC samples have an unrealistic appearance. The work uses the term *convergent learning* to refer to the expected outcome where short-run and long-run MCMC samples have similar appearance, and the term *non-convergent learning* to refer to the unexpected but prevalent outcome where models have realistic short-run samples and oversaturated long-run samples. Convergent learning is essential for our defense strategy because long-run samples must be realistic for classifiers to maintain high accuracy after Langevin transformation.

As observed in (Nijkamp et al., 2020), we were unable to learn a convergent model when updating θ with the Adam optimizer (Kingma & Ba, 2015). Despite the drawbacks of Adam for convergent learning, it is a very effective tool for obtaining realistic short-run synthesis. Drawing inspiration from classifier training in (Keskar & Socher, 2017), we learn a convergent EBM in two phases. The first phase uses Adam for θ updates to achieve realistic short-run synthesis. The second phase uses SGD for θ updates to align short-run and long-run MCMC samples to correct the degenerate steady-state from the Adam phase. This modification allows us to learn the convergent EBMs for complex datasets such as CIFAR-10 using 20% of the computational budget in (Nijkamp et al., 2020). We use the light-weight EBM from (Nijkamp et al., 2019) as our network architecture. See Figure 2.

We use long run chains for our EBM defense to remove adversarial signals while maintaining image features. The steady-state convergence property ensures adversarial signals will eventually vanish, while metastable behaviors preserve features of the initial state. Theoretical perspectives on our defense can be found in Appendix C and a comparison of convergent and non-convergent EBM defense can be found in Appendix D.

3 ATTACK AND DEFENSE FORMULATION

3.1 CLASSIFICATION WITH STOCHASTIC TRANSFORMATIONS

Let $T(x)$ be a stochastic pre-processing transformation for a state $x \in \mathbb{R}^D$. Given a fixed input x, the transformed state $T(x)$ is a random variable over $\mathbb{R}^D$. In this work, $T(x)$ is obtained from K steps of the Langevin update (3) starting from $X_0 = x$. One can compose $T(x)$ with a deterministic classifier $f(x) \in \mathbb{R}^J$ (for us, a naturally-trained classifier) to define a new classifier $F(x) \in \mathbb{R}^J$ as

$$F(x) = E_{T(x)}[f(T(x))]. \tag{4}$$

$F(x)$ is a deterministic classifier even though $T(x)$ is stochastic. The predicted label for x is then $c(x) = \arg\max_j F(x)_j$. In this work, $f(x)$ denotes logits and $F(x)$ denotes expected logits, although other choices are possible (e.g. softmaxes). We refer to (4) as an Expectation-Over-Transformation (EOT) defense. The classifier $F(x)$ in (4) is simply the target of the EOT attack (Athalye et al., 2018). The importance of the EOT formulation is well-established for adversarial attacks, but its importance for adversarial defense has not yet been explored. Although direct evaluation of $F(x)$ is generally impossible, the law of large numbers ensures that the finite-sample approximation of $F(x)$ given by

$$\hat{F}_H(x) = \frac{1}{H}\sum_{h=1}^{H} f(\hat{x}_h) \quad \text{where,} \quad \hat{x}_h \sim T(x) \text{ i.i.d.,} \tag{5}$$

can approximate $F(x)$ to any degree of accuracy for a sufficiently large sample size H. In other words, $F(x)$ is intractable but trivial to accurately approximate via $\hat{F}_H(x)$ given enough computation.

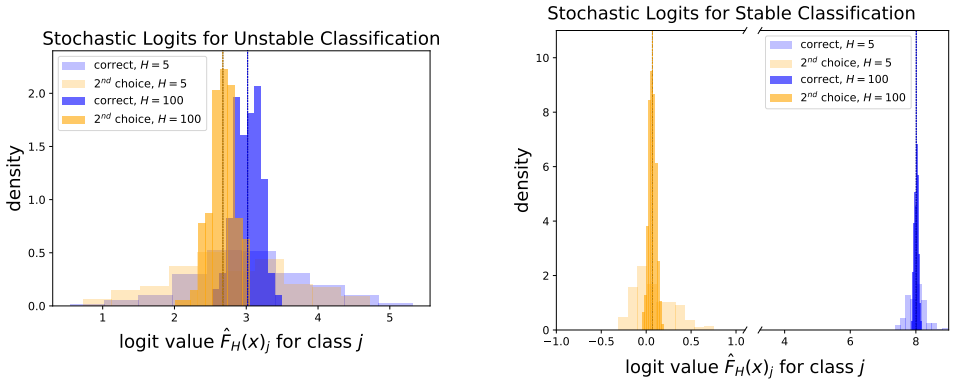

Figure 3: The histograms display different realizations of the logits $\hat{F}_H(x)$ for the correct class and the second most probable class for images x_1 (*left*) and x_2 (*right*) over different choices of H. In both cases, $F(x)$ (dashed vertical lines) gives correct classification. However, the overlap between the logit histograms of $\hat{F}_H(x_1)$ indicate a high probability of misclassification even for large H, while $\hat{F}_H(x_2)$ gives correct prediction even for small H because the histograms are well-separated. The EOT defense formulation (4) is essential for securing borderline images such as x_1.

In the literature both attackers (Athalye et al., 2018; Tramer et al., 2020) and defenders (Song et al., 2018; Srinivasan et al., 2019; Yang et al., 2019; Grathwohl et al., 2020) evaluate stochastic classifiers of the form $f(T(x))$ using either $\hat{F}_1(x)$ or $\hat{F}_{H_{\text{adv}}}(x)$ where H_{adv} is the number of EOT attack samples, typically around 10 to 30. This evaluation is not sound when $\hat{F}_H(x)$ has a small but plausible chance of misclassification because randomness alone could cause x to be identified as an adversarial image even though $\hat{F}_H(x)$ gives the correct prediction on average (see Figure 3). In experiments with EBM defense, we identify many images x that exhibit variation in the predicted label

$$\hat{c}_H(x) = \arg\max_j \hat{F}_H(x)_j \tag{6}$$

for smaller $H \approx 10$ but which have consistently correct prediction for larger $H \approx 150$. Fair evaluation of stochastic defenses must be based on the deterministic EOT defense F in (4) and attackers must use sufficiently large H to ensure that $\hat{F}_H(x)$ accurately approximates $F(x)$ before declaring that an attack against an image x is successful. In practice, we observe that $\hat{F}_{150}$ is sufficiently stable to evaluate F over several hundred attack steps of Algorithm 1 for our EBM defense.

3.2 Attacking Stochastic Classifiers

Let $L(F(x), y) \in \mathbb{R}$ be the loss (e.g. cross-entropy) between a label $y \in \{1, \ldots, J\}$ the outputs $F(x) \in \mathbb{R}^J$ of a classifier (e.g. the logits) for an image $x \in \mathbb{R}^D$. For a given pair of observed data (x^+, y), an untargeted white-box adversarial attack searches for the state

$$x_{\text{adv}}(x^+, y) = \arg\max_{x \in S} L(F(x), y) \tag{7}$$

that maximizes loss for predicting y in a set $S \subset \mathbb{R}^D$ centered around x^+. In this work, a natural image x^+ will have pixels intensities from 0 to 1 (i.e. $x^+ \in [0,1]^D$). One choice of S is the intersection of the image hypercube $[0,1]^D$ and the l_∞-norm ε-ball around x^+ for suitably small $\varepsilon > 0$. Another option is the intersection of the hypercube with the l_2-norm ε-ball.

The Projected Gradient Descent (PGD) attack (Madry et al., 2018) is the standard benchmark when S is the ε-ball in the l_p norm. PGD begins at a random $x_0 \in S$ and maximizes (7) by iteratively updating x_i with

$$x_{i+1} = \prod_S (x_i + \alpha g(x_i, y)), \quad g(x, y) = \arg\max_{\|v\|_p \leq 1} v^\top \Delta(x, y), \tag{8}$$

where Π_S denotes projection onto S, $\Delta(x, y)$ is the attack gradient, and $\alpha > 0$ is the attack step size. Standard PGD uses the gradient $\Delta_{\text{PGD}}(x, y) = \nabla_{x_i} L(F(x_i), y)$.

The EOT attack (Athalye et al., 2018) circumvents the intractability of F by attacking finite sample logits $\hat{F}_{H_{\text{adv}}}$, where H_{adv} is the number of EOT attack samples, with the gradient

$$\Delta_{\text{EOT}}(x, y) = \nabla_x L(\hat{F}_{H_{\text{adv}}}(x), y). \tag{9}$$

The EOT attack is the natural adaptive attack for our EOT defense formulation. Another challenge when attacking a preprocessing defense is the computational infeasibility or theoretical impossibility of differentiating $T(x)$. The Backward Pass Differentiable Approximation (BPDA) technique (Athalye et al., 2018) uses an easily differentiable function $g(x)$ such that $g(x) \approx T(x)$ to attack $F(x) = f(T(x))$. One calculates the attack loss using $L(f(T(x)), y)$ on the forward pass but calculates the attack gradient using $\nabla_x L(f(g(x)), y)$ on the backward pass. A simple but effective form of BPDA is the identity approximation $g(x) = x$. This approximation is reasonable for preprocessing defenses that seek to remove adversarial signals while preserving the main features of the original image. When $g(x) = x$, the BPDA attack gradient is $\Delta_{\text{BPDA}}(x, y) = \nabla_z L(f(z), y)$ where $z = T(x)$. Intuitively, this attack obtains an attack gradient with respect to the purified image and applies it to the original image.

Combining the EOT attack and BPDA attack with identity $g(x) = x$ gives the attack gradient

$$\Delta_{\text{BPDA+EOT}}(x, y) = \frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} \nabla_{\hat{x}_h} L\left(\frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} f(\hat{x}_h), y\right), \quad \hat{x}_h \sim T(x) \text{ i.i.d.} \tag{10}$$

The BPDA+EOT attack represents the strongest known attack against preprocessing defense, as shown by its effective use in recent works (Srinivasan et al., 2019; Tramer et al., 2020). We use $\Delta_{\text{BPDA+EOT}}(x, y)$ in (8) as our primary attack to evaluate the EOT defense (4). Pseudo-code for our adaptive attack can be found in Algorithm 1 in Appendix A.

4 Experiments

We use two different network architectures in our experiments. The first network is the lightweight EBM from (Nijkamp et al., 2019). The second network is a 28×10 WideResNet (Zagoruyko

& Komodakis, 2016) classifier. The EBM and classifier are trained independently on the same dataset. No adversarial training or other training modifications are used for either model. We use the parameters from Algorithm 1 for all evaluations unless otherwise noted. Code and pre-trained models to reproduce these results are publicly available.

4.1 PGD Attack from Base Classifier for CIFAR-10 Dataset

We first evaluate our defense using adversarial images created from a PGD attack on the classifier $f(x)$. Since this attack does not incorporate the Langevin sampling from $T(x)$, the adversarial images in this section are relatively easy to secure with Langevin transformation. This attack serves as a benchmark for comparing our defense to the IGEBM (Du & Mordatch, 2019) and JEM (Grathwohl et al., 2020) models that also evaluate adversarial defense with a ConvNet EBM (1). For all methods, we evaluate the base classifier and the EBM defense for $K = 10$ Langevin steps (as in prior defenses) and $K = 1500$ steps (as in our defense). The results are displayed in Table 1.

Table 1: CIFAR-10 accuracy for our EBM defense and prior EBM defenses against a PGD attack from the base classifier $f(x)$ with l_∞ perturbation $\varepsilon = 8/255$. (**evaluated on 1000 images*)

	Base Classifier $f(x)$		EBM Defense, $K = 10$		EBM Defense, $K = 1500$	
	Nat.	Adv.	Nat.	Adv.	Nat.	Adv.
Ours	0.9530	0.0000	0.9586	0.0001	0.8412	0.7891
(Du & Mordatch, 2019)	0.4714	0.3219	0.4885	0.3674	0.487*	0.375*
(Grathwohl et al., 2020)	0.9282	0.0929	0.9093	0.1255	0.755*	0.238*

Our natural classifier $f(x)$ has a high base accuracy but no robustness. The JEM base classifier has high natural accuracy and minor robustness, while the IGEBM base classifier has significant robustness but very low natural accuracy. Short-run sampling with $K = 10$ Langevin steps does not significantly increase robustness for any model. Long-run sampling with $K = 1500$ steps provides a dramatic increase in defense for our method but only minor increases for the prior methods. Further discussion of the IGEBM and JEM defenses can be found in Appendix F.

4.2 BPDA+EOT Attack

In this section, we evaluate our EBM defense using the adaptive BPDA+EOT attack (10) designed specifically for our defense. This attack is recently used in (Tramer et al., 2020) to evaluate the preprocessing defense (Yang et al., 2019) that is very similar to our method.

Table 2: Defense vs. whitebox attacks with l_∞ perturbation $\varepsilon = 8/255$ for CIFAR-10.

Defense	$f(x)$ Train Ims.	$T(x)$ Method	Attack	Nat.	Adv.
Ours	Natural	Langevin	BPDA+EOT	0.8412	0.5490
(Madry et al., 2018)	Adversarial	–	PGD	0.873	0.458
(Zhang et al., 2019)	Adversarial	–	PGD	0.849	0.5643
(Carmon et al., 2019)	Adversarial	–	PGD	0.897	0.625
(Song et al., 2018)	Natural	Gibbs Update	BPDA	0.95	0.09
(Srinivasan et al., 2019)	Natural	Langevin	PGD	–	0.0016
(Yang et al., 2019)	Transformed	Mask + Recon.	BPDA+EOT	0.94	0.15

CIFAR-10. We ran 5 random restarts of the BPDA+EOT attack in Algorithm 1 with the the listed parameters on the entire CIFAR-10 test set. In particular, the attacks use adversarial perturbation $\varepsilon = 8/255$ and attack step size $\alpha = 2/255$ in the l_∞ norm. One evaluation of the entire test set took approximately 2.5 days using 4x RTX 2070 Super GPUs. We compare our results to a representative selection of AT and AP defenses in Table 2. We include the training method for the classifier, transformation (if any), and the strongest attack for each defense.

Our EBM defense is stronger than standard AT (Madry et al., 2018) and comparable to modified AT such as (Zhang et al., 2019). Although our results are not on par with state-of-the-art AT (Carmon et al., 2019), our defense is the first method that can effectively secure naturally-trained classifiers.

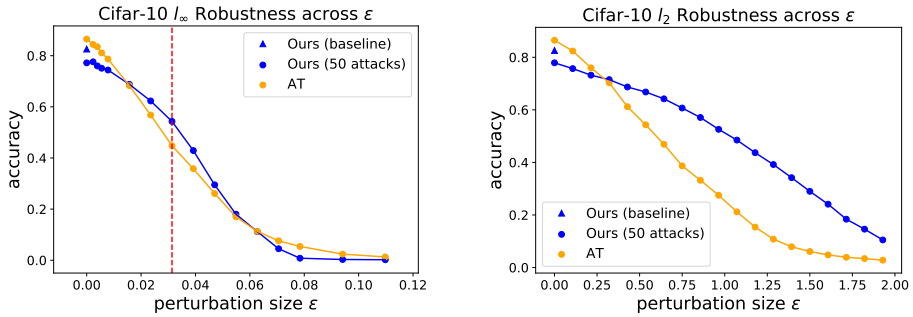

Figure 4: Accuracy across perturbation ε for l_∞ and l_2 attacks against our defense and AT (Madry et al., 2018).

We now examine the effect of the perturbation ε, number of attacks N, and number of EOT attack replicates H_{adv} on the strength of the BPDA+EOT attack. To reduce the computational cost of the diagnostics, we use a fixed set of 1000 randomly selected test images for diagnostic attacks.

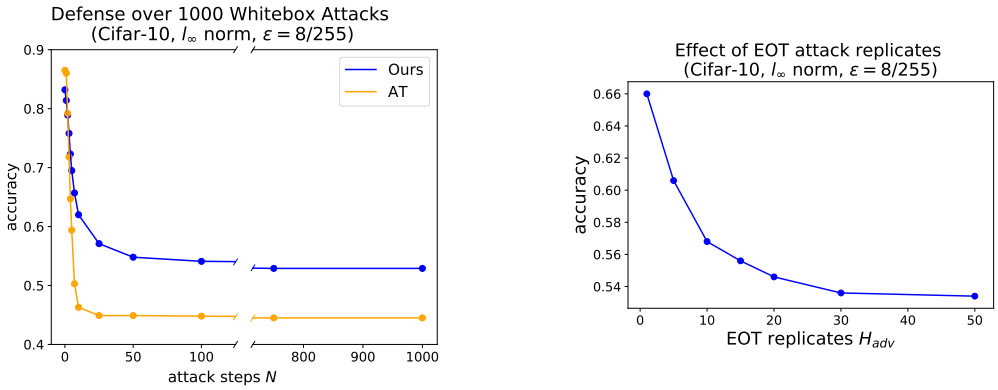

Figure 5: Effect of number of attack steps N and number of EOT replicates H_{adv}.

Figure 4 displays the robustness of our model and standard AT (Madry et al., 2018) for l_∞ and l_2 attacks across perturbation size ε. Our model is attacked with BPDA+EOT while the AT model is attacked with PGD. Our defense is more robust than AT for a range of medium-size perturbations in the l_∞ norm and much more robust than AT for medium and large ε in the l_2 norm.

Figure 5 visualizes the effect of increasing the computational power of the attacker. The left figure compares our defense with AT over 1000 attacks using an increased number $H_{\text{adv}} = 20$ of attack replicates. The majority of breaks happen within the first 50 attacks as used in the CIFAR-10 experiment in Table 2, while a small number of breaks occur within a few hundred attack steps. It is likely that some breaks from long-run attacks are the result of lingering stochastic behavior from $\hat{F}_{H_{\text{def}}}(x)$ rather than the attack itself. The right figure shows the effect of the number of EOT attack replicates over 50 attacks. The strength of the EOT attack saturates when using 20 to 30 replicates. A small gap in attack strength remains between the 15 replicates used in our attacks and the strongest possible attack. Some of this effect is likely mitigated by our use of 5 random restarts.

The diagnostics indicate that the report in Table 2 is a fair evaluation. Increasing H_{def} would secure some borderline images that break due to random effects while increasing N and H_{adv} would lead to stronger attacks. The evaluation in Table 2 pushes the limits of what is computationally feasible with widely available resources. Our defense report is an accurate approximation of the defense of our

ideal classifier $F(x)$ in (4) against the BPDA+EOT attack (10) although we acknowledge that more computation would yield a more accurate estimate.

SVHN and CIFAR-100. The attack and defense parameters for our method are identical to those used in the CIFAR-10 experiments. We compare our results with standard AT. Overall, our defense performs well for datasets that are both simpler and more complex than CIFAR-10. In future work, further stabilization of image appearance after Langevin sampling could yield significant benefits for settings where precise details need to be preserved for accurate classification. The AT results for CIFAR-100 are from (Balaji et al., 2019) and the results for SVHN are from our implementation.

Table 3: Defense vs. whitebox attacks with l_∞ perturbation $\varepsilon = 8/255$ for SVHN and CIFAR-100.

	SVHN		CIFAR-100	
	Nat.	Adv.	Nat.	Adv.
Ours	0.9223	0.6755	0.5166	0.2610
(Madry et al., 2018)	0.8957	0.5039	0.5958	0.2547

5 Related Work

Adversarial training learns a robust classifier using adversarial images created during each weight update. The method is introduced in (Madry et al., 2018). Many variations of AT have been explored, some of which are related to our defense. (He et al., 2019) applies noise injection to each network layer to increase robustness via stochastic effects. Similarly, Langevin updates with our EBM can be interpreted as a ResNet (He et al., 2016) with noise injected layers as discussed in (Nijkamp et al., 2019). Semi-supervised AT methods (Uesato et al., 2019; Carmon et al., 2019) use unlabeled images to improve robustness. Our EBM also leverages unlabeled data for defense.

Adversarial preprocessing is a strategy where auxiliary transformations are applied to adversarial inputs before they are given to the classifier. Prior forms of pre-processing defenses include rescaling (Xie et al., 2018), thermometer encoding (Buckman et al., 2018), feature squeezing (Xu et al., 2018), activation pruning (Dhillon et al., 2018), reconstruction (Samangouei et al., 2018), ensemble transformations (Raff et al., 2019), addition of Gaussian noise (Cohen et al., 2019), and reconstruction of masked images (Yang et al., 2019). Prior preprocessing methods also include energy-based models such as Pixel-Defend (Song et al., 2018) and MALADE (Srinivasan et al., 2019) that draw samples from a density that differs from (1). It was shown in (Athalye et al., 2018; Tramer et al., 2020) that many preprocessing defenses can be totally broken or dramatically weakened by simple adjustments to the standard PGD attack, namely the EOT and BPDA techniques. Furthermore, many preprocessing defenses such as (Cohen et al., 2019; Raff et al., 2019; Yang et al., 2019) also modify classifier training so that the resulting defenses are analogous to AT, as discussed in Appendix E. No prior preprocessing defense is competitive with AT when applied to a naturally-trained classifier.

Energy-based models are a probabilistic method for unsupervised modeling. Early energy-based image models include the FRAME (Zhu et al., 1998) and RBM (Hinton, 2002). The EBM (1) used in this work is introduced in (Xie et al., 2016) and important observations about the learning process are presented in (Nijkamp et al., 2020; 2019). Preliminary investigations for using the EBM (1) for adversarial defense are presented in (Du & Mordatch, 2019; Grathwohl et al., 2020) but the results are not competitive with AT (see Section 4.1). Our work builds on the convergent learning methodology in (Nijkamp et al., 2020) to apply long-run Langevin sampling as a defense technique.

6 Conclusion

This work demonstrates that Langevin sampling with a convergent EBM is an effective defense for naturally-trained classifiers. Our defense is founded on an improvement to EBM training that enables efficient learning of stable long-run sampling for complex datasets. We evaluate our defense using non-adaptive and adaptive whitebox attacks for the CIFAR-10, CIFAR-100, and SVHN datasets. The evaluations show that our defense is competitive with adversarial training. Securing naturally-trained classifiers with post-training defense is a long-standing open problem that our work resolves.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ICML*, 2018.

Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910:08051*, 2019.

Giancarlo Benettin, Luigi Galgani, and Jean-Marie Streclyn. Kolmogorov entropy and numerical experiments. *Physical Review A*, 14(6):2338–2345, 1976.

Anton Bovier and Frank den Hollander. Metastability: A potential theoretic approach. *International Congress of Mathematicians*, 3:499–518, 2006.

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. *ICLR*, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *NeurIPS*, 2019.

Jeremy M. Cohen, Elan Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *ICML*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*, 2020.

Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. *ICLR*, 2018.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *NeurIPS*, 2019.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, Kevin Swersky, and Mohammad Norouzi. Your classifier is secretly an energy-based model and you should treat it like one. *ICLR*, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *CVPR*, 2019.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Nitish Shirish Keskar and Richard Socher. Improving generalization by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Ying-Cheng Lai, Zonghua Liu, Lora Billiings, and Ira B. Schartz. Noise-induced unstable dimension variability and transition to chaos in random dynamical systems. *Physical Review E*, 67, 2003.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run MCMC toward energy-based model. *NeurIPS*, 2019.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of MCMC-based maximum likelihood learning of energy-based models. *AAAI*, 2020.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. *CVPR*, 2019.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ICLR*, 2018.

Vignesh Srinivasan, Arturo Marban, Klaus-Robert Muller, Wojciech Samek, and Shinichi Nakajima. Defense against adversarial attacks by langevin dynamics. *ICML*, 2019.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002:08347*, 2020.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *NeurIPS*, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *ICLR*, 2018.

Jianwen Xie, Yang Lu, Song-Chun Zhu Zhu, and Ying Nian Wu. A theory of generative convnet. *ICML*, 2016.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *NDSS*, 2018.

Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *ICML*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled tradeoff between robustness and accuracy. *ICML*, 2019.

Song Chun Zhu, Ying Nian Wu, and David Mumford. Filters, random fields and maximum entropy (FRAME): towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

A ATTACK ALGORITHM

Algorithm 1 BPDA+EOT adaptive attack to evaluate EOT defense (4)

Require: Natural images $\{x_m^+\}_{m=1}^M$, EBM $U(x)$, classifier $f(x)$, Langevin noise $\tau = 0.01$, Langevin updates $K = 1500$, number of attacks $N = 50$, attack step size $\alpha = \frac{2}{255}$, maximum perturbation size $\varepsilon = \frac{8}{255}$, EOT attack samples $H_{\text{adv}} = 15$, EOT defense samples $H_{\text{def}} = 150$

Ensure: Defense record $\{d_m\}_{m=1}^M$ for each image.

```
for m=1:M do
    Calculate large-sample predicted label of the natural image ĉ_{H_def}(x_m^+) with (6).
    if ĉ_{H_def}(x_m^+) ≠ y_m then
        Natural image misclassified. d_m ← False. End loop iteration m.
    else
        d_m ← True.
    end if
    Randomly initialize X_0 in the l_p ε-ball centered at x_m^+ and project to [0,1]^D.
    for n=1:(N+1) do
        Calculate small-sample predicted label ĉ_{H_adv}(X_{n-1}) with (6).
        Calculated attack gradient Δ_{BPDA+EOT}(X_{n-1}, y_m) with (10).
        if ĉ_{H_adv}(X_{n-1}) ≠ y_m then
            Calculate large-sample predicted label ĉ_{H_def}(X_{n-1}) with (6).
            if ĉ_{H_def}(X_{n-1}) ≠ y_m then
                The attack has succeeded. d_m ← False. End loop iteration m.
            end if
        end if
        Use Δ_{BPDA+EOT}(X_{n-1}, y_m) with the l_p ε-bounded PGD update (8) to obtain X_n.
    end for
end for
```

Algorithm 1 is pseudo-code the attack and defense framework described in Section 3. One notable aspect of the algorithm is the inclusion of an EOT defense phase to verify potentially successful attacks. Images which are identified as broken for the smaller sample that is used to generate EOT attack gradients are checked again using a much large EOT defense sample to ensure that the break is due to the adversarial state and not random finite-sample effects. This division is done for purely computational reasons. It is extremely expensive to use 150 EOT attack replicates but much less expensive to use 15 EOT attack replicates as a screening method and to carefully check candidates for breaks when they are identified from time to time using 150 EOT defense replicates. In our experiments we find that the EOT attack is close to its maximum strength after about 15 to 20 replicates are used, while about 150 EOT defense replicates are needed for consistent prediction of F over several hundred attacks. Ideally, the same number of chains should be used for both EOT attack and defense, in which case the separate verification phase would not be necessary.

B IMPROVED LEARNING OF CONVERGENT EBMS

Algorithm 2 gives pseudo-code for our improvement of the convergent learning method in (Nijkamp et al., 2020). This implementation allows efficient learning of convergent EBM's for complex datasets.

C ERASING ADVERSARIAL SIGNALS WITH MCMC SAMPLING

This section discusses two theoretical perspectives that justify the use of an EBM for purifying adversarial signals: memoryless and chaotic behaviors from sampling dynamics. We emphasize that the discussion applies primarily to long-run behavior of a Langevin image trajectory. Memoryless and chaotic properties do not appear to emerge from short-run evolution. Throughout our experiments, we never observe significant defense benefits from short-run Langevin sampling.

Algorithm 2 ML with Adam to SGD Switch for Convergent Learning of EBM (1)

Require: ConvNet potential $U(x;\theta)$, number of training steps J, step to switch from SGD to Adam J_{SGD}, initial weight θ_1, training images $\{x_i^+\}_{i=1}^{N_{\text{data}}}$, data perturbation $\tau_{\text{data}} = 0.02$, step size $\tau = 0.01$, Langevin steps $K = 100$, Adam learning rate $\gamma_{\text{Adam}} = 0.0001$, SGD learning rate $\gamma_{\text{SGD}} = 0.00005$.

Ensure: Weights θ_{J+1} for energy $U(x;\theta)$.

Set optimizer $g = \text{Adam}(\gamma_{\text{Adam}})$. Initialize persistent image bank as N_{data} uniform noise images.

for j=1:(J+1) **do**

  if $j = J_{\text{SGD}}$ **then**

    Set optimizer $g \leftarrow \text{SGD}(\gamma_{\text{SGD}})$.

  end if

1. Draw batch images $\{x_{(i)}^+\}_{i=1}^m$ from training set, where (i) indicates a randomly selected index for sample i, and get samples $X_i^+ = x_{(i)} + \tau_{\text{data}} Z_i$, where $Z_i \sim \text{N}(0, I_D)$ i.i.d.
2. Draw initial negative samples $\{Y_i^{(0)}\}_{i=1}^m$ from persistent image bank. Update $\{Y_i^{(0)}\}_{i=1}^m$ with the Langevin equation

$$Y_i^{(k)} = Y_i^{(k-1)} - \frac{\tau^2}{2}\frac{\partial}{\partial y} U(Y_i^{(k-1)};\theta_j) + \tau Z_{i,k},$$

   where $Z_{i,k} \sim \text{N}(0, I_D)$ i.i.d., for K steps to obtain samples $\{X_i^-\}_{i=1}^m = \{Y_i^{(K)}\}_{i=1}^m$. Update persistent image bank with images $\{Y_i^{(K)}\}_{i=1}^m$.
3. Update the weights by $\theta_{j+1} = \theta_j - g(\Delta\theta_j)$, where g is the optimizer and

$$\Delta\theta_j = \frac{\partial}{\partial\theta}\left(\frac{1}{n}\sum_{i=1}^n U(X_i^+;\theta_j) - \frac{1}{m}\sum_{i=1}^m U(X_i^-;\theta_j)\right)$$

   is the ML gradient approximation.

end for

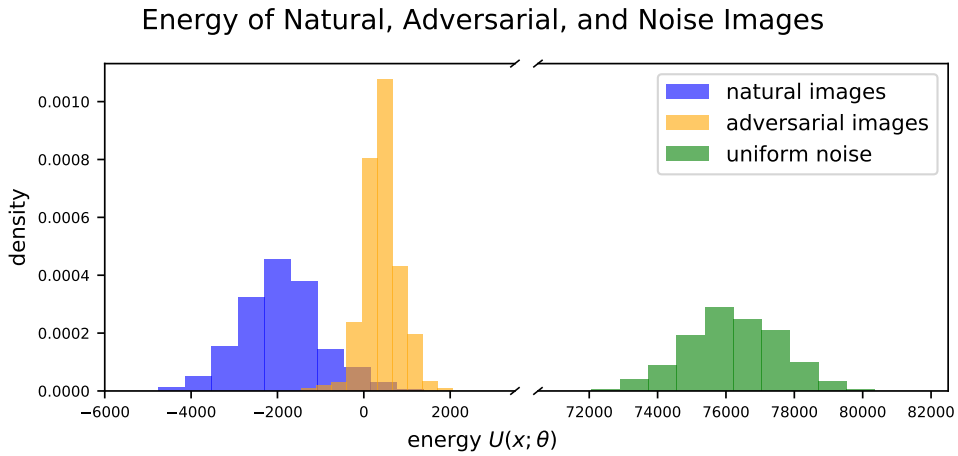

Figure 6: Energy $U(x;\theta)$ of natural, adversarial, and noise images.

C.1 Memoryless Dynamics

The first justification of EBM defense is that iterative probabilistic updates will move an image from a low-probability adversarial region to a high-probability natural region, as previously discussed in (Song et al., 2018; Srinivasan et al., 2019; Grathwohl et al., 2020). Comparing the energy of adversarial and natural images shows that adversarial images tend to have somewhat higher energy,

which is evidence that adversarial images are improbable deviations from the EBM manifold learned from natural images (see Figure 6).

The theoretical foundation for removing adversarial signals with MCMC sampling comes from the well-known steady-state convergence property of Markov chains. The Langevin update (3) is designed to converge to the distribution $p(x; \theta)$ learned from unlabeled data after an infinite number of steps. The steady-state property guarantees that adversarial signals will be erase from the sampled state as long as enough steps are used. This property is actually too extreme because full MCMC mixing would completely undermine classification by causing samples to jump between class modes.

Fortunately the quasi-equilibrium and metastable properties of MCMC sampling (Bovier & den Hollander, 2006) can be as useful as its equilibrium properties. Although slow-mixing and high autocorrelation of MCMC chains are often viewed as a major shortcoming, these properties enable defense by preserving class appearance while sampling erases of adversarial signals. Successful classification of long-run MCMC samples occurs when the metastable regions of the EBM $p(x; \theta)$ are aligned with the label information learned by the classifier network $f(x)$. Our method relies on a balance between the memoryless properties of MCMC sampling that erase noise and the metastable properties of MCMC sampling that preserve the initial state.

C.2 Chaotic Dynamics

Chaos theory gives another perspective for justifying the erasure of adversarial signals using long-run iterative transformations. Intuitively, a deterministic system is chaotic if an initial infinitesimal perturbation grows exponentially in time so that paths of nearby points become distant as the system evolves. The same concept can be extended to stochastic systems. The SDE

$$\frac{dX}{dt} = V(X) + \eta\xi(t), \tag{11}$$

where $\xi(t)$ is Brownian motion and $\eta \geq 0$, that encompasses the Langevin equation is known to exhibit chaotic behavior in many contexts for sufficiently large η (Lai et al., 2003).

One can determine whether a dynamical system is chaotic or ordered by measuring the maximal Lyapunov exponent of the system. Order systems have a maximal Lyapunov exponent that is either negative or 0, while chaotic systems have positive Lyapunov exponents. The SDE (11) will have a maximal exponent of at least 0 since dynamics in the direction of gradient flow are neither expanding or contracting. One can therefore detect whether a Langevin equation yields ordered or chaotic dynamics by examining whether its corresponding maximal Lyapunov exponent is 0 or positive.

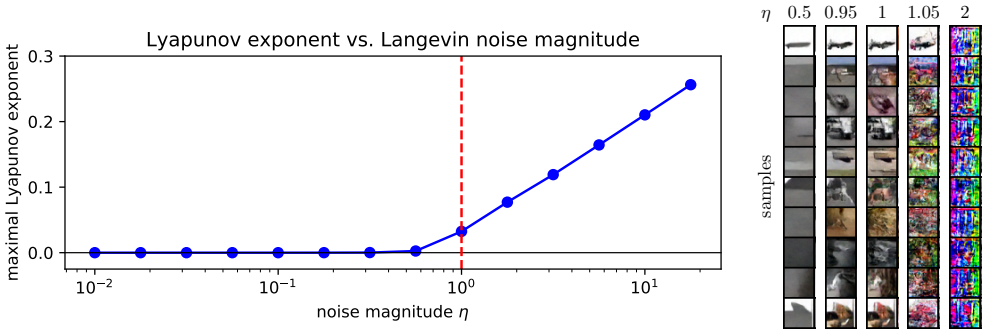

Figure 7: *Left:* Maximal Lyapunov exponent for different values of η. The value $\eta = 1$ which corresponds to our training and defense sampling dynamics is just above the transition from the ordered region where the maximal exponent is 0 to the chaotic region that where the maximal exponent is positive. *Right:* Appearance of steady-state samples for different values of η. Oversaturated images appear for low values of η, while noisy images appear for high η. Realistic synthesis is achieve in a small window around $\eta = 1$ where gradient and noise forces are evenly balanced.

We use the classical method (Benettin et al., 1976) to calculate the maximal Lyapunov exponent of the altered form Langevin transformation (3) given by

$$T_\eta^*(X) = X - \frac{\tau^2}{2} \nabla_X U(X;\theta) + \eta\tau Z_k$$

for a variety of noise strengths η. Our results exhibit the predicted transition from noise to chaos. The value $\eta = 1$ corresponding to our training and defense algorithms is just beyond the transition from the ordered region to the chaotic region. Our defense dynamics occur in a critical interval where stable forces promoting pattern formation (but also oversaturation) and noise forces that disrupt pattern formation are evenly balanced. The results are shown in Figure 7. We believe that the unpredictability of paths under T^* is an effective defense against BPDA because informative attack gradients cannot be generated through chaotic transformation. Other chaotic transformations, either or stochastic or deterministic, might be an interesting line of research as a class of defense methods.

D Effect of Number of Langevin Steps and Comparison of Convergent and Non-Convergent Defenses

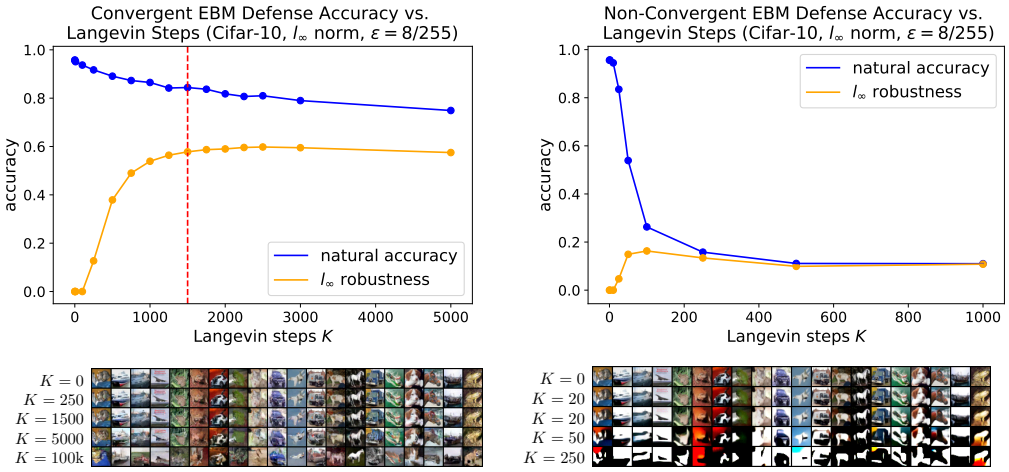

Figure 8: Accuracy on natural and adversarial images created from the BPDA+EOT attack (10) for EBM defense with different number of Langevin steps, and images sampled from the EBM. *Left:* Defense with a convergent EBM. Using approximately 1500 Langevin steps yields a good balance of natural and robust accuracy. *Right:* Defense with non-convergent EBM. Oversaturated long-run images prevent non-convergent EBM defense from achieving high natural or robust accuracy.

In this appendix, we examine the effect of the number of Langevin steps on defense accuracy for the CIFAR-10 dataset (see Figure 8). Each point displays either the baseline accuracy of our stochastic classifier (orange) or the the results of a BPDA+EOT attack (blue) on 1000 test images. The attacks used to make this diagram used a reduced load of $H_{\text{adv}} = 7$ replicates for EOT attacks so the defense accuracies are slightly higher than the full attack results presented later. Short-run Langevin with $K \leq 100$ steps yields almost no adversarial robustness. Increasing the number of steps gradually increases robustness until the defense saturates at around $K = 2000$. We chose $K = 1500$ steps in our experiments as a good tradeoff between robustness, natural accuracy, and computational cost.

For comparison, we run the same experiment using a non-convergent EBM. The network structure and training are identical to our convergent model, except that the SGD phase is removed and Adam is used throughout training. The non-convergent EBM defense cannot achieve high natural accuracy with long-run sampling because of the oversaturated features that emerge. Without a high natural accuracy, it is impossible to obtain good defense results. Convergent EBM's that can produce realistic long-run samples are a key ingredient for the success of our method.

We can further examine the difference between convergent and non-convergent EBM defenses by measuring the FID score of samples initialized from the data (see Figure 9). The convergent EBM

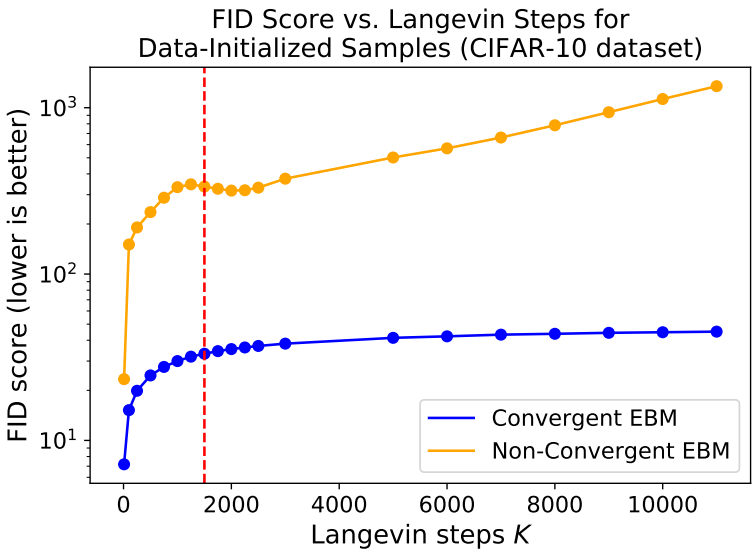

Figure 9: FID scores of samples from a convergent and non-convergent EBM that are initialized from training data. The FID of the convergent model remains reasonably low because the steady-state of the convergent model is aligned with the data distribution. The FID of the non-convergent model quickly increases because the oversaturated steady-state of the non-convergent model differs significantly from the data distribution. Maintaining a sampling distribution close to the data distribution is essential for achieving high prediction accuracy of a natural classifier on transformed states.

maintains a reasonably low FID score across many steps so that long-run samples can be accurately classified. The non-convergent EBM experiences a quick increase in FID as oversaturated samples appear. Labels for the oversaturated samples cannot be accurately predicted by a naturally-trained classifier, preventing successful defense.

E A Note on Modified Classifier Training for Preprocessing Defenses

Many works that report significant robustness via defensive transformations (Cohen et al. (2019); Raff et al. (2019); Yang et al. (2019) and others) also modify classifier learning by training with transformed images rather than natural images. Prior defensive transformations that are strong enough to remove adversarial signals have the side effect of greatly reducing the accuracy of a naturally trained classifier. Therefore, the signals introduced by these defensive transformations (high Gaussian noise, ensemble transformations, ME reconstruction, etc.) can be also considered "adversarial" signals (albeit perceptible ones) because they heavily degrade natural classifier accuracy much like the attack signals. From this perspective, modifying training to classify transformed images is a direct analog of AT where the classifier is trained to be robust to "adversarial" signals from the defensive transformation itself rather than a PGD attack. Across many adversarial defense works, identifying a defensive transformation that removes attack signals while preserving high accuracy of a natural classifier has universally proven elusive.

Our experiments train the classifier with natural images alone and not with images generated from our defensive transformation of Langevin sampling. Our approach represents the first successful defense based purely on transformation and validates an entirely different approach compared to defenses which modify training of the base classifier (Madry et al. (2018); Cohen et al. (2019); Raff et al. (2019); Yang et al. (2019) etc.). To our knowledge, showing that natural classifiers can be secured with post-training defensive transformation is a contribution that is unique in the literature. Our approach has the potential of securing images for many applications using a single defense model, while AT and relatives must learn a robust model for each application.

F DISCUSSION OF IGEBM AND JEM DEFENSES

We hypothesize that the non-convergent behavior of the IGEBM (Du & Mordatch, 2019) and JEM (Grathwohl et al., 2020) models limits their use as an EBM defense method. Long-run samples from both models have oversaturated and unrealistic appearance (see Figure 10). Non-convergent learning problems are caused by training implementation rather than model formulation. Convergent learning may be a path to robustness using the IGEBM and JEM defense methods.

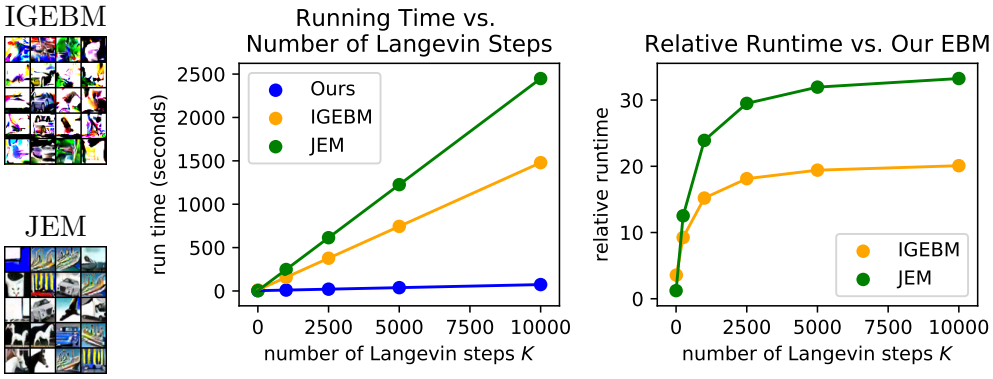

Figure 10: *Left:* Approximate steady-state samples of the IGEM and JEM models. Both exhibit oversaturation from non-convergent learning that can interfere with defense capabilities. *Right:* Comparison of running time for Langevin sampling with a batch of 100 images. The small scale and fast sampling of our EBM are important for the computational feasibility of our defense.

Both prior works use very large networks to maximize scores on generative modeling metrics. As a result, sampling from these models can be up to 30 times slower than sampling from our lightweight EBM structure from (Nijkamp et al., 2019) (see Figure 10). The computational feasibility of our method currently relies on the the small scale of our EBM. Given the effectiveness of the weaker and less expensive PGD attack in Section 4.1 and the extreme computational cost of sampling with large EBM models, we do not to apply BPDA+EOT to the IGEBM or JEM defense.

The original evaluations of the IGEBM and JEM model use end-to-end backpropagation through the Langevin dynamics when generating adversarial examples. On the other hand, the relatively weak attack in Section 4.1 is as strong or much stronger than the theoretically ideal end-to-end attack. Gradient obfuscation from complex second-order differentiation might hinder the strength of end-to-end PGD when attacking Langevin defenses.

The IGEBM defense overcomes oversaturation by restricting sampling to a ball around the input image, but this likely prevents sampling from being able to manifest its defensive properties. An adversarial signal will be partially preserved by the boundaries of the ball regardless of how many sampling steps are used. Unrestricted sampling, as performed in our work and the JEM defense, is essential for removing adversarial signals.