Stochastic Security: Adversarial Defense Using Long-Run Dynamics of Energy-Based Models

Mitch Hill*
University of Central Florida
University of California, Los Angeles
mitchell.hill@ucf.edu

Jonathan Mitchell*
University of California, Los Angeles
jcmitchell@ucla.edu

Song-Chun Zhu
University of California, Los Angeles
sczhu@stat.ucla.edu

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 pre-existing and highly vulnerable classifiers.

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, 2) an Expectation-Over-Transformation (EOT) defense that resolves theoretical ambiguities for 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 adversarially-trained classifiers on Cifar-10, SVHN, and Cifar-100. Code and pre-trained models are available at `https://github.com/point0bar1/ebm-defense`.

1 Motivation and Contributions

The outputs of deep networks are known to be very sensitive to small perturbations to the input. 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 a human [10]. 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. Untargeted attacks seek to cause incorrect prediction without preference among alternatives to the correct label. The effectiveness of white-box attacks raises important questions about the nature of the cognitive abilities of deep networks and the security risks posed by deployment of deep networks in everyday life.

Preprint. Under review.
*equal contributions

Many methods have been introduced to create adversarial examples. Strong iterative attacks, such as Projected Gradient Descent (PGD) [19], are capable of reducing the accuracy of a standard 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) [19]. Another defense strategy, which we will refer to as *adversarial purification*, uses iterative refinement to purify an image and remove adversarial signals before classifying the purified samples [23, 24, 30]. The trajectory in the image space that results from iterative refinement distinguishes adversarial purification from defenses that pre-process before classification using a one-time treatment such as rotation [12], discretization [5], or reconstruction [22]. Adversarial purification is an attractive defense strategy compared with adversarial training because it could be used to secure existing and naturally-trained classifiers. While previous investigations into adversarial purification often modify the training of the classifier network in addition to performing purification, in this work we focus on defending classifiers that are trained with natural images and no learning modification.

Markov Chain Monte Carlo (MCMC) sampling using an Energy-Based Model (EBM) with a ConvNet potential [28] has recently emerged as a method for adversarial purification [9, 11]. However, the proposed defenses are not competitive with adversarial training (see Table 1 and [7]). In the present work we demonstrate that EBM defense of a naturally-trained classifier can be significantly stronger than standard adversarial training in [19] and competitive compared to modified adversarial training such as [32, 6].

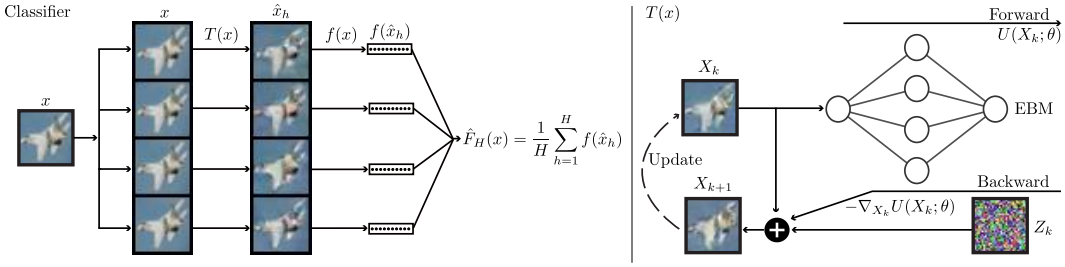

Figure 1: *Left:* Visualization of calculating our stochastic logits $\hat{F}_H(x)$ from (8). The input image x is replicated H times and parallel Langevin updates with a ConvNet EBM are performed on each replicate to generate $\{\hat{x}_h\}_{h=1}^H$. Purified samples are sent in parallel to our naturally-trained 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 (7) 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.

It is essential to properly incorporate the purification process when designing white-box attacks to reliably evaluate the success of adversarial purification. The work [1] revealed that many purification defenses can be overcome with minor adjustments to the standard PGD attack. Both stochastic behavior from purification and the computational difficulty of end-to-end backpropagation can be circumvented to attack the classifier through the refinement process. In this paper we carefully address [1] to evaluate adversarial purification with an EBM using attacks with the greatest known efficiency and effectiveness.

Despite the weakness of adversarial purification to attack strategies from [1], we believe that EBM purification can be a solid defense method. Our defense tools are a classifier trained with labeled natural images and EBM trained with unlabeled natural images from the dataset. For prediction, we perform MCMC 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. The stochastic dynamics of long-run MCMC sampling constitute a memoryless and chaotic trajectory that removes adversarial signals, while metastable sampling behaviors preserve image classes over long-run trajectories. Balancing these factors can lead to effective adversarial defense. Realistic long-run sampling for a ConvNet EBM has only recently been introduced [20] and our work builds on these observations to provide the first analysis of adversarial purification with long-run MCMC. We find that long-run sampling as opposed to short-run sampling is essential for the success of purification defense with an EBM. Our main contributions are:

- A simple but effective adjustment to improve the convergent learning procedure in [20]. 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 well-known EOT attack [1] 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 adversarial training.

2 Improved Convergent Learning of Energy-Based Models

The Energy-Based Model introduced in [28] 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. The Langevin update

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

where $Z_k \sim \mathrm{N}(0, I_D)$ and $\tau > 0$, is often used to obtain the samples $\{X_i^-\}_{i=1}^m$. A full Langevin implementation requires an additional momentum and Metropolis-Hastings step but these are often not used and are not needed for small τ as the Metropolis-Hastings acceptance approaches 1.

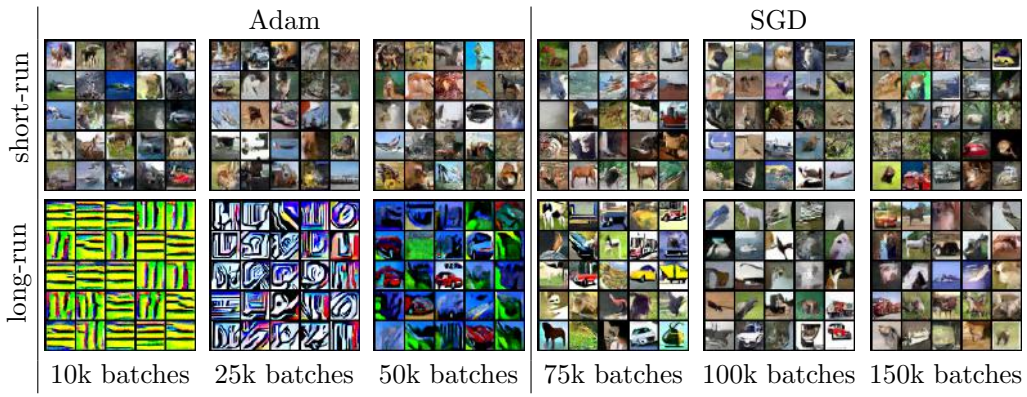

Figure 2: Comparison of long-run and short-run samples over model updates for our improved method of convergent learning. The model is updated in a non-convergent learning phase with the Adam optimizer for the first 50,000 batches. The majority of short-run synthesis realism is learned during this phase, 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 is gradually aligned with the short-run distribution.

The recent work [20] 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. Our purification strategy depends on the realism of long-run samples and successful convergent learning is crucial for our approach (see Figure 4). Although the work presents preliminary results for convergent learning of image potentials, the experiments are limited and computationally intensive.

We now present a simple but effective adjustment for improving convergent learning. As observed in [20], we were unable to learn a convergent model when updating θ using the Adam optimizer [17]. One possible explanation is that the invariance of the Adam update with respect to the scale of the gradient obfuscates the temperature of the underlying model. Despite the drawbacks of Adam for convergent learning, it is a very effective tool for non-convergent learning with short-run synthesis. Drawing inspiration from classifier training in [16], we propose to learn a convergent EBM by using Adam for weight updates early in training to achieve realistic short-run synthesis and use SGD later in training to align short-run and long-run MCMC samples to correct the degenerate steady-state from the Adam phase. This allows us to learn the convergent EBM's for complex datasets such as Cifar-10 using a budget of 100 MCMC steps per network weight update as opposed to the 500 MCMC steps used in [20]. We use the light-weight EBM from [21] as our network architecture. See Figure 2 for an illustration and Appendix A training details. Despite its simplicity, this learning modification provides a crucial foundation for efficient EBM learning with complex datasets to enable defense applications.

3 Attack and Defense Formulation

This section begins with a brief review of the norm-bounded attack framework and the Projected Gradient Descent (PGD) [19] attack that is a standard benchmark for deterministic defenses. We then present a framework for evaluating stochastic defenses. In particular, we resolve ambiguities in prior evaluations of stochastic defenses by introducing a deterministic objective that removes the possibility of a stochastic defense breaking due to stochasticity rather than the attack. Our notation and attack methodology largely follow [1, 25].

3.1 PGD Attack

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

$$x_{\text{adv}}(x^+, y) = \arg\max_{x \in S} L(F(x), y) \tag{4}$$

that maximizes loss for predicting y in a set $S \subset \mathbb{R}^D$ centered around x^+. In this work, natural images x^+ will have pixels intensities from 0 to 1 (i.e. $x^+ \in [0,1]^D$). Typically one expects that the visual appearance of x differs only slightly from the original image x^+ for all $x \in S$. 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 adversarial attack is successful if it finds a perturbed image $x^* \in S$ that causes a mislabeling $c(x^*) \neq y$. For consistency of notation, a natural image misclassification $c(x^+) \neq y$ is considered a trivial successful attack.

The Projected Gradient Descent (PGD) attack [19] 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 (4) by iteratively updating x_i with

$$x_{i+1} = \prod_S (x_i + \alpha g(x_i, y)), \quad g(x, y) = \arg\max_{\|v\|_p \leq 1} v^\top \Delta(x, y), \tag{5}$$

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

$$\Delta_{\text{PGD}}(x, y) = \nabla_{x_i} L(F(x_i), y). \tag{6}$$

Intuitively, the attack alternates between ascending along L in the steepest direction of the l_p metric then projecting the resulting state back into S so that the perturbation remains imperceptible. The attack gradient is $g(x, y) = \text{sign}\Delta(x, y)$ for an l_∞ attack and $g(x, y) = \Delta(x, y)/\|\Delta(x, y)\|_2$ for an l_2 attack.

3.2 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 $\mathbb{R}^D$. Stochastic transformations encompass deterministic transformations for which $T(x)$ is a constant random variable. One can compose $T(x)$ with a deterministic classifier $f(x) \in \mathbb{R}^J$ (in our case, a naturally-trained classifier network) to define a new classifier $F(x) \in \mathbb{R}^J$ as

$$F(x) = E_{T(x)}[f(T(x))]. \tag{7}$$

We emphasize that $F(x)$ is a deterministic classifier even though $T(x)$ is stochastic. Additionally, the classifier $F(x)$ will differ depending on whether $f(x)$ gives logit, softmax, or log softmax output of the existing network. In our experiments we did not observe significant differences between these alternatives and throughout this paper $f(x)$ will denote logits and $F(x)$ will denote expected logits. We refer to (7) as an Expectation-Over-Transformation (EOT) defense. The classifier $F(x)$ in (7) is simply the target of the EOT attack [1]. 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

$$\hat{F}_H(x) = \frac{1}{H} \sum_{h=1}^{H} f(\hat{x}_h) \quad \text{where,} \quad \hat{x}_h \sim T(x) \text{ i.i.d.}, \tag{8}$$

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.

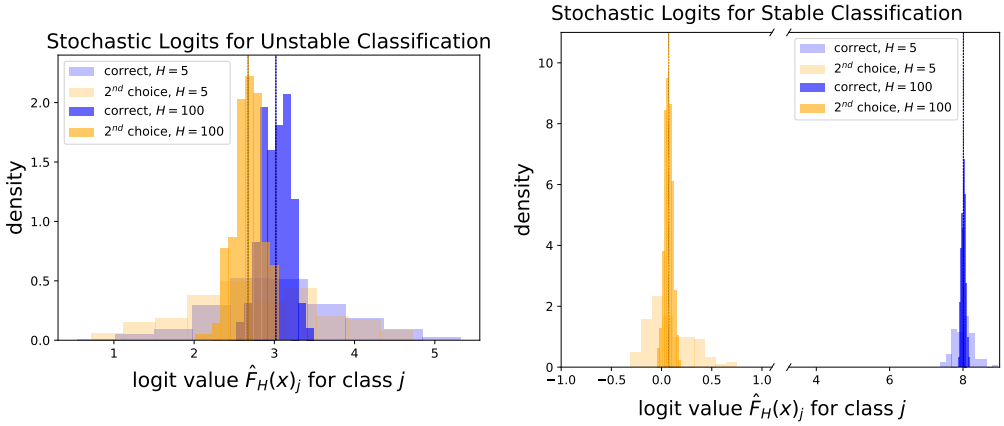

Figure 3: Demonstration of importance of EOT defense. The histograms display different realizations of the logits $\hat{F}_H(x)$ for the correct class and the second most probable class for images x_1 (*left*) and x_2 (*right*) and different choices of H. In both cases, $F(x)$ (given approximately by the dashed vertical lines) gives correct classification. However, the overlap between the logit histograms of $\hat{F}_H(x_1)$ indicate a relatively high probability of misclassification even for large H, while $\hat{F}_H(x_2)$ gives correct prediction virtually always even for small H because the histograms are well-separated. The EOT defense formulation (7) is essential for securing borderline images such as x_1.

In the literature it appears that both attackers [1, 25] and defenders [23, 24, 30, 11] 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 x could be identified as an adversarial image from randomness alone from repeated attacks even when $\hat{F}_H(x)$ gives the correct prediction on average. In experiments with EBM purification, we identify many images x that exhibit significant variation in the predicted label

$$\hat{c}_H(x) = \arg\max_j \hat{F}_H(x)_j \tag{9}$$

for smaller $H \approx 10$ but which have consistently correct prediction for larger $H \approx 150$. We strongly believe that fair evaluation of stochastic defenses must be based on the deterministic EOT defense F in (7) and that attackers must use sufficiently large H to ensure that $\hat{F}_H(x)$ 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.

We refer to the use of a transformation $T(x)$ before classification to improve robustness as *adversarial pre-processing*. A distinct subgroup of adversarial pre-processing methods are those which iteratively apply a transformation T^*:

$$T(x) = X_K, \quad X_k = T^*(X_{k-1}), \quad X_0 = x, \quad k = 1, \dots, K. \tag{10}$$

We refer to the use of iterative transformations to improve robustness as *adversarial purification*. This distinction is of interest because certain properties of an image trajectory $\{X_k\}_{k=0}^K$ can be used as a theoretical justification of defense. The stochastic transformation $T^*(x)$ studied in this work is a Langevin update (3) with an energy function $U(x)$ trained on unlabeled images from the same dataset that is used to train the deterministic classifier $f(x)$. The properties of this transformation that relate to adversarial defense are discussed in Section B.

3.3 Attacking Stochastic and Non-Differentiable Classifiers

This section discusses strategies for modifying the PGD attack (5) to address complexities caused by the stochastic transformation $T(x)$ in the EOT defense (7). The first strategy, known as the EOT attack [1], circumvents the intractability of F by attacking the finite sample logits $\hat{F}_{H_{\text{adv}}}$, where H_{adv} is the number of EOT attack samples, with the gradient

$$\Delta_{\text{EOT}}(x, y) = \nabla_x L(\hat{F}_{H_{\text{adv}}}(x), y). \tag{11}$$

This approach is very effective for attacking stochastic defenses. The EOT attack is the natural adaptive attack for our EOT defense formulation. In particular, both rely on $\hat{F}_H(x)$ to approximate $F(x)$. We note that EOT attacks in other works sometimes use different averaging methods (e.g. averaging the loss rather than the logits). However, our EOT defense explicitly uses the average logits for prediction and aligning the EOT attack averaging method with the EOT defense averaging method appears to produce the best results.

The second challenge of attacking a purification defense is the computational infeasibility or theoretical impossibility of differentiating $T(x)$. For the moment suppose that $T(x)$ is deterministic. The Backward Pass Differentiable Approximation (BPDA) technique [1] 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 that we use in this paper is the identity approximation $g(x) = x$. This approximation is reasonable for purification 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.

Given the noise $\{Z_k\}_{k=1}^K$, one could differentiate through the Langevin transformation (3) to the original state x so that BPDA would not be necessary. Standard automatic differentiation software can calculate the second-order information needed to differentiate through the gradient flow. While this approach is used to attack EBM defenses in [9, 11], it is computationally infeasible when more than around $K \approx 10$ Langevin steps are used for purification due to the extreme computational cost of second-order differentiation. BPDA allows us to evaluate purification defenses that use approximately $K = 1500$ Langevin steps, as needed for effective purification. Interestingly, a simple PGD transfer attack created from the classifier is as effective or more effective for attacking prior EBM defenses than the end-to-end attack through the Langevin updates used in the original evaluations (see Section 4.1 and [7]). Although the end-to-end attack is theoretically stronger, in practice the difficulty of second-order computation appears to significantly obfuscate attack gradients. BPDA, or even simpler attacks, are often very effective and essential for efficient evaluation of purification defenses.

We now bring together the strategies for dealing with the stochasticity and non-differentiability of $T(x)$. Combining the EOT attack and BPDA attack with identity $g(x) = x$ gives the attack gradient

$$\Delta_{\text{BPDA+EOT}}(x, y) = \frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} \nabla_{\hat{x}_h} L \left(\frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} f(\hat{x}_h), y \right), \quad \hat{x}_h \sim T(x) \text{ i.i.d.} \tag{12}$$

which we believe is the adaptive attack most naturally suited for our defense. To our knowledge, the BPDA+EOT attack represents the strongest known attack against purification defense, as demonstrated by its effectiveness in recent works such as [24, 25]. While there are many small variations in how one could define an BPDA+EOT attack (most notably the choice of EOT averaging method), our intention is to align our attack as closely as possible with our defense. We use $\Delta_{\text{BPDA+EOT}}(x, y)$ in (5) as our primary attack to evaluate the EOT defense (7).

3.4 Attack and Defense Algorithm

Algorithm 1 BPDA+EOT adaptive attack to evaluate EOT defense (7)

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 (9).
    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 (9).
        Calculated attack gradient Δ_{BPDA+EOT}(X_{n-1}, y_m) with (12).
        if ĉ_{H_adv}(X_{n-1}) ≠ y_m then
            Calculate large-sample predicted label ĉ_{H_def}(X_{n-1}) with (9).
            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 (5) to obtain X_n.
    end for
end for
```

Algorithm 1 summarizes the attack and defense framework described in this section. 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 achieves 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.

3.5 Effect of Number of Langevin Steps on Defense

In this section, we examine the effect of the number of Langevin steps on defense accuracy for the Cifar-10 dataset (see Figure 4). The classifier is trained using natural labeled images and the EBM

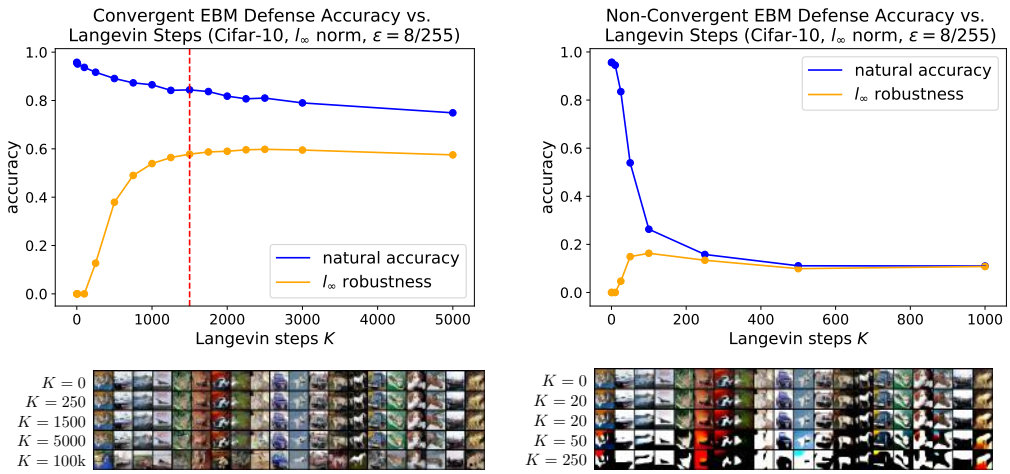

Figure 4: Accuracy on natural and adversarial images created from the BPDA+EOT attack (12) for EBM defense with different number of Langevin steps, and images sampled from the EBM. *Left:* Defense with a convergent EBM. Using approximately 1500 Langevin steps yields a good balance of natural and robust accuracy. *Right:* Defense with non-convergent EBM. Oversaturated long-run images prevent non-convergent EBM defense from achieving high natural or robust accuracy.

is trained using natural unlabeled images. Each point displays either the baseline accuracy of our stochastic classifier or the the results of a BPDA+EOT attack 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.

4 Experiments

We use two different network structures in our experiments. The first network is the lightweight EBM from [21]. The small scale of the EBM is essential for our experiments because many MCMC steps must be used before purification is effective as a defense and the cost quickly becomes prohibitive for large models. Our second network is WideResNet [31] classifier with depth 28 and width 10. The EBM and classifier are trained independently on the same dataset. The EBM is learned in the standard unconditional ML framework from unlabeled natural images while the classifier is trained to minimize cross-entropy loss on labeled natural images. No form of adversarial training is used for either model. We use the parameters from Algorithm 1 for all evaluations unless otherwise noted. Code and pre-trained models are available at `https://github.com/point0bar1/ebm-defense`.

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 should be relatively easy to secure with purification. We use this attack as a benchmark for comparing our defense to prior methods [9, 11] that evaluate adversarial defense with a ConvNet EBM (1). Both prior works use a conditional EBM for both sampling and classification rather than a separately trained unconditional EBM and classifier. The IGEBM defense [9] performs sampling on

each conditional model, while the JEM defense [11] performs sampling with an unconditional EBM that is derived from the conditional EBM. Furthermore, the IGEBM defense restricts sampling to a small l_∞ ball around the input image. For all methods, we evaluate the base classifier and the EBM defense for 10 Langevin steps (as in prior defenses) and 1500 steps (as in our defense). The results are displayed in Table 1.

	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
IGEBM [9]	0.4714	0.3219	0.4885	0.3674	0.487*	0.375*
JEM [11]	0.9282	0.0929	0.9093	0.1255	0.755*	0.238*

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*)

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. Our results are consistent with the original IGEBM evaluation, although we further observe that the majority of robustness comes from the base EBM classifier and not purification. Our evaluation of the JEM model shows significantly less defense that the original evaluation, and our results are consistent with the JEM evaluation in [7] that uses the same attack methodology as we do. Further discussion of the IGEBM and JEM defenses is included in Appendix C. Throughout our experiments, we never observe any significant defense benefits from short-run Langevin and we were only able to implement successful long-run Langevin defense with convergent EBM's.

4.2 BPDA+EOT Attack

In this section, we evaluate our EBM defense using the adaptive BPDA+EOT attack in (12). To our knowledge, this is the strongest attack against our model given the current understanding of adversarial attacks. The attack technique is recently used in [25] to evaluate the purification defense [30] that is very similar to our method. We focus our attention on thoroughly investigating this attack.

4.2.1 Cifar-10

We first present our main result on the Cifar-10 dataset. 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 with the parameters in Algorithm 1. We compare our results to a representative selection of adversarial training and adversarial preprocessing defenses in Table 2. We include the training method for the classifier, transformation (if any), and the strongest attack for each defense.

Defense	$f(x)$ Train Ims.	$T(x)$ Method	Attack	Nat.	Adv.
Ours	Natural	Langevin	BPDA+EOT	0.8412	0.5490
AT [19]	Adversarial	–	PGD	0.873	0.458
TRADES [32]	Adversarial	–	PGD	0.849	0.5643
Semi-sup. AT [6]	Adversarial	–	PGD	0.897	0.625
PixelDefend [23]	Natural	Gibbs Update	BPDA	0.95	0.09 [1]
MALADE [24]	Natural	Langevin	PGD	–	0.0016
ME-Net [30]	Transformed	Mask + Recon.	BPDA+EOT	0.94	0.15 [25]

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

Our EBM defense is significantly stronger than standard adversarial training [19] and comparable to modified adversarial training such as [32]. Although our results are not on par with state-of-the-art adversarial training such as [6], we note that our defense method has the additional advantage of securing naturally trained networks. Prior pre-processing methods such as [23, 24, 30] (and several others) do not provide significant defense without adversarial training. In fact, combining pre-processing with adversarial training often results in *lower* robustness that standard adversarial training [25]. Our primary intention in this work is to demonstrate that effective defense of naturally-trained classifiers is possible.

4.2.2 Attack Diagnostics

In this section, we 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 from Section 4.2.1. To reduce the computational cost of the diagnostics, we use a fixed set of 1000 randomly selected test images for all evaluations in this section.

Figure 5 displays the robustness of our model and standard adversarial training [19] 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 model is more robust than adversarial training for a range of medium-size distortions for the l_∞ norm and much more robust than adversarial training for medium and large ε in the l_2 norm.

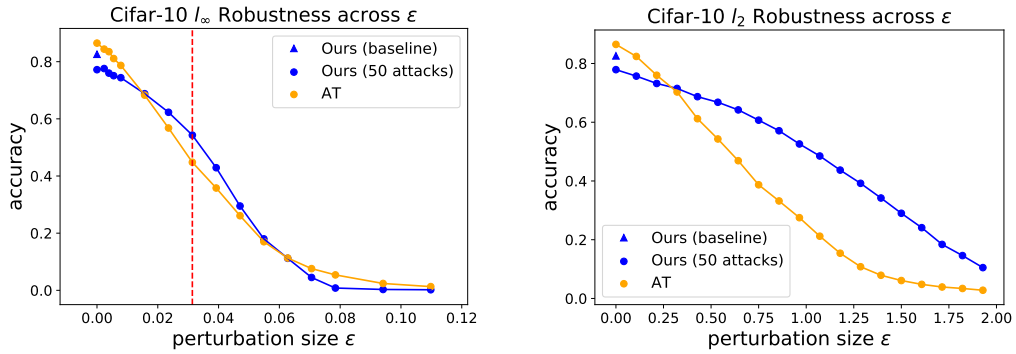

Figure 5: Accuracy across perturbation ε for l_∞ and l_2 attacks against our defense and AT [19].

The plots in Figure 5 include two evaluations of baseline accuracy at $\varepsilon = 0$. The true baseline (given by the triangle) uses $H_{\text{def}} = 150$ replicates to evaluate each natural image a single time. The pseudo-baseline (given by the circle) shows the result of applying Algorithm 1 for 50 steps without an attack. This can involve multiple evaluations of $\hat{F}_{H_{\text{def}}}(x)$ which cause false breaks, since even 150 replicates may not be enough for accurate evaluation of $F(x)$ for close borderline images. Attrition from false breaks probably also affects our results in Table 2. The discrepancy can be avoided by simply increasing H_{def}, but this is expensive in practice.

Figure 6 visualizes the effect of increasing the computational power of the attacker. The left figure compares our defense and adversarial training over 1000 attacks. The attack also uses a slightly increased number $H_{\text{adv}} = 20$ of attack replicates. The majority of breaks happen within the first 50 attacks as used in Section 4.2.1, while a small number of breaks occur within a few hundred attack steps. We note this 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 after about 20 to 30 replicates are used. 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 the 5 random restarts that were used in Section 4.2.1.

Overall, the diagnostics indicate that the defense report in Table 2 is a fair evaluation of our model. On one hand, increasing H_{def} would likely secure a small proportion of borderline images that still break due to random effects. On the other hand, increasing N and H_{adv} would likely to slightly stronger attacks. Our evaluation in Table 2 is already pushes the limits of what is feasible given

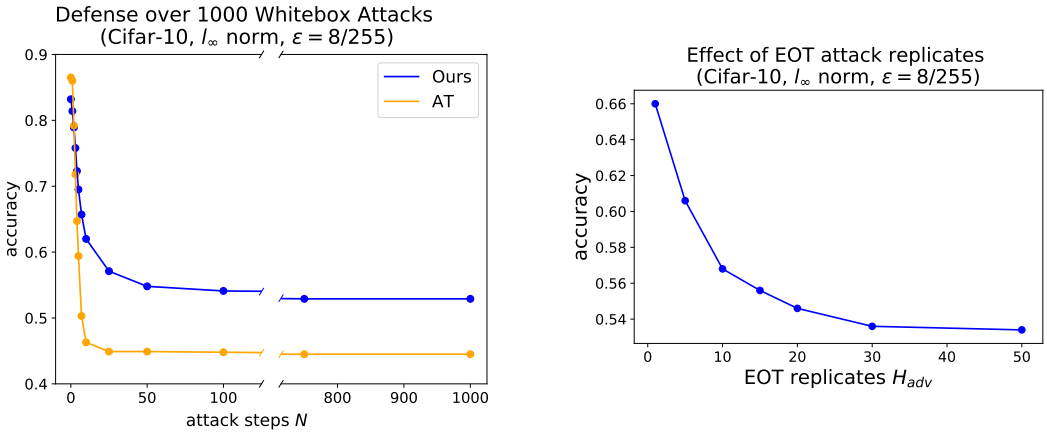

Figure 6: Effect of number of attack steps N and number of EOT replicates H_{adv}.

widely-available computational resources. We are confident that our defense report is a reasonable approximation of the defense of our ideal classifier $F(x)$ in (7) against the BPDA+EOT attack (12), although we acknowledge that more computation would yield a more accurate estimate.

4.2.3 SVHN and Cifar-100

We now present defense to the SVHN and Cifar-100 datasets. For comparison, we implement standard adversarial training for both datasets using 10 attacks per classifier and update with perturbation $\varepsilon = 8/255$ and step size $\alpha = 2/255$. We found that SVHN defense improved by using a larger step size $\tau = 0.0175$. Otherwise, the attack and defense parameters for our method are identical to those used in Section 4.2.1.

	SVHN		Cifar-100	
	Nat.	Adv.	Nat.	Adv.
Ours	0.9223	0.6755	0.5166	0.2610
AT [19]	0.8957	0.5039	0.5958 [2]	0.2547 [2]

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

Our defense method performs well on both datasets. The SVHN results for our defense are much stronger than adversarial training. Our defense also outperforms adversarial training on Cifar-100 by a small margin. This difference likely occurs because learning stable representations of aligned digits is relatively easy compared to natural image concepts which exhibit a much greater variety of appearances and contexts. 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 across Langevin iterations could yield significant benefits for natural image settings where precise details need to be preserved for accurate classification.

5 Related Work

Adversarial training learns a robust classifier using adversarial images created during each weight update. The method is introduced in [19]. Adversarial training has proven to be the most reliable defense method and many variations have been explored. We briefly discuss two adversarial training strategies that relate to our defense. The work [14] uses noise injection into each network layer to increase robustness via stochastic effects. Similarly, Langevin updates with our EBM can be interpreted as a ResNet [13] with noise injected layers as discussed in [21]. Semi-supervised adversarial training methods [26, 6] show that using unlabeled data can dramatically improve robustness. The EBM in our method similarly introduces an unsupervised representation of the data to facilitate defense.

Adversarial preprocessing is a strategy where auxiliary transformations are applied to adversarial inputs before they are given to the classifier. Some forms of pre-processing techniques for defense include rescaling [27], thermometer encoding [5], feature squeezing [29], activation pruning [8], and reconstruction [22]. Adversarial defense with iterative transformations, which we call adversarial purification, is a subclass of preprocessing defenses. Prior adversarial purification methods include energy-based methods such as Pixel-Defend [23] and MALADE [24] that sample with a density that differs from (1). ME-Net [30] iteratively removes and reconstruct image features. It was shown in [1, 25] that existing preprocessing defenses can be totally broken or dramatically weakened by simple adjustments to the standard PGD attack, namely the EOT and BPDA techniques. No prior preprocessing defense has emerged to compete with adversarial training.

Energy-based models are a probabilistic method for unsupervised modeling. Early energy-based image models include the FRAME [33] and RBM [15]. The EBM (1) used in this work is introduced in [28] and important observations about the learning process are presented in [20, 21]. Preliminary investigations for using the EBM (1) for adversarial defense are presented in [9, 11] but the results are not competitive with adversarial training (see Section 4.1). Our work builds on the convergent learning methodology in [20] to apply long-run Langevin sampling as a defense technique.

6 Conclusion

In this work we demonstrate the first use of an EBM as an effective defense method for naturally trained image classifiers against white box attacks. Our defense is founded on an improvement to EBM training that enables efficient learning of stable long-run sampling for complex datasets. We demonstrate the capabilities of 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. Our work is an important step for viable alternative approaches to adversarial defense that focus on securing pre-existing classifiers.

Acknowledgments and Disclosure of Funding

This work is partially supported by CRISP Center under the DARPA JUMP Program GI18518.156870.

References

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

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

[3] Giancarlo Benettin, Luigi Galgani, and Jean-Marie Streclyn, *Kolmogorov entropy and numerical experiments*, Physical Review A **14** (1976), no. 6, 2338–2345.

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

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

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

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

[8] 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).

[9] Yilun Du and Igor Mordatch, *Implicit generation and generalization in energy-based models*, NeurIPS (2019).

[10] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, ICLR (2015).

[11] 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).

[12] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten, *Countering adversarial images using input transformations*, ICLR (2018).

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

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

[15] Geoffrey E. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation **14** (2002), no. 8, 1771–1800.

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

[17] Diederik P. Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, ICLR (2015).

[18] 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).

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

[20] 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).

[21] 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).

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

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

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

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

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

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

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

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

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

[31] Sergey Zagoruyko and Nikos Komodakis, *Wide residual networks*, arXiv preprint arXiv:1605.07146 (2016).

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

[33] 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** (1998), no. 2, 107–126.

A Improved Learning of Convergent EBM's

Algorithm 2 gives pseudo-code for our improvement of the convergent learning method in [20]. This implementation allows efficient learning of convergent EBM's for complex datasets.

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 τ_{data}, step size τ, Langevin steps K, Adam learning rate γ_{Adam}, SGD learning rate γ_{SGD}.

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

B 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.

B.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 [23, 24, 11]. 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 7).

The theoretical foundation of probabilistic purification 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. 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 [4] 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

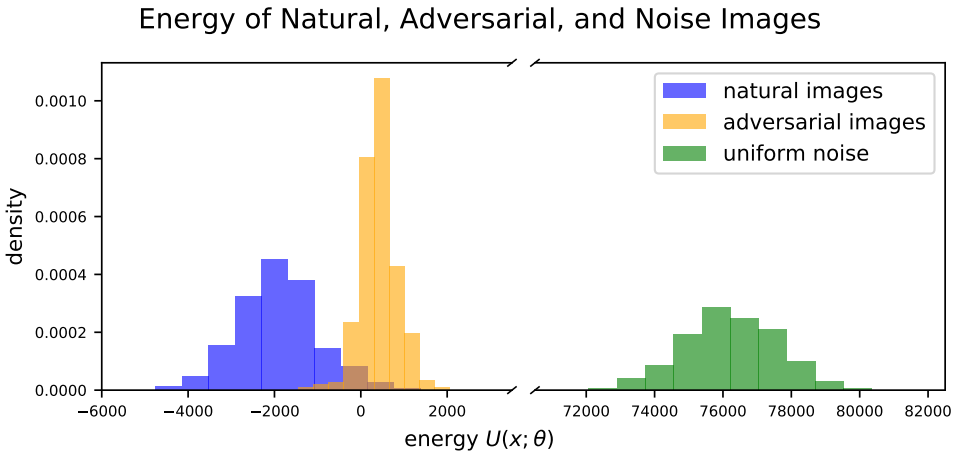

Figure 7: Energy $U(x;\theta)$ of natural, adversarial, and noise images.

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 careful balance between the memoryless properties of MCMC sampling that erase noise and the metastable properties of MCMC sampling that preserve the initial state.

B.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

$$\frac{dX}{dt} = V(X) + \eta\xi(t), \tag{13}$$

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 [18]. 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 (13) 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.

We use the classical method [3] 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 purification 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 8. We believe that the unpredictability of paths under T^* is an effective defense against BPDA because informative attack gradients cannot be generated through chaotic purification. Other chaotic transformations, either or stochastic or deterministic, might be an interesting line of research as a class of defense methods.

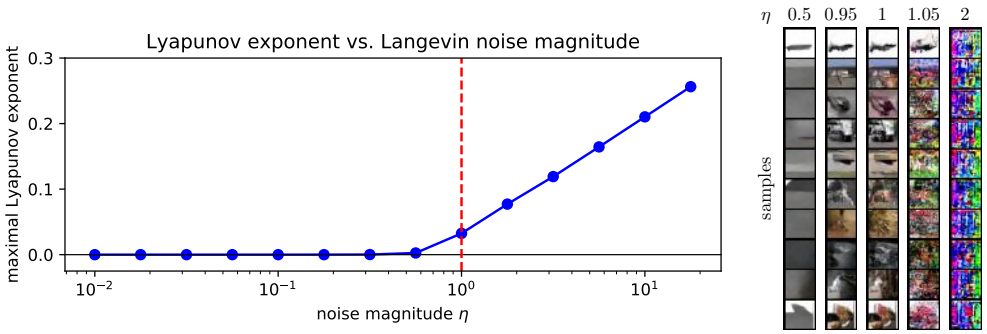

Figure 8: *Left:* Maximal Lyapunov exponent for different values of η. The value $\eta = 1$ which corresponds to our training and purification sampling dynamics is just above the transition from the ordered region where the maximal exponent is 0 to the chaotic region that where the maximal exponent is 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.

C Discussion of IGEBM and JEM Defenses

We hypothesize that the non-convergent behavior of the IGEBM [9] and JEM [11] models limits their use as an EBM defense method. Long-run samples from both models have oversaturated and unrealistic appearance (see Figure 9). 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.

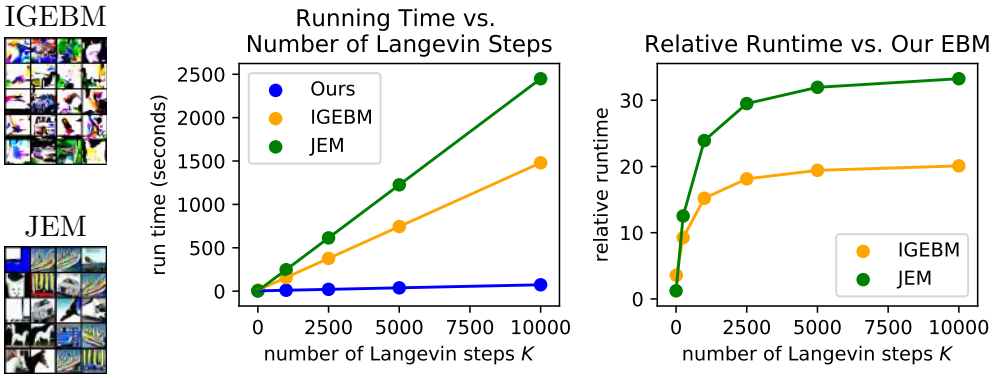

Figure 9: *Left:* Approximate steady-state samples of the IGEM and JEM models. Both exhibit oversaturation from non-convergent learning that can interfere with defense capabilities. *Right:* Comparison of running time for Langevin sampling with a batch of 100 images. The small scale and fast sampling 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 [21] (see Figure 9). 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.