Towards Bridging the Performance Gaps of Joint Energy-based Models

Xiulong Yang, Qing Su, and Shihao Ji
Georgia State University
{xyang22,qsu3,sji}@gsu.edu

Abstract

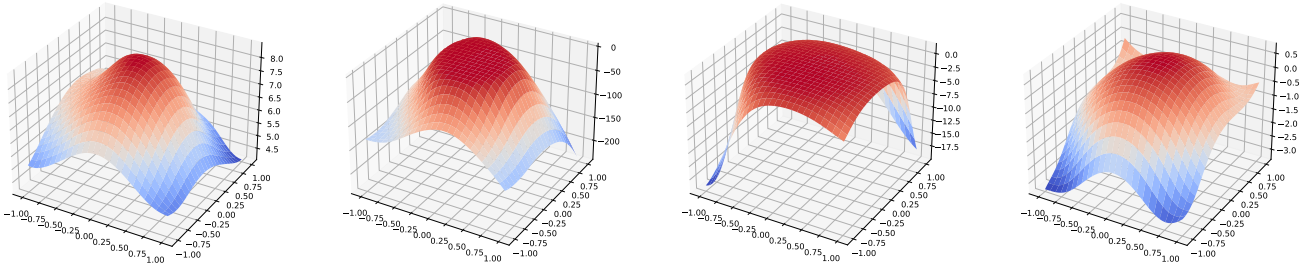
Can we train a hybrid discriminative-generative model with a single network? This question has recently been answered in the affirmative, introducing the field of Joint Energy-based Model (JEM) [17, 48], which achieves high classification accuracy and image generation quality simultaneously. Despite recent advances, there remain two performance gaps: the accuracy gap to the standard softmax classifier, and the generation quality gap to state-of-the-art generative models. In this paper, we introduce a variety of training techniques to bridge the accuracy gap and the generation quality gap of JEM. 1) We incorporate a recently proposed sharpness-aware minimization (SAM) framework to train JEM, which promotes the energy landscape smoothness and the generalization of JEM. 2) We exclude data augmentation from the maximum likelihood estimate pipeline of JEM, and mitigate the negative impact of data augmentation to image generation quality. Extensive experiments on multiple datasets demonstrate our SADA-JEM achieves state-of-the-art performances and outperforms JEM in image classification, image generation, calibration, out-of-distribution detection and adversarial robustness by a notable margin. Our code is available at https://github.com/sndnyang/SADAJEM.

1. Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performances in a wide range of learning tasks, including image classification, image generation, object detection, and language understanding [21, 30]. Among them, energy-based models (EBMs) have seen a flurry of interest recently, partially inspired by the impressive results of IGEBM [10] and JEM [17], which exhibit the capability of training generative models within a discriminative framework. Specifically, JEM [17] reinterprets the standard softmax classifier as an EBM and achieves impressive performances in image classification and generation simultaneously. Furthermore, these EBMs enjoy improved performance on out-of-distribution detection, calibration, and adversarial robustness. The follow-up works (e.g., [18, 48]) further improve the training in terms of speed, stability and accuracy.

Despite the recent advances and the appealing property of training a single network for hybrid modeling, training JEM is still challenging on complex high-dimensional data since it requires an expensive MCMC sampling. Furthermore, models produced by JEM still have an accuracy gap to the standard softmax classifier and a generation quality gap to the GAN-based approaches.

In this paper, we introduce a few simple yet effective training techniques to bridge the accuracy gap and generation quality gap of JEM. Our hypothesis is that both performance gaps are the symptoms of lack of generalization of JEM trained models. We therefore analyze the trained models under the lens of loss geometry. Figure 1 visualizes the energy landscapes of different models by the technique introduced in [34]. Since different models are trained with different loss functions, visualizing their loss functions is meaningless for the purpose of comparison. Therefore, the LSE energy functions (i.e., Eq. 4) of different models are visualized. Comparing Figure 1(a) and (b), we find that JEM converges to extremely sharp local maxima of the energy landscape as manifested by the significantly large y-axis scale. By incorporating the recently proposed sharpness-aware minimization (SAM) [12] to JEM, the energy landscape of trained model (JEM+SAM) becomes much smoother as shown in Figure 1(c). This also substantially improves the image classification accuracy and generation quality. To further improve the energy landscape smoothness, we exclude data augmentation from the maximum likelihood estimate pipeline of JEM, and visualize the energy landscape of SADA-JEM in Figure 1(d), which achieves the smoothest landscape among all the models considered. This further improves image generation quality dramatically while retaining or sometimes improving classification accuracy. Since our method improves the performance of JEM primarily in the framework of sharpness-aware optimization, we refer it as SADA-JEM, a Sharpness-Aware Joint Energy-based Model with single branched Data Augmentation.



(a) Softmax Classifier (b) JEM (c) JEM+SAM (d) SADA-JEM

Figure 1. Visualizing the energy landscapes [34] of different models trained on CIFAR10. Note the dramatic scale differences of the y-axes, indicating SADA-JEM identifies the smoothest local optimum among all the methods considered.

Our main contributions are summarized as follows:

1. We investigate the energy landscapes of different models and find that JEM leads to the sharpest one, which potentially undermines the generalization of trained models.
2. We incorporate the sharpness-aware minimization (SAM) framework to JEM to promote the energy landscape smoothness, and thus model generalization.
3. We recognize the negative impact of data augmentation in the training pipeline of JEM, and introduce two data loaders for image classification and image generation separately, which improves image generation quality significantly.
4. Extensive experiments on multiple datasets show that SADA-JEM achieves the state-of-the-art discriminative and generative performances, while outperforming JEM in calibration, out-of-distribution detection and adversarial robustness by a notable margin.

2. Related Work

Energy-Based Models (EBMs) [33] stem from the observation that any probability density function $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ can be expressed via a Boltzmann distribution as

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x})\right)}{Z(\boldsymbol{\theta})}, \tag{1}$$

where $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ is an energy function that maps input $\boldsymbol{x} \in \mathcal{X}$ to a scalar, and $Z(\boldsymbol{\theta}) = \int_{\boldsymbol{x}} \exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x})\right)$ is the normalizing constant w.r.t. $\boldsymbol{x}$ (also known as the partition function). Ideally, an energy function should assign low energy values to the samples drawn from data distribution, and high values otherwise.

The key challenge of EBM training is to estimate the intractable partition function $Z(\boldsymbol{\theta})$, and thus the maximum likelihood estimate of parameters $\boldsymbol{\theta}$ is not straightforward. Specifically, the derivative of the log-likelihood of $\boldsymbol{x} \in \mathcal{X}$ w.r.t. $\boldsymbol{\theta}$ can be expressed as

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}\right] - \mathbb{E}_{p_d(\boldsymbol{x})}\left[\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}\right], \tag{2}$$

where $p_d(\boldsymbol{x})$ is the real data distribution (i.e., training dataset), and $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the estimated probability density function, sampling from which is challenging due to the intractable $Z(\boldsymbol{\theta})$.

Prior works have developed a number of sampling-based approaches to sample from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ efficiently, such as MCMC and Gibbs sampling [25]. By utilizing the gradient information, Stochastic Gradient Langevin Dynamics (SGLD) [46] has been employed recently to speed up the sampling from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ [10, 17, 39]. Specifically, to sample from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, the SGLD follows

$$\begin{aligned} &\boldsymbol{x}^0 \sim p_0(\boldsymbol{x}), \\ &\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\alpha}{2}\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{x}^t)}{\partial \boldsymbol{x}^t} + \alpha \boldsymbol{\epsilon}^t, \quad \boldsymbol{\epsilon}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \end{aligned} \tag{3}$$

where $p_0(\boldsymbol{x})$ is typically a uniform distribution over $[-1, 1]$, whose samples are refined via a noisy gradient decent with step-size α over a sampling chain.

Joint Energy-based Model (JEM) [17] reinterprets the standard softmax classifier as an EBM and trains a single network for hybrid discriminative-generative modeling. Specifically, Grathwohl et al. [17] were the first to recognize the logits $f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]$ from a standard softmax classifier can be considered as an energy function over $(\boldsymbol{x}, y)$, and thus the joint density can be defined as $p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = e^{f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]}/Z(\boldsymbol{\theta})$, where $Z(\boldsymbol{\theta})$ is an unknown normalizing constant (regardless of $\boldsymbol{x}$ or y). Then the density of $\boldsymbol{x}$ can be derived by marginalizing over y: $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_y p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \sum_y e^{f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]}/Z(\boldsymbol{\theta})$. Subsequently, the corresponding energy function of $\boldsymbol{x}$ can be identified as

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\log \sum_y e^{f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]} = -\mathrm{LSE}(f_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{4}$$

where LSE$(\cdot)$ denotes the Log-Sum-Exp function.

To optimize the model parameter $\boldsymbol{\theta}$, JEM maximizes the logarithm of joint density function $p_{\boldsymbol{\theta}}(\boldsymbol{x}, y)$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{5}$$

where the first term denotes the cross-entropy objective for classification, and the second term can be optimized by the maximum likelihood learning of EBM as shown in Eq. 2.

However, JEM suffers from high training instability even with a large number of SGLD sampling steps K (e.g., $K = 20$). After divergence, JEM requires to restart the SGLD sampling with a doubled K. Recently, JEM++ [48] proposes a number of new training techniques to improve JEM's accuracy, training stability and speed altogether, including the proximal gradient clipping, YOPO-based SGLD sampling acceleration, and informative initialization. Furthermore, JEM++ enables batch norm [27] in the backbone models, while IGEBM and JEM have to exclude batch norm due to the high training instability incurred by it.

Flat Minima and Generalization A great number of previous works have investigated the relationship between the flatness of local minima and the generalization of learned models [6, 12, 29, 32, 34, 45]. Now it is widely accepted and empirically verified that flat minima tend to give better generalization performance. Based on these observations, several recent regularization techniques are proposed to search for the flat minima of loss landscapes [6, 12, 32, 45]. Among them, the sharpness-aware minimization (SAM) [12] is a recently introduced optimizer that demonstrates promising performance across all kinds of models and tasks, such as ResNet [21], Vision Transformer (ViT) [6] and Language Modeling [3]. Furthermore, score matching-based methods [26, 41–43] also explore the behaviour of flat minima in generative models and learn unnormalized statistical models by matching the gradient of the log probability density of the model distribution to that of the data distribution. To the best of our knowledge, we are the first to explore the sharpness-aware optimization to improve both the discriminative and generative performance of EBMs.

3. SADA-JEM

3.1. Sharpness-Aware Minimization

To train a generalizable model, SAM [12] proposes to search for model parameters $\boldsymbol{\theta}$ whose entire neighborhoods have uniformly low loss values by optimizing a minimax objective:

$$\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{train}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{6}$$

where ρ is the radius of the L_2-ball centered at model parameters $\boldsymbol{\theta}$, and λ is a hyperparameter for L_2 regularization on $\boldsymbol{\theta}$. To solve the inner maximization problem, SAM employs the Taylor expansion to develop an efficient first-order approximation to the optimal $\boldsymbol{\epsilon}^*$ as:

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) &= \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(\boldsymbol{\theta}) + \epsilon^T \nabla_{\boldsymbol{\theta}} L_{train}(\boldsymbol{\theta}) \\ &= \rho \nabla_{\boldsymbol{\theta}} L_{train}(\boldsymbol{\theta}) / \|\nabla_{\boldsymbol{\theta}} L_{train}(\boldsymbol{\theta})\|_2, \end{aligned} \tag{7}$$

which is a scaled L_2 normalized gradient at the current model parameters $\boldsymbol{\theta}$. Once $\hat{\boldsymbol{\epsilon}}$ is determined, SAM updates $\boldsymbol{\theta}$ based on the gradient $\nabla_{\boldsymbol{\theta}} L_{train}(\boldsymbol{\theta})|_{\boldsymbol{\theta}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} + 2\lambda\boldsymbol{\theta}$ at an updated parameter location $\boldsymbol{\theta} + \hat{\boldsymbol{\epsilon}}$. More recently, Kwon et al. [32] propose an Adaptive SAM (ASAM) with the objective:

$$\min_{\boldsymbol{\theta}} \max_{\|T_{\boldsymbol{\theta}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho} L_{train}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{8}$$

where $T_{\boldsymbol{\theta}}$ is an element-wise operator $T_{\boldsymbol{\theta}} = \text{diag}(|\theta_1|, |\theta_2|, \ldots, |\theta_k|)$ with $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]$. Similar to SAM, the Taylor expansion is leveraged in ASAM to derive a first-order approximation to the optimal $\boldsymbol{\epsilon}^*$ with $\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) = \rho T_{\boldsymbol{\theta}} \,\text{sign}(\nabla L_{train}(\boldsymbol{\theta}))$.

As we observed from Figure 1(a) and (b), models trained by JEM converge to very sharp local optima, which potentially undermines the generalization of JEM. We therefore incorporate the framework of SAM to the original training pipeline of JEM [17] in order to improve the generalization of trained models. Specifically, instead of the traditional maximum likelihood training, we optimize the joint density function of JEM in a minimax objective:

$$\max_{\boldsymbol{\theta}} \min_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \log p_{(\boldsymbol{\theta}+\boldsymbol{\epsilon})}(\boldsymbol{x}, y) + \lambda \|\boldsymbol{\theta}\|_2^2. \tag{9}$$

For the outer maximization that involves $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, SGLD is again used to sample from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ as in the original JEM.

3.2. Image Generation without Data Augmentation

Data augmentation is a critical technique in supervised deep learning and self-supervised contrastive learning [5, 31]. Not surprisingly, JEM also utilizes data augmentation in its training pipeline, such as horizontal flipping, random cropping, and padding. Specifically, let T denote a data augmentation operator. The actual objective function of JEM is

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \log p_{\boldsymbol{\theta}}(y|T(\boldsymbol{x})) + \log p_{\boldsymbol{\theta}}(T(\boldsymbol{x})), \tag{10}$$

which shows that JEM maximizes the likelihood function $p_{\boldsymbol{\theta}}(T(\boldsymbol{x}))$ rather than $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. From our empirical studies, horizontal flipping has little impact on the image generation quality, while cropping and padding play a bigger role because the generated images contain cropping and padding effects, which hurt the quality of generated images. This is consistent with GANs [14], which observed that any augmentation that is applied to the training dataset will get inherited in the generated images. Based on this observation,

Algorithm 1 SADA-JEM Training: Given network f_θ, SGLD step-size α, SGLD noise σ, SGLD steps K, replay buffer B, reinitialization frequency γ, SAM noise bound ρ, and learning rate lr

while not converged **do**
  Sample $\boldsymbol{x}^+$ and y from training dataset
  Sample $\widehat{\boldsymbol{x}}_0 \sim B$ with probability $1-\gamma$, else $\widehat{\boldsymbol{x}}_0 \sim p_0(\boldsymbol{x})$
  for $t \in [1, 2, \ldots, K]$ **do**
    $\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{x}}_{t-1} - \alpha \cdot \frac{\partial E(\widehat{\boldsymbol{x}}_{t-1})}{\partial \widehat{\boldsymbol{x}}_{t-1}} + \sigma \cdot \mathcal{N}(0, I)$
  end for
  $\boldsymbol{x}^- = \text{StopGrad}(\widehat{\boldsymbol{x}}_K)$
  $L_{\text{gen}}(\theta) = E(\boldsymbol{x}^+) - E(\boldsymbol{x}^-)$
  $L(\boldsymbol{\theta}) = L_{\text{clf}}(\boldsymbol{\theta}) + L_{\text{gen}}(\boldsymbol{\theta})$ with $L_{\text{clf}}(\boldsymbol{\theta}) = \text{xent}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$
  # Apply SAM optimizer as following:
  Compute gradient $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ of the training loss
  Compute $\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})$ with ρ as in Eq. 7
  Compute gradient $\boldsymbol{g} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})}$
  Update model parameters: $\boldsymbol{\theta} = \boldsymbol{\theta} - lr \cdot \boldsymbol{g}$
  Push $\boldsymbol{x}^-$ to B
end while

we exclude the data augmentation from $p_{\boldsymbol{\theta}}(T(\boldsymbol{x}))$ and only retain the data augmentation for classification given its pervasive success in image classification. To this end, our final objective function of SADA-JEM becomes:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \log p_{\boldsymbol{\theta}}(y|T(\boldsymbol{x})) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}), \qquad (11)$$

where the first term is calculated using a mini-batch with data augmentation, and the second term is calculated using a mini-batch without data augmentation, which can be implemented efficiently by using two data loaders. StyleGAN2-ADA [28] proposes a type of "non-leaking" data augmentation to prevent the discriminator from overfitting, and thus improves the image quality. However, from our empirical studies, we find that this technique hurts the performance of both image quality and classification accuracy.

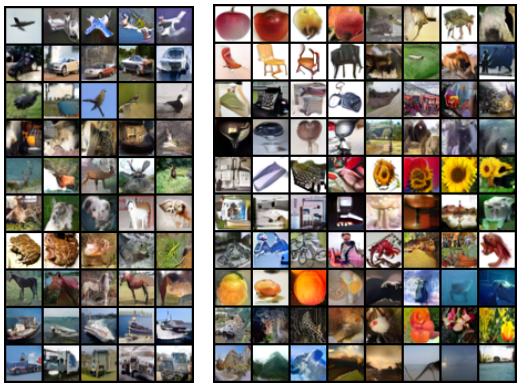
Algorithm 1 provides the pseudo-code of SADA-JEM training, which follows a similar design of JEM [17] and JEM++ [48] with a replay buffer. For brevity, only one real sample and one generated sample are used to optimize model parameters $\boldsymbol{\theta}$. But it is straightforward to generalize the pseudo-code below to a mini-batch setting, which we use in the experiments. It is worth mentioning that we adopt the Informative Initialization in JEM++ to initialize the Markov chain from $p_0(\boldsymbol{x})$, which enables the batch norm and plays a crucial role in the tradeoff between the number of SGLD sampling steps K and overall performance, including the classification accuracy and training stability.

4. Experiments

We train SADA-JEM with the Wide-ResNet 28-10 [49] backbone on CIFAR10 and CIFAR100, and evaluate its performance on a set of discriminative and generative tasks, including image classification, generation, calibration, out-of-distribution (OOD) detection, and adversarial robustness. Our code is built on top of JEM++ [48] [1] (given its improved performance over JEM) and SAM [2]. For a fair comparison, our experiments largely follow the settings of JEM and JEM++, with details provided in the supplementary material. All our experiments are conducted using PyTorch on a single Nvidia RTX GPU.

4.1. Hybrid Modeling

We first compare the performance of SADA-JEM with state-of-the-art hybrid models, stand-alone discriminative models, and generative models on CIFAR10 and CIFAR100, with the results reported in Table 1 and 2. Inception Score (IS) [40] and Fréchet Inception Distance (FID) [23] are employed to measure the quality of generated images. It can be observed from Table 1 that SADA-JEM ($K=5$) outperforms JEM ($K=20$) and JEM++ ($M=20$) in classification accuracy (95.5%) and the FID score (9.41) on CIFAR10, where the FID score of SADA-JEM is a dramatic improvement over that of JEM/JEM++'s (37.1). Similarly, Table 2 shows that the improvement of SADA-JEM over JEM/JEM++ on CIFAR100 is also significant: the FID score is improved from 33.7 to 14.4. Moreover, we find that SADA-JEM is superior in training stability too. For instance, SADA-JEM ($K=5$) outperforms JEM++ ($M=20$) in classification accuracy, while exhibiting a much higher training stability than JEM/JEM++ [3]. Example images generated by SADA-JEM for CIFAR10 and CIFAR100 are provided in Figure 2.



(a) CIFAR10 (b) CIFAR100

Figure 2. Generated samples from SADA-JEM.

[1] https://github.com/sndnyang/jempp
[2] https://github.com/davda54/sam
[3] JEM ($K=20$) and JEM++ ($M=5$) can easily diverge at early epochs.

Table 1. Results on CIFAR10

Model	Acc % ↑	IS ↑	FID ↓
SADA-JEM (K=5)	95.5	8.77	9.41
SADA-JEM (K=10)	96.0	8.63	11.4
SADA-JEM (K=20)	96.1	8.40	13.1
Single Hybrid Model			
IGEBM (K=60) [10]	49.1	8.30	37.9
JEM (K=20)* [17]	92.9	8.76	38.4
JEM++ (M=5)* [48]	91.1	7.81	37.9
JEM++ (M=10) [48]	93.5	8.29	37.1
JEM++ (M=20) [48]	94.1	8.11	38.0
JEAT [51]	85.2	8.80	38.2
Other EBMs			
CF-EBM (K=50) [50]	-	-	16.7
ImCD (K=40) [9]	-	7.85	25.1
DiffuRecov (K=30) [13]	-	8.31	9.58
VAEBM (K=6) [47]	-	8.43	12.2
VERA [18]	93.2	8.11	30.5
Other Models			
Softmax	96.2	-	-
Softmax + SAM	**97.2**	-	-
SNGAN [37]	-	8.59	21.7
StyleGAN2-ADA [28]	-	**9.74**	**2.92**

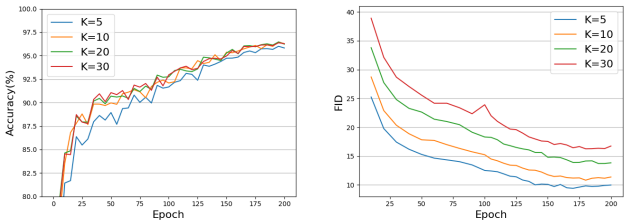
* The training is unstable and regularly diverged.

Table 2. Results on CIFAR100

Model	Acc % ↑	IS ↑	FID ↓
SADA-JEM (K=5)	75.0	**11.63**	14.4
SADA-JEM (K=10)	76.4	10.95	15.1
SADA-JEM (K=20)	77.3	10.78	19.9
JEM (K=20)* [17]	72.2	10.22	38.1
JEM++ (M=5)* [48]	72.1	8.05	38.9
JEM++ (M=10)* [48]	74.2	9.97	34.5
JEM++ (M=20)* [48]	75.9	10.07	33.7
VERA (α=100)* [18]	72.2	8.25	29.5
VERA (α=1)* [18]	48.7	7.84	25.1
Softmax	81.3	-	-
Softmax + SAM	**83.4**	-	-
SNGAN [37]	-	9.30	15.6
BigGAN [4]	-	11.0	**11.7**

* No official IS and FID scores are reported. We run the official code with the default settings and report the results.

One interesting phenomenon we observed from our experiments is that the image quality often drops as number of SGLD sampling steps K increases, as shown in Figure 3(b). A similar observation has been reported in IGEBM [10], where the authors found that a large K can facilitate the convergence of SGLD to high likelihood modes of an energy landscape, but often leads to saturated images and thus degraded image quality. Unlike IGEBM, SADA-JEM is a hybrid model that trains one single network for image classification and generation. As we can see from Figure 3, as K increases the classification accuracy of SADA-JEM increases (insignificantly), while the image quality drops. Therefore, it seems there is a performance trade-off between classification accuracy and image generation quality, and SADA-JEM's performances on both tasks are not always positively correlated after certain points (e.g., K). This is an interesting observation that we believe is worthy of further investigation.



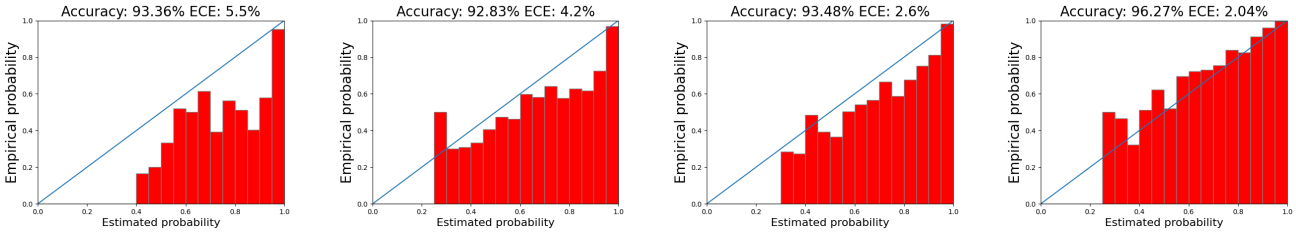
(a) Accuracy (b) FID (the lower the better)

Figure 3. The learning curves of SADA-JEM on CIFAR10 with different SGLD sampling steps K.

4.2. Calibration

While modern classifiers are growing more accurate, recent works show that their predictions could be overconfident due to increased model capacity [19]. Typically, the confidence of a model's prediction can be defined as $\max_y p(y|\boldsymbol{x})$ and is used to decide whether to output a prediction or not. However, incorrect but confident predictions can be catastrophic for safety-critical applications, which necessitates calibration of uncertainty especially for models of large capacity. As such, a well-calibrated but less accurate model can be considerably more useful than a more accurate but less-calibrated model.

In this experiment, all models are trained on the CIFAR10 dataset for a fair comparison. We compare the Expected Calibration Error (ECE) score [19] of SADA-JEM to those of the standard softmax classifier and JEM. We utilize the *reliability diagram* to visualize the discrepancy between the true probability and the confidence, with the results shown in Figure 4. We find that the model trained by SADA-JEM ($K = 10$) achieves a much smaller ECE (2.04% vs. 4.2% of JEM and 5.5% of softmax classifier), demonstrating SADA-JEM's predictions are better calibrated than the competing methods. Similar to image quality, we notice that a larger K also undermines the calibration quality slightly. Due to page limit, more results are relegated to the supplementary material.



(a) Softmax (w/o BN) (b) JEM (K=20) (c) JEM++ (M=10) (d) SADA-JEM (K=10)

Figure 4. Calibration results on CIFAR10. The smaller ECE is, the better.

Table 3. OOD detection results. Models are trained on CIFAR10. Values are AUROC.

$s_{\boldsymbol{\theta}}(\boldsymbol{x})$	Model	SVHN	CIFAR10 Interp	CIFAR100	CelebA
$\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$	WideResNet [35]	.91	-	.87	.78
	IGEBM [10]	.63	.70	.50	.70
	JEM (K=20) [17]	.67	.65	.67	.75
	JEM++ (M=20) [48]	.85	.57	.68	.80
	VERA [18]	.83	**.86**	.73	.33
	ImCD [9]	.91	.65	.83	-
	SADA-JEM (K=5)	.91	.79	.90	.82
	SADA-JEM (K=10)	.95	.81	.90	.88
	SADA-JEM (K=20)	**.98**	.83	**.92**	**.95**
$\max_y p_{\boldsymbol{\theta}}(y\|\boldsymbol{x})$	WideResNet	.93	.77	.85	.62
	IGEBM [10]	.43	.69	.54	.69
	JEM (K=20) [17]	.89	.75	.87	.79
	JEM++ (M=20) [48]	.94	.77	.88	**.90**
	SADA-JEM (K=5)	.92	.77	.88	.81
	SADA-JEM (K=10)	.93	.78	.89	.78
	SADA-JEM (K=20)	**.96**	**.80**	**.91**	.84

4.3. Out-Of-Distribution Detection

Formally, the OOD detection is a binary classification problem, which outputs a score $s_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}$ for a given query $\boldsymbol{x}$. The model should be able to assign lower scores to OOD examples than to in-distribution examples such that it can be used to distinguish OOD examples from in-distribution ones. Following the settings of JEM [17], we use the Area Under the Receiver-Operating Curve (AUROC) [22] to evaluate the performance of OOD detection. In our experiments, two score functions are considered: the input density $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ [38], and the predictive distribution $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ [22].

Input Density We can use the input density $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ as $s_{\boldsymbol{\theta}}(\boldsymbol{x})$. Intuitively, examples with low $p(\boldsymbol{x})$ are considered to be OOD samples. Quantitative results can be found in Table 3 (top row), where CIFAR10 is the in-distribution data, and SVHN, an interpolated CIFAR10, CIFAR100 and CelebA are the out-of-distribution data, respectively. Moreover, the corresponding visualization are shown in Table 4. As we can see, SADA-JEM performs better in distinguishing the in-distribution samples from OOD ones, outperforming JEM, JEM++ and most of the other models by significant margins.

Predictive Distribution Another useful OOD score function is the maximum probability from a classifier's predictive distribution: $s_{\boldsymbol{\theta}}(\boldsymbol{x}) = \max_y p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. Hence, OOD performance using this score is highly correlated with a model's classification accuracy. Table 3 (bottom row) reports the results of this method. Again, SADA-JEM outperforms JEM and all the other models in majority of cases.

Table 3 (top row) also shows that JEM and JEM++ have even worse performance than a standard classifier in OOD detection. This is likely because both JEM and JEM++ maximize $p_{\boldsymbol{\theta}}(T(\boldsymbol{x}))$ with data augmentation T, which undesirably enlarges the span of estimated $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and makes it less distinguishable to the OOD samples. In contrast, VERA, ImCD, and SADA-JEM exclude the data augmentation from their training pipelines, and consistently, they all demonstrate improved OOD detection performance over JEM and JEM++.

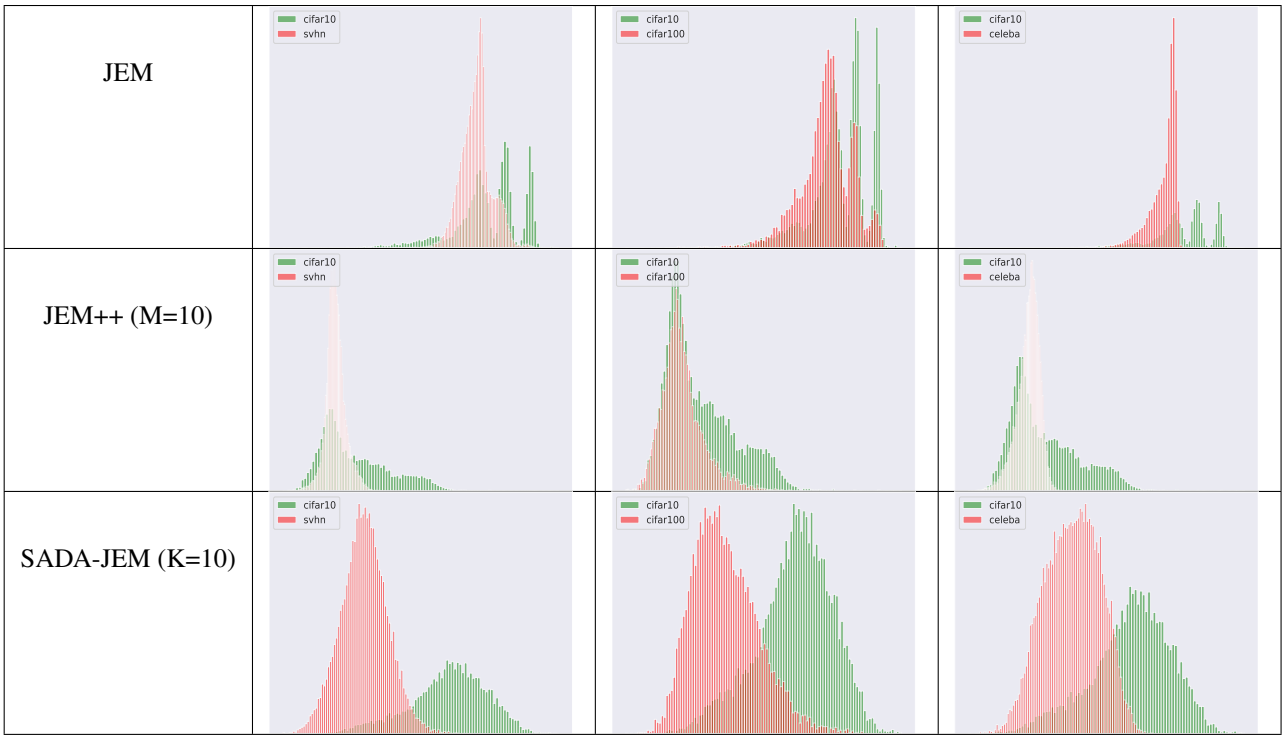


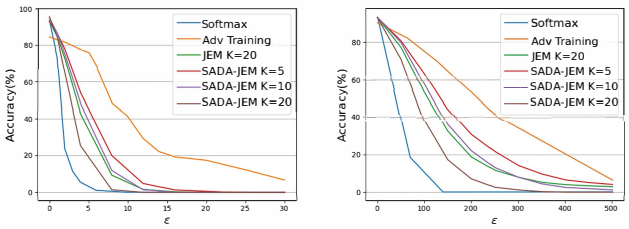
Table 4. Histograms of $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ for OOD detection. Green corresponds to in-distribution dataset, while red corresponds to OOD dataset.

4.4. Robustness

DNNs are known to be vulnerable to adversarial examples [16, 44] in the form of tiny but sensitive perturbations to the inputs that trick the model to yield incorrect predictions. To mitigate this security threat posed by adversarial examples, a variety of defense algorithms have been proposed in the past few years to improve the robustness of deep networks [1, 7, 11, 15, 20, 36]. Existing works [17, 24] have verified empirically that JEM is more robust than the softmax classifiers trained in standard procedures. Since SADA-JEM promotes the smoothness of energy landscape, it would be interesting to measure if SADA-JEM can also improve model robustness.

The white-box PGD attack [36] under an L_∞ or L_2-norm constraint is the most common approach to evaluate the robustness of a classifier. However, Athalye et al. [2] found that the defense methods using gradient obfuscation always report overrated robustness, and the defense can be overcome with minor adjustments to the standard PGD attacks. Therefore, to better evaluate the robustness of EBMs, Mitch Hill et al. [24] proposed the Expectation-Over-Transformation (EOT) attack and Backward Pass Differentiable Approximation (BPDA) attack specifically for EBMs. We therefore employ these two attacks in our experiments with the results reported in Figure 5.

As we can see, SADA-JEM achieves a similar robustness as JEM under the L_∞ and L_2 PGD attacks, while both are more robust than the standard softmax classifiers. Moreover, we find that a larger K undermines the robustness significantly, even though it can boost the accuracy on clean data. Similar observation has been reported by Yao et al. [51], who found that EBM learns a smooth energy function around real data by increasing the energy of SGLD-sampled points; however, a larger K can generate samples of lower energy which are closer to real data distribution, and thus leads to a sharper energy landscape around real data after optimizing on both real and generated samples. As such, the models trained with a larger K are less robust than the ones with a smaller K. In addition, JEM ($K = 20$) diverges regularly, and it needs to restart the training by doubling K (e.g., $K = 40$), while SADA-JEM ($K = 20$) is very stable. With a smaller K SADA-JEM achieves even higher robustness than JEM ($K = 20$).



(a) L_∞ Robustness (b) L_2 Robustness

Figure 5. Adversarial robustness under the PGD attacks.

4.5. Ablation Study

We study the impacts of SAM and data augmentation (DA) to the performance of SADA-JEM on image classification and generation in this section. The results on CIFAR10 are reported in Table 5. It can be observed that SAM can improve the classification accuracy and generation quality of JEM/JEM++, while the improvements on classification accuracy are more pronounced. Secondly, by further excluding data augmentation T from $p_{\boldsymbol{\theta}}(T(\boldsymbol{x}))$ of JEM++, which leads to SADA-JEM, the FID score is improved dramatically from 35.0 to 11.4. Prior works on EBMs [17, 39, 48] include DA to their training pipelines to stabilize the training. However, DA introduces the artifacts to the training images, leading to foggy synthesized images. As a result, by excluding DA, SADA-JEM optimizes on $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and improves image generation quality significantly, while still being very stable due to the SAM optimizer. We further experiment replacing SAM in SADA-JEM with the energy L_2 regularization proposed in IGEBM [10] to weakly regularize energy magnitudes of both positive and negative samples. We found that the L_2 regularization fails to improve the classification accuracy and degrades the training stability.

We also study the impact of the noise radius ρ to the performance of SADA-JEM in image classification and generation, with the results reported in Table 6. It can be observed that SAM with $\rho = 0.2$ achieves an overall good performance in classification and image quality, and thus is chosen as default in all our experiments.

Table 5. Ablation study of SADA-JEM. All the models are trained on CIFAR10 with $K = 10$.

Ablation	Acc% ↑	FID ↓
JEM	89.5	36.2
JEM +SAM	90.1	35.0
JEM++	93.5	37.1
JEM++ +SAM	94.1	36.6
JEM++ w/o DA	93.6	12.9
JEM++ w/o DA +L_2*	93.4	-
SADA-JEM	**96.0**	**11.4**

* It fails to generate realistic images after 110 epochs.

5. Limitations

It is challenging to train SGLD-based EBMs, including IGEBM, JEM, JEM++ and SADA-JEM, on complex high-dimensional data. IGEBM, JEM and many prior works have investigated methods to stabilize the training of EBM, but they require an extremely expensive SGLD sampling with a large K. Our SADA-JEM can stabilize the training on CIFAR10 and CIFAR100 with a small K (e.g., $K = 5$). However, when the image resolution scales up (e.g., from 32x32 to 224x224), SADA-JEM has to increase K accordingly to improve image generation quality. Hence, the trade-off between generation quality and computational complexity still limits the application of SADA-JEM to large-scale benchmarks, including ImageNet [8].

Table 6. Ablation study of SADA-JEM on ρ. All the models are trained on CIFAR10 with $K = 10$.

Ablation	Acc % ↑	FID ↓
ASAM ($\rho = 0.5$)	94.2	12.1
ASAM ($\rho = 1$)	94.5	11.9
ASAM ($\rho = 2$)	94.8	11.7
ASAM ($\rho = 4$)	95.3	11.5
ASAM ($\rho = 8$)	Diverged after 2nd epoch	
SAM ($\rho = 0.05$)	94.8	**10.9**
SAM ($\rho = 0.1$)	95.5	11.4
SAM ($\rho = 0.2$)	**96.0**	11.4
SAM ($\rho = 0.4$)	95.1	14.1
SAM ($\rho = 0.8$)	91.9	19.5

Besides, the computation bottleneck of SADA-JEM is not SAM as SAM is only used to optimize model parameters $\boldsymbol{\theta}$ (the outer maximization in Eq. 9). Instead, the K SGLD sampling steps (typically $K = 10$) is the most expensive operation (the inner minimization in Eq. 9). SAM doubles the cost of $\boldsymbol{\theta}$ optimization, which is insignificant compared to K SGLD steps. Overall, the training speed of SADA-JEM is comparable to JEM/JEM++. Therefore, a more efficient sampling method is required to scale up SADA-JEM to large-scale applications.

6. Conclusion

We propose SADA-JEM to bridge the classification accuracy gap and the generation quality gap of JEM. By incorporating the framework of SAM to JEM and excluding the undesirable data augmentation from the training pipeline of JEM, SADA-JEM promotes the energy landscape smoothness and hence the generalization of trained models. Our experiments verify the effectiveness of these techniques on multiple benchmarks and demonstrate the state-of-the-art results in most of the tasks of image classification, generation, uncertainty calibration, OOD detection and adversarial robustness. As for the future work, we are interested in improving the scalability of EBMs to large-scale benchmarks, such as ImageNet and NLP tasks.

7. Acknowledgement

We would like to thank the anonymous reviewers for their comments and suggestions, which helped improve the quality of this paper. We would also gratefully acknowledge the support of Cisco Systems, Inc. for its university research fund to this research.

References

[1] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 7

[3] Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 3

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 3

[6] Xiangning Chen, Cho-jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[7] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *International Conference on Learning Representations (ICLR) 2020*, 2020. 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 8

[9] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved Contrastive Divergence Training of Energy Based Models. In *International Conference on Machine Learning (ICML)*, 2021. 5, 6

[10] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. In *Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 5, 6, 8

[11] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 7

[12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 1, 3

[13] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. Learning Energy-Based Models by Diffusion Recovery Likelihood. In *ICLR*, 2021. 5

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems (NeurIPS)*, 2014. 3

[15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[17] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[18] Will Sussman Grathwohl, Jacob Jin Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No mcmc for me: Amortized sampling for fast and stable training of energy-based models. In *ICLR*, 2021. 1, 5, 6

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. 5

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. 6

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems (NeurIPS)*, 2017. 4

[24] Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations (ICLR)*, 2021. 7

[25] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002. 2

[26] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005. 3

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. 3

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 4, 5

[29] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[30] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2012. 3

[32] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2021. 3

[33] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006. 2

[34] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In *Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3

[35] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 7

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[38] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018. 6

[39] Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent short-run mcmc toward energy-based model. In *Neural Information Processing Systems (NeurIPS)*, 2019. 2, 8

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Neural Information Processing Systems (NeurIPS)*, 2016. 4

[41] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems (NeurIPS)*, 2019. 3

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 3

[43] Kevin Swersky, David Buchman, Nando D Freitas, Benjamin M Marlin, et al. On autoencoders and score matching for energy based models. In *International Conference on Machine Learning (ICML)*, 2011. 3

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 7

[45] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning (ICML)*, 2020. 3

[46] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011. 2

[47] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[48] Xiulong Yang and Shihao Ji. JEM++: Improved Techniques for Training JEM. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 4, 5, 6, 8

[49] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 4

[50] Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[51] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, and Zhenguo Li. Towards Understanding the Generative Capability of Adversarially Robust Classifiers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 5, 7