

Why Spectral Normalization Stabilizes GANs: Analysis and Improvements

Zinan Lin¹ Vyas Sekar¹ Giulia Fanti¹

Abstract

Spectral normalization (SN) (Miyato et al., 2018) is a widely-used technique for improving the stability and sample quality of Generative Adversarial Networks (GANs). However, there is currently limited understanding of why SN is effective. In this work, we show that SN controls two important failure modes of GAN training: exploding and vanishing gradients. Our proofs illustrate a (perhaps unintentional) connection with the successful LeCun initialization (LeCun et al., 1998). This connection helps to explain why the most popular implementation of SN for GANs (Miyato et al., 2018) requires no hyper-parameter tuning, whereas stricter implementations of SN (Gouk et al., 2018; Farnia et al., 2018) have poor empirical performance out-of-the-box. Unlike LeCun initialization which only controls gradient vanishing at the beginning of training, SN preserves this property throughout training. Building on this theoretical understanding, we propose a new spectral normalization technique: Bidirectional Scaled Spectral Normalization (BSSN), which incorporates insights from later improvements to LeCun initialization: Xavier initialization (Glorot & Bengio, 2010) and Kaiming initialization (He et al., 2015). Theoretically, we show that BSSN gives better gradient control than SN. Empirically, we demonstrate that it outperforms SN in sample quality and training stability on several benchmark datasets.

1. Introduction

Generative adversarial networks (GANs) are state-of-the-art deep generative models, perhaps best known for their ability to produce high-resolution, photorealistic images (Goodfellow et al., 2014). The objective of GANs is to produce

¹Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. Correspondence to: Zinan Lin <zinanl@andrew.cmu.edu>, Vyas Sekar <vsekar@andrew.cmu.edu>, Giulia Fanti <gfanti@andrew.cmu.edu>.

random samples from a target data distribution, given only access to an initial set of training samples. This is achieved by learning two functions: a generator G , which maps random input noise to a generated sample, and a discriminator D , which tries to classify input samples as either real (i.e., from the training dataset) or fake (i.e., produced by the generator). In practice, these functions are implemented by deep neural networks (DNNs), and the competing generator and discriminator are trained in an alternating process known as *adversarial training*. Theoretically, given enough data and model capacity, GANs converge to the true underlying data distribution (Goodfellow et al., 2014).

Although GANs have been very successful in improving the sample quality of data-driven generative models (Karras et al., 2017; Brock et al., 2018), their adversarial training also contributes to instability. That is, small hyper-parameter changes and even randomness in the optimization can cause training to fail. Many approaches have been proposed for improving the stability of GANs, including different architectures (Radford et al., 2015; Karras et al., 2017; Brock et al., 2018), loss functions (Arjovsky & Bottou, 2017; Arjovsky et al., 2017; Gulrajani et al., 2017; Wei et al., 2018), and various types of regularizations/normalizations (Miyato et al., 2018; Brock et al., 2016; Salimans & Kingma, 2016). One of the most successful proposals to date is called *spectral normalization* (SN) (Miyato et al., 2018; Gouk et al., 2018; Farnia et al., 2018). SN forces each layer of the generator to have unit spectral norm during training. This has the effect of controlling the Lipschitz constant of the discriminator, which is empirically observed to improve the stability of GAN training (Miyato et al., 2018).

Despite the successful applications of SN (Brock et al., 2018; Lin et al., 2020; Zhang et al., 2018; Jolicoeur-Martineau, 2018; Yu et al., 2019; Miyato & Koyama, 2018; Lee et al., 2018), to date, it remains unclear precisely why this specific normalization is so effective.

In this paper, we show that SN controls two important failure modes of GAN training: exploding gradients and vanishing gradients. These problems are well-known to cause instability in GANs (Arjovsky et al., 2017; Brock et al., 2018), leading either to bad local minima or stalled training prior to convergence. We make three primary contributions:

(1) *Analysis of why SN avoids exploding gradients (§ 3).*



Figure 1. The interesting connections we find between spectral normalizations and prior initialization techniques: (1) The insights from LeCun initialization (LeCun et al., 1998) can help explain why SN avoids exploding gradients; (2) Motivated from newer initialization techniques (Glorot & Bengio, 2010; He et al., 2015), we proposed BSSN to further improve SN.

Poorly-chosen architectures and hyper-parameters, as well as randomness during training, can amplify the effects of large gradients on training instability, ultimately leading to generalization error in the learned discriminator. We theoretically prove that SN imposes an upper bound on gradients during GAN training, mitigating these effects.

(2) *Analysis of why SN avoids vanishing gradients (§ 4).* Small gradients during training are known to cause GANs (and other DNNs) to converge to bad models (LeCun et al., 1998; Arjovsky et al., 2017). The well-known LeCun initialization, first proposed over two decades ago, mitigates this effect by carefully choosing the variance of the initial weights (LeCun et al., 1998). We prove theoretically that SN controls the variance of weights in a way that closely parallels LeCun initialization. Whereas LeCun initialization only controls the gradient vanishing problem at the beginning of training, we show empirically that SN preserves this property throughout training. Our analysis also explains why a strict implementation of SN (Farnia et al., 2018) has poor out-of-the-box performance on GANs and requires additional tuning to avoid the vanishing gradient problem, whereas the implementation of SN in (Miyato et al., 2018) requires no tuning.

(3) *Improving SN with the above theoretical insights (§ 5).* Given this new understanding of the connections between SN and LeCun initialization, we propose Bidirectional Scaled Spectral Normalization (BSSN), a new normalization technique that combines two key insights (Fig. 1): (a) It introduces a novel bidirectional spectral normalization inspired by *Xavier initialization*, which improved on LeCun initialization by controlling not only the variances of internal outputs, but also the variance of backpropagated gradients (Glorot & Bengio, 2010). We theoretically prove that BSSN mimics Xavier initialization to give better gradient control than SN. (b) BSSN introduces a new scaling of weights inspired by *Kaiming initialization*, a newer initialization technique that has better performance in practice (He et al., 2015). We show that BSSN achieves better sample quality and training stability than SN on several benchmark datasets, including CIFAR10, STL10, CelebA, and ImageNet.

Note that better gradient control should *not* be the only reason behind the success of SN (more discussions in § 6).

However, our theoretical results do show a connection between gradient control, initialization techniques, and spectral normalization. Empirical results of the two improvements we propose demonstrate the practical value of this new theoretical understanding.

2. Background and Preliminaries

The instability of GANs is believed to be predominantly caused by poor discriminator learning (Arjovsky & Bottou, 2017; Salimans et al., 2016). We therefore focus in this work on the discriminator, and the effects of SN on discriminator learning. We adopt the same model as (Miyato et al., 2018). Consider a discriminator with L internal layers:

$$D_\theta(x) = a_L \circ l_{w_L} \circ a_{L-1} \circ l_{w_{L-1}} \circ \dots \circ a_1 \circ l_{w_1}(x) \quad (1)$$

where x denotes the input to the discriminator and $\theta = \{w_1, w_2, \dots, w_L\}$ the weights; a_i ($i = 1, \dots, L - 1$) is the activation function in the i -th layer, which is usually element-wise ReLU or leaky ReLU in GANs (Goodfellow et al., 2014). a_L is the activation function for the last layer, which is sigmoid for the vanilla GAN (Goodfellow et al., 2014) and identity for WGAN-GP (Gulrajani et al., 2017); l_{w_i} is the linear transformation in i -th layer, which is usually fully-connected or a convolutional neural network (Goodfellow et al., 2014; Radford et al., 2015). Like prior work on the theoretical analysis of (spectral) normalization (Miyato et al., 2018; Farnia et al., 2018; Santurkar et al., 2018), we do not model bias terms.

Lipschitz regularization and spectral normalization. Prior work has shown that regularizing the Lipschitz constant of the discriminator $\|D_\theta\|_{\text{Lip}}$ improves the stability of GANs (Arjovsky et al., 2017; Gulrajani et al., 2017; Wei et al., 2018). For example, WGAN-GP (Gulrajani et al., 2017) adds a gradient penalty ($\|\nabla D_\theta(\tilde{x})\| - 1\|^2$) to the loss function, where $\tilde{x} = \alpha x + (1 - \alpha)G(z)$ and $\alpha \sim \text{Uniform}(0, 1)$ to ensure that the Lipschitz constant of the discriminator is bounded by 1.

Spectral normalization (SN) takes a different approach. For fully connected layers (i.e., $l_{w_i}(x) = w_i x$), it regularizes the weights w_i to ensure that spectral norm $\|w_i\|_{\text{sp}} = 1$ for all $i \in [1, L]$, where the spectral norm $\|w_i\|_{\text{sp}}$ is defined as the largest singular value of w_i . This bounds the Lipschitz constant of the discriminator since $\|D_\theta\|_{\text{Lip}} \leq \prod_{i=1}^L \|l_{w_i}\|_{\text{Lip}} \cdot \prod_{i=1}^L \|a_i\|_{\text{Lip}} \leq \prod_{i=1}^L \|w_i\|_{\text{sp}} \cdot \prod_{i=1}^L \|a_i\|_{\text{Lip}} \leq 1$, as $\|l_{w_i}\|_{\text{Lip}} \leq \|w_i\|_{\text{sp}}$ and $\|a_i\|_{\text{Lip}} \leq 1$ for networks with (leaky) ReLU as activation functions for the internal layers and identity/sigmoid as the activation function for the last layer (Miyato et al., 2018). Prior work has theoretically connected the generalization gap of neural networks to the product of the spectral norms of the layers (Bartlett et al., 2017; Neyshabur et al., 2017). These insights led to

multiple implementations of spectral normalization (Farnia et al., 2018; Gouk et al., 2018; Yoshida & Miyato, 2017; Miyato et al., 2018), with the implementation of (Miyato et al., 2018) achieving particular success on GANs. SN can be viewed as a special case of more general techniques for enhancing stability of neural network training by controlling the spectrum of the network’s input-output Jacobian (Pennington et al., 2017), e.g., through techniques like Jacobian clamping (Odena et al., 2018), which constrains the values of the maximum and minimum singular values in the generator during training.

In practice, spectral normalization (Farnia et al., 2018; Miyato et al., 2018) is implemented by dividing the weight matrix w_i by its spectral norm: $\frac{w_i}{u_i^T w_i v_i}$, where u_i and v_i are the left/right singular vectors of w_i corresponding to its largest singular value. As observed by Gouk et al. (Gouk et al., 2018), there are two approaches in the SN literature for instantiating the matrix w_i for convolutional neural networks (CNNs). In a CNN, since convolution is a linear operation, convolutional layers can equivalently be written as a multiplication by an expanded weight matrix \tilde{w}_i that is derived from the raw weights w_i . Hence in principle, spectral normalization should normalize each convolutional layer by $\|\tilde{w}_i\|_{\text{sp}}$ (Gouk et al., 2018; Farnia et al., 2018). We call this canonical normalization SN_{Conv} as it controls the spectral norm of the convolution layer.

However, the spectral normalization that is known to outperform other regularization techniques and improves training stability for GANs (Miyato et al., 2018), which we call SN_w , does not implement SN in a strict sense. Instead, it uses $\|w_i^{c_{out} \times (c_{in} k_w k_h)}\|_{\text{sp}}$; that is, it first reshapes the convolution kernel $w_i \in \mathbb{R}^{c_{out} c_{in} k_w k_h}$ into a matrix \hat{w}_i of shape $c_{out} \times (c_{in} k_w k_h)$, and then normalizes with the spectral norm $\|\hat{w}_i\|_{\text{sp}}$, where c_{in} is the number of input channels, c_{out} is the number of output channels, k_w is the kernel width, and k_h is the kernel height. Miyato et al. showed that their implementation implicitly penalizes w_i from being too sensitive in one specific direction (Miyato et al., 2018). However, this does not explain why SN_w is more stable than other Lipschitz regularization techniques, and as observed in (Gouk et al., 2018), it is unclear how SN_w relates to SN_{Conv} . Despite this, SN_w has empirically been immensely successful in stabilizing the training of GANs (Brock et al., 2018; Lin et al., 2020; Zhang et al., 2018; Jolicoeur-Martineau, 2018; Yu et al., 2019; Miyato & Koyama, 2018; Lee et al., 2018). Even more puzzling, we show in § 4 that the canonical approach SN_{Conv} has comparatively poor out-of-the-box performance when training GANs.

Hence, two questions arise: (1) Why is SN so successful at stabilizing the training of GANs? (2) Why is SN_w proposed by (Miyato et al., 2018) so much more effective than the canonical SN_{Conv} ?

In this work, we show that both questions are related to two well-known phenomena: vanishing and exploding gradients. These terms describe a problem in which gradients either grow or shrink rapidly during training (Bengio et al., 1994; Pascanu et al., 2012; 2013; Bernstein et al., 2020), and they are known to be closely related to the instability of GANs (Arjovsky & Bottou, 2017; Brock et al., 2018). We provide an example to illustrate how vanishing or exploding gradients cause training instability in GANs in App. I.

3. Exploding Gradients

In this section, we show that spectral normalization prevents gradient explosion by bounding the gradients of the discriminator. Moreover, we show that the common choice to normalize all layers equally achieves the tightest upper bound for a restricted class of discriminators. We use $\theta \in \mathbb{R}^d$ to denote a vector containing all elements in $\{w_1, \dots, w_L\}$. In the following analysis, we assume linear transformations are fully-connected layers $l_{w_i}(x) = w_i x$ as in (Miyato et al., 2018), though the same analysis can be applied to convolutional layers. Following prior work on the theoretical analysis of (spectral) normalization (Miyato et al., 2018; Farnia et al., 2018; Santurkar et al., 2018), we assume no bias in the network (i.e., Eq. (1)) for simplicity.

To highlight the effects of the spectral norm of each layer on the gradient and simplify the exposition, we will compute gradients with respect to $w'_i = \frac{w_i}{u_i^T w_i v_i}$ in the following discussion. In reality, gradients are computed with respect to w_i ; we defer this discussion to App. C, where we show the relevant extension.

How SN controls exploding gradients. The following proposition shows that under this simplifying assumption, spectral normalization controls the magnitudes of the gradients of the discriminator with respect to θ . Notice that simply controlling the Lipschitz constant of the discriminator (e.g., as in WGAN (Arjovsky & Bottou, 2017)) does not imply this property; it instead ensures small (sub)gradients with respect to the input, x .

Proposition 1 (Upper bound of gradient’s Frobenius norm for spectral normalization). *If $\|w_i\|_{\text{sp}} \leq 1$ for all $i \in [1, L]$, then we have $\|\nabla_{w_t} D_\theta(x)\|_F \leq \|x\| \prod_{i=1}^L \|a_i\|_{\text{Lip}}$, and the norm of the overall gradient can be bounded by $\|\nabla_\theta D_\theta(x)\|_F \leq \sqrt{L} \|x\| \prod_{i=1}^L \|a_i\|_{\text{Lip}}$.*

(*Proof in App. A*). Note that under the assumption that internal activation functions are ReLU or leaky ReLU, if the activation function for the last layer is identity (e.g., for WGAN-GP (Gulrajani et al., 2017)), the above bounds can be simplified to $\|\nabla_{w_t} D_\theta(x)\|_F \leq \|x\|$ and $\|\nabla_\theta D_\theta(x)\| \leq \sqrt{L} \|x\|$, and if the activation for the last layer is sigmoid (e.g., for vanilla GAN (Goodfellow et al., 2014)), the above bounds become $\|\nabla_{w_t} D_\theta(x)\|_F \leq 0.25 \|x\|$ and $\|\nabla_\theta D_\theta(x)\| \leq$

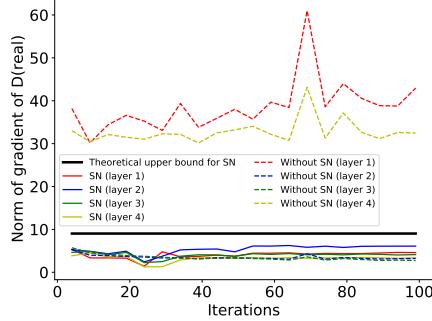


Figure 2. Gradient norms of each discriminator layer in MNIST.

$0.25\sqrt{L}\|x\|$. A comparable bound can also be found to limit the norm of the Hessian, which we defer to App. D.

The bound in Prop. 1 has a significant effect in practice. Fig. 2 shows the norm of the gradient for each layer of a GAN trained on MNIST with and without spectral normalization. Without spectral normalization, some layers have extremely large gradients throughout training, which makes the overall gradient large. With spectral normalization, the gradients of all layers are upper bounded as shown in Prop. 1. We see similar results in other datasets and network architectures (App. J).

Optimal spectral norm allocation. Common implementations of SN advocate setting the spectral norm of *each layer* to the same value (Miyato et al., 2018; Farnia et al., 2018). However, the following proposition states that we can set the spectral norms of different layers to different constants, without changing the network’s behavior on the input samples, as long as the *product* of the spectral norm bounds is the same.

Proposition 2. *For any discriminator $D_\theta = a_L \circ l_{w_L} \circ a_{L-1} \circ l_{w_{L-1}} \circ \dots \circ a_1 \circ l_{w_1}$ and $D'_\theta = a_L \circ l_{c_L \cdot w_L} \circ a_{L-1} \circ l_{c_{L-1} \cdot w_{L-1}} \circ \dots \circ a_1 \circ l_{c_1 \cdot w_1}$ where the internal activation functions $\{a_i\}_{i=1}^{L-1}$ are ReLU or leaky ReLU, and positive constant scalars c_1, \dots, c_L satisfy that $\prod_{i=1}^L c_i = 1$, we have*

$$D_\theta(x) = D'_\theta(x) \quad \forall x \text{ and} \\ \frac{\partial^n D_\theta(x)}{\partial x^n} = \frac{\partial^n D'_\theta(x)}{\partial x^n} \forall x, \forall n \in \mathbb{Z}^+.$$

(Proof in App. B). Given this observation, it is natural to ask if there is any benefit to setting the spectral norms of each layer equal. It turns out that the answer is yes, under some assumptions that appear to approximately hold in practice. Let

$$\mathcal{D} \triangleq \left\{ D_\theta = a_L \circ l_{w_L} \circ \dots \circ a_1 \circ l_{w_1} : \frac{\|\nabla_{w_i} D_\theta(x)\|_F}{\|\nabla_{w_j} D_\theta(x)\|_F} = \frac{\|w_j\|_{sp}}{\|w_i\|_{sp}}, \right. \\ \left. a_i \in \{\text{ReLU, leaky ReLU}\} \forall i, j \in [1, L] \right\}. \quad (2)$$

This intuitively describes the set of all discriminators for which scaling up the weight of one layer proportionally

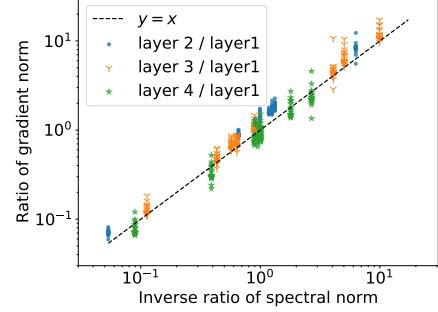


Figure 3. Ratio of gradient norm v.s. inverse ratio of spectral norm in MNIST.

increases the gradient norm of all other layers; the definition of this set is motivated by our upper bound on the gradient norm (App. A). The following theorem shows that when optimizing over set \mathcal{D} , choosing every layer to have the same spectral norm gives the smallest possible gradient norm, for a given set of parameters.

Theorem 1. *Consider a given set of discriminator parameters $\theta = \{w_1, \dots, w_L\}$. For a vector $c = \{c_1, \dots, c_L\}$, we denote $\theta_c \triangleq \{c_t w_t\}_{t=1}^L$. Let $\lambda_\theta = \prod_{i=1}^L \|w_i\|_{sp}^{1/L}$ denote the geometric mean of the spectral norms of the weights. Then we have*

$$\left\{ \frac{\lambda_\theta}{\|w_1\|_{sp}}, \dots, \frac{\lambda_\theta}{\|w_L\|_{sp}} \right\} \\ = \arg \min_{c: D_{\theta_c} \in \mathcal{D}, \prod_{i=1}^L c_i = 1, c_i \in \mathbb{R}^+} \|\nabla_{\theta_c} D_{\theta_c}(x)\|_F$$

(Proof in App. E). The key constraint in this theorem is that we optimize only over discriminators in set \mathcal{D} in Eq. (2). To show that this constraint is realistic (i.e., SN GAN discriminator optimization tends to choose models in \mathcal{D}), we trained a spectrally-normalized GAN with four hidden layers on MNIST, computing the ratios of the gradient norms at each layer and the ratios of the spectral norms, as dictated by Eq. (2). We computed these ratios at different epochs during training, as well as for different randomly-selected rescalings of the spectral normalization vector c . Each point in Fig. 3 represents the results averaged over 64 real samples at a specific epoch of training for a given (random) c . Vertical series of points are from different epochs of the same run, therefore their ratio of spectral norms is the same. The fact that most of the points are near the diagonal line suggests that training naturally favors discriminators that are in or near \mathcal{D} ; we confirm this intuition in other experimental settings in App. K. This observation, combined with Thm. 1, suggests that it is better to force the spectral norms of every layer to be equal. Hence, existing SN implementations (Miyato et al., 2018; Farnia et al., 2018) chose the correct, uniform normalization across layers to upper bound discriminator’s gradients.

4. Vanishing Gradients

An equally troublesome failure mode of GAN training is vanishing gradients (Arjovsky & Bottou, 2017). Prior work has proposed new objective functions to mitigate this problem (Arjovsky & Bottou, 2017; Arjovsky et al., 2017; Gulrajani et al., 2017), but these approaches do not fully solve the problem (see Fig. 11). In this section, we show that SN also helps to control vanishing gradients.

How SN controls vanishing gradients. Gradients tend to vanish for two reasons. First, gradients vanish when the objective function *saturates* (LeCun et al., 1998; Arjovsky & Bottou, 2017), which is often associated with function parameters growing too large. Common loss functions (e.g., hinge loss) and activation functions (e.g., sigmoid, tanh) saturate for inputs of large magnitude. Large parameters tend to amplify the inputs to the activation functions and/or loss function, causing saturation. Second, gradients vanish when function parameters (and hence, internal outputs) grow too small. This is because backpropagated gradients are scaled by the function parameters (App. A).

These insights motivated the LeCun initialization technique (LeCun et al., 1998). The key idea is that to prevent gradients from vanishing, we must ensure that the outputs of each neuron do not vanish or explode. If the inputs to a neural unit are uncorrelated random variables with variance 1, then to ensure that the unit’s output also has variance (approximately) 1, the weight parameters should be zero-mean random variables with variance of $\frac{1}{n_i}$, where n_i denote the fan-in (number of incoming connections) of layer i (LeCun et al., 1998). Hence, LeCun initialization prevents gradient vanishing by controlling the variance of the individual parameters. In the following theorem, we show that SN enforces a similar condition.

Theorem 2 (Parameter variance of SN). *For a matrix $A \in \mathbb{R}^{m \times n}$ with i.i.d. entries a_{ij} from a symmetric distribution with zero mean (e.g., zero-mean Gaussian or uniform), we have*

$$\text{Var}\left(\frac{a_{ij}}{\|A\|_{\text{sp}}}\right) \leq \frac{1}{\max\{m, n\}}. \quad (3)$$

Furthermore, if $m, n \geq 2$ and $\max\{m, n\} \geq 3$, and a_{ij} are from a zero-mean Gaussian, we have

$$\frac{L}{\max\{m, n\} \log(\min\{m, n\})} \leq \text{Var}\left(\frac{a_{ij}}{\|A\|_{\text{sp}}}\right) \leq \frac{1}{\max\{m, n\}},$$

where L is a constant which does not depend on m, n .

(*Proof in App. F*). In other words, spectral normalization forces zero-mean parameters to have a variance that scales inversely with $\max\{m, n\}$. The proof relies on a characterization of extreme values of random vectors drawn uniformly from the surface of a high-dimensional unit ball. Many fully-connected, feed-forward neural networks have a fixed width

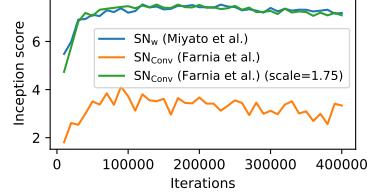


Figure 4. Inception score of different SN variants in CIFAR10.

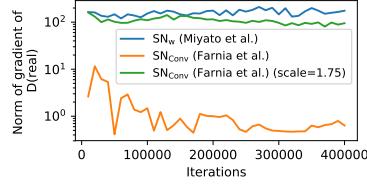


Figure 5. Gradient norms of different SN variants in CIFAR10.

across hidden layers, so $\max\{m, n\}$ corresponds precisely to the fan-in of any neuron in a hidden layer, implying that SN has an effect like LeCun initialization.

Why SN_w works better than SN_{Conv} . In a CNN, the interpretation of $\max\{m, n\}$ depends on how SN is implemented. Recall that the implementation SN_w by (Miyato et al., 2018) does not strictly implement SN, but a variant that normalizes by the spectral norm of $\hat{w}_i = w_i^{c_{out} \times (c_{in} k_w k_h)}$. In architectures like DCGAN (Radford et al., 2015), the larger dimension of \hat{w}_i for *hidden layers* tends to be $c_{in} k_w k_h$, which is exactly the fan-in. This means that SN gets the right variance for hidden layers in CNN.

Perhaps surprisingly, we find empirically that the strict implementation SN_{Conv} of (Farnia et al., 2018) does *not* prevent gradient vanishing. Figs. 4 and 5 shows the gradients of SN_{Conv} vanishing when trained on CIFAR10, leading to a comparatively poor inception score, whereas the gradients of SN_w remain stable. To understand this phenomenon, recall that SN_{Conv} normalizes by the spectral norm of an expanded matrix \tilde{w}_i derived from w_i . Thm. 2 does not hold for \tilde{w}_i since its entries are not i.i.d. (even at initialization); hence it cannot be used to explain this effect. However, Corollary 1 in (Tsuzuku et al., 2018) shows that $\|\hat{w}_i\|_{\text{sp}} \leq \|\tilde{w}_i\|_{\text{sp}} \leq \alpha \|\hat{w}_i\|_{\text{sp}}$, where α is a constant only depends on kernel size, input size, and stride size of the convolution operation. This result has two implications:

(1) $\|\tilde{w}_i\|_{\text{sp}} \leq \alpha \|\hat{w}_i\|_{\text{sp}}$: Although SN_w does not strictly normalize the matrix with the actual spectral norm of the layer, it does upper bound the spectral norm of the layer. Therefore, all our analysis in § 3 still applies for SN_w by changing the spectral norm constant from 1 to $\alpha \|\hat{w}_i\|_{\text{sp}}$. This means that SN_w can still prevent gradient explosion.

(2) $\|\hat{w}_i\|_{\text{sp}} \leq \|\tilde{w}_i\|_{\text{sp}}$: This implies that SN_{Conv} normalizes by a factor that is at least as large as SN_w . In fact, we observe empirically that $\|\tilde{w}_i\|_{\text{sp}}$ is strictly larger than $\|\hat{w}_i\|_{\text{sp}}$ during training (App. L.3). This means that for the same

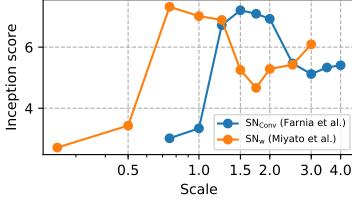


Figure 6. Inception score of scaled SN in CIFAR10.

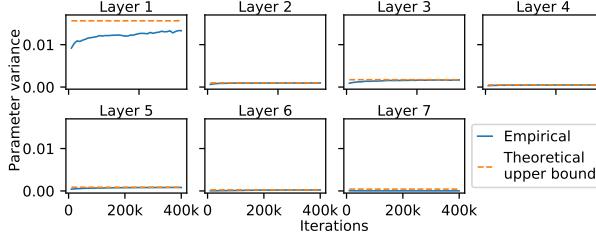


Figure 7. Parameter variances throughout training in CIFAR10. The blue lines show the parameter variances of different layers when SN is applied, and the original line shows our theoretical upper bound given in Eq. (3).

w_i , a discriminator using SN_{Conv} will have smaller outputs than the discriminator using SN_w . We hypothesize that the different scalings explain why SN_{Conv} has vanishing gradients but SN_w does not.

To confirm this hypothesis, for SN_w and SN_{Conv} , we propose to multiply all the normalized weights by a scaling factor s , which is fixed throughout the training. Fig. 6 shows that SN_{Conv} seems to be a shifted version of SN_w . SN_{Conv} with $s = 1.75$ has similar inception score (Fig. 4) to SN_w , as well as similar gradients (Fig. 5) and parameter variances (App. L.4) throughout training. This, combined with Thm. 2, suggests that SN_w inherently finds the correct scaling for the problem, whereas “proper” spectral normalization SN_{Conv} requires additional hyper-parameter tuning.

SN has good parameter variances throughout training. Our theoretical analysis only applies at initialization, when the parameters are selected randomly. However, unlike LeCun initialization which only controls the variance at initialization, we find empirically that Eq. (3) for SN appears to hold *throughout training* (Fig. 7). As a comparison, if trained without SN, the variance increases and the gradient decreases, which makes sample quality bad (App. L.2). This explains why in practice GANs trained with SN are stable *throughout training*.

5. Extensions of Spectral Normalization

Given the above theoretical insights, we propose an extension of spectral normalization called Bidirectional Scaled Spectral Normalization (BSSN). It combines two key ideas: bidirectional normalization and weight scaling.

5.1. Bidirectional Normalization

Glorot and Bengio (Glorot & Bengio, 2010) built on the intuition of LeCun (LeCun et al., 1998) to design an improved initialization, commonly called *Xavier initialization*. Their key observation was that to limit gradient vanishing (and explosion), it is not enough to control only feed-forward outputs; we should also control the variance of backpropagated gradients. Let n_i, m_i denote the fan-in and fan-out of layer i . (In fully-connected layers, $n_i = m_{i-1}$ = the width of layer i .) Whereas LeCun chooses initial parameters with variance $\frac{1}{n_i}$, Glorot and Bengio choose them with variance $\frac{2}{n_i + m_i}$, a compromise between $\frac{1}{n_i}$ (to control output variance) and $\frac{1}{m_i}$ (to control variance of backpropagated gradients).

The first component of BSSN is Bidirectional Spectral Normalization (BSN), which applies a similar intuition to improve the spectral normalization of Miyato *et al.* (Miyato et al., 2018). For fully connected layers, BSN keeps the normalization the same as SN_w (Miyato et al., 2018). For convolution layers, instead of normalizing by $\|w^{c_{out} \times (c_{in} k_w k_h)}\|_{\text{sp}}$, we normalize by $\sigma_w \triangleq \frac{\|w^{c_{out} \times (c_{in} k_w k_h)}\|_{\text{sp}} + \|w^{c_{in} \times (c_{out} k_w k_h)}\|_{\text{sp}}}{2}$, where $\|w^{c_{in} \times (c_{out} k_w k_h)}\|_{\text{sp}}$ is the spectral norm of the reshaped convolution kernel of dimension $c_{in} \times (c_{out} k_w k_h)$. For calculating these two spectral norms, we use the same power iteration method in (Miyato et al., 2018). The following theorem gives the theoretical explanation.

Theorem 3 (Parameter variance of BSN). *For a convolutional kernel $w \in \mathbb{R}^{c_{out} c_{in} k_w k_h}$ with i.i.d. entries w_{ij} from a symmetric distribution with zero mean (e.g. zero-mean Gaussian or uniform) where $k_w k_h \geq \max \left\{ \frac{c_{out}}{c_{in}}, \frac{c_{in}}{c_{out}} \right\}$, and σ_w defined as above, we have*

$$\text{Var} \left(\frac{w_{ij}}{\sigma_w} \right) \leq \frac{2}{c_{in} k_w k_h + c_{out} k_w k_h} .$$

Furthermore, if $c_{in}, c_{out} \geq 2$ and $c_{in} k_w k_h, c_{out} k_w k_h \geq 3$, and w_{ij} are from a zero-mean Gaussian distribution, there exists a constant L that does not depend on $c_{in}, c_{out}, k_w, k_h$ such that

$$\begin{aligned} & \frac{L}{c_{in} k_w k_h \log(c_{out}) + c_{out} k_w k_h \log(c_{in})} \\ & \leq \text{Var} \left(\frac{w_{ij}}{\sigma_w} \right) \leq \frac{2}{c_{in} k_w k_h + c_{out} k_w k_h} . \end{aligned}$$

(*Proof in App. G*). Note that in convolution layers, $n_i = c_{in} k_w k_h$ and $m_i = c_{out} k_w k_h$. Therefore, BSN sets the variance of parameters to scale as $\frac{2}{n_i + m_i}$, as dictated by Xavier initialization. Moreover, BSN naturally inherits the benefits of SN discussed in § 4 (e.g., controlling variance throughout the training).

5.2. Weight Scaling

The second component of BSSN is to multiply all the normalized weights by a constant scaling factor (i.e., as we did in Fig. 6). We call the combination of BSN and this weight scaling *Bidirectional Scaled Spectral Normalization* (BSSN). Note that scaling can also be applied independently to SN, which we call Scaled Spectral Normalization (SSN). The scaling is motivated by the following reasons.

(1) The analysis in LeCun and Xavier initialization assumes that the activation functions are linear, which is not true in practice. More recently, Kaiming initialization was proposed to include the effect of non-linear activations (He et al., 2015). The result is that we should set the variance of parameters to be $2/(1 + a^2)$ times the ones in LeCun or Xavier initialization, where a is the negative slope of leaky ReLU. This suggests the importance of a constant scaling.

(2) However, we found that the scaling constants proposed in LeCun/Kaiming initialization do not always perform well for GANs. Even more surprisingly, there are *multiple modes* of good scaling. Fig. 8 shows the sample quality of LeCun initialization with different scaling on the discriminator. We see that there are at least two good modes of scaling: one at around 0.2 and the other at around 1.2. This phenomenon cannot be explained by the analysis in LeCun/Kaiming initialization.

Recall that SN has similar properties as LeCun initialization (§ 4). Interestingly, we see that SSN also has two good modes of scaling (Fig. 8). Although the best scaling constants for LeCun initialization and SN are very different, there indeed exists an interesting mode correspondence in terms of parameter variances (App. M). We hypothesize that the shift of good scaling from Kaiming initialization we see here could result from adversarial training, and defer the theoretical analysis to future work. These results highlight the need for a separate scaling factor.

(3) The bounds in Thm. 2 and Thm. 3 only imply that in SN and BSN the *order* of parameter variance w.r.t. the network size is correct, but constant scaling is unknown.

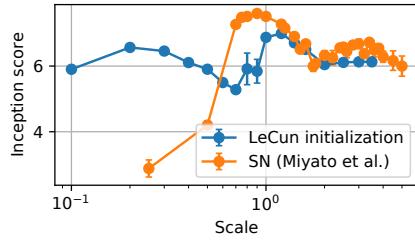


Figure 8. Inception score of SSN and scaled LeCun initialization in CIFAR10. Mean and standard error of the best score during training across multiple runs are shown.

5.3. Results

In this section we verify the effectiveness of BSSN with extensive experiments. The code for reproducing the results is at <https://github.com/fjxmlzn/BSN>.

(Miyato et al., 2018) already compares SN with many other regularization techniques like WGAN-GP (Gulrajani et al., 2017), batch normalization (Ioffe & Szegedy, 2015), layer normalization (Ba et al., 2016), weight normalization (Salimans & Kingma, 2016), and orthogonal regularization (Brock et al., 2016), and SN is shown to outperform them all. Therefore, we focus on comparing the performance of SN with BSSN here. Additionally, to isolate the effects of the two components proposed in BSSN, we include comparison against bidirectional normalization without scaling (BSN) and scaling without bidirectional normalization (SSN).

We conduct experiments across *different datasets* (from low-resolution to high-resolution) and *different network architectures* (from standard CNN to ResNets). More specifically, we conduct experiments on CIFAR10, STL10, CelebA, and ImageNet (ILSVRC2012), following the same settings in (Miyato et al., 2018). All experimental details are attached in Apps. N to S. The results are summarized in Table 1.

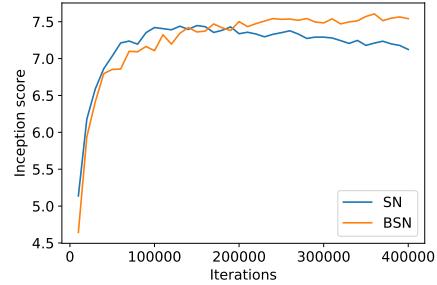


Figure 9. Inception score in CIFAR10. The results are averaged over 5 random seeds, with $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

BSN v.s. SN (showing the effect of bidirectional normalization § 5.1). By comparing BSN with SN in Table 1, we can see that BSN outperforms SN by a large margin in all metrics except in ILSVRC2012 (discussed later).

More importantly, the superiority of BSN is stable across hyper-parameters. In App. N, we vary the learning rates (α_g, α_d) and momentum parameters of generator and discriminator, and the number of discriminator updates per generator update (n_{dis}). We see that BSN consistently outperforms SN in most of the cases.

Moreover, BSN is more stable in the entire training process. We see that as training proceeds, the sample quality of SN often drops, whereas the sample quality of BSN appears to monotonically increase (Fig. 9, more in Apps. P to R). In most cases, BSN not only outperforms SN in final sample quality (i.e., at the end of training), but also in *peak* sample quality. This means that BSN makes the training process more stable, which is the purpose of SN (and BSN).

	CIFAR10		STL10		CelebA	ILSVRC2012	
	IS ↑	FID ↓	IS ↑	FID ↓	FID ↓	IS ↑	FID ↓
Real data	11.26	9.70	26.70	10.17	4.44	197.37	15.62
SN	7.12 ± 0.07	31.43 ± 0.90	9.05 ± 0.05	44.35 ± 0.54	9.43 ± 0.09	12.84 ± 0.33	75.06 ± 2.38
SSN	7.38 ± 0.06	29.31 ± 0.23	9.28 ± 0.03	43.52 ± 0.26	8.50 ± 0.20	12.84 ± 0.33	73.21 ± 1.92
BSN	7.54 ± 0.04	26.94 ± 0.58	9.25 ± 0.01	42.98 ± 0.54	9.05 ± 0.13	1.77 ± 0.13	265.20 ± 19.01
BSSN	7.54 ± 0.04	26.94 ± 0.58	9.25 ± 0.01	42.90 ± 0.17	9.05 ± 0.13	13.23 ± 0.16	69.04 ± 1.46

Table 1. Inception score (IS) and FID on CIFAR10, STL10, CelebA, and ILSVRC2012. The last three rows are proposed in this work, with BSSN representing our final proposal—a combination of BSN and SSN. Each experiment is conducted with 5 random seeds except that the last three rows on ILSVRC2012 is conducted with 3 random seeds. Mean and standard error across these random seeds are reported. We follow the common practice of excluding IS in CelebA as the inception network is pretrained on ImageNet, which is very different from CelebA. The bold font marks the best numbers in that column.

SSN v.s. SN (showing the effect of scaling § 5.2). By comparing SSN with SN in Table 1, we see that scaling consistently improves (or has the same metric) *in all cases*. This verifies our intuition in § 5.2 that the inherent scaling in SN is not optimal, and a extra constant scaling is needed to get the best results.

BSSN v.s. BSN (showing the effect of scaling § 5.2). By comparing BSSN with BSN in Table 1, we see that in some cases the optimal scale of BSN happens to be 1 (e.g., in CIFAR10), but in other cases, scaling is critical. For example, in ILSVRC2012, BSN without any scaling has the same gradient vanishing problem we observe for SN_{Conv} (Farnia et al., 2018) in § 4, which causes bad sample quality. BSSN successfully solves the gradient vanishing problem and achieves the best sample quality.

Additional results. Because of the space constraints, we defer other results (e.g., generated images, training curves, more comparisons and analysis) to Apps. N to S.

Summary. In summary, both designs we proposed can effectively stabilize training and achieve better sample quality. Combining them together, BSSN achieves the best sample quality in most of the cases. This demonstrates the practical value of the theoretical insights in § 3 and 4.

6. Discussion

Other reasons contributing to the stability of SN. In the paper we present one possible reason (i.e., SN avoids exploding and vanishing gradients), and show such correlation through extensive theoretical and empirical analysis. However, there could exists many other parallel factors. For example, SN paper (Miyato et al., 2018) points out that SN could speed up training by encouraging the weights to be updated along directions orthogonal to itself. This is orthogonal to the reasons we discuss in the paper.

Related work. A related result to our upper bound was shown in (Santurkar et al., 2018), which shows that batch normalization (BN) makes the scaling of the Hessian along

the direction of the gradient smaller, thereby making gradients more predictive. Given Prop. 1, we can apply the reasoning from (Santurkar et al., 2018) to explain why spectrally-normalized GANs are robust to different learning rates as shown in (Miyato et al., 2018). However, our insights regarding the gradient vanishing problem are the more surprising result; this notion is not discussed in (Santurkar et al., 2018). An interesting question for future work is whether BN similarly controls vanishing gradients.

In parallel to this work, some other approaches have been proposed to improve SN. For example, (Fang et al., 2021) finds out that even with SN, the condition numbers of the weights can still be large, which causes the instability. To solve the issue, they borrow the insights from linear algebra and propose precondition layers to improve the condition numbers and therefore promote stability.

Future directions. Our results suggest that SN stabilizes GANs by controlling exploding and vanishing gradients in the discriminator. However, our analysis also applies to the training of any feed-forward neural network. This connection partially explains why SN helps train generators as well as discriminators (Zhang et al., 2018; Brock et al., 2018), and why SN is more generally useful in training neural networks (Farnia et al., 2018; Gouk et al., 2018; Yoshida & Miyato, 2017). We focus on GANs in this paper because SN seems to have a disproportionately beneficial effect on GANs (Miyato et al., 2018). Formally extending this analysis to understand the effects of adversarial training is an interesting direction for future work.

Related to the weight initialization and training dynamics, a series of work (Pennington et al., 2017; Saxe et al., 2013) has shown that Gaussian weights or ReLU activations cannot achieve dynamical isometry (all singular values of the network Jacobian are near 1), a desired property for training stability. Orthogonal weight initialization may be better at achieving the goal. In this paper, we focus the theoretical analysis and experiments on Gaussian weights and ReLU activations as they are the predominant implementations

in GANs. We defer the study of other networks to future work.

Acknowledgements

This work was supported in part by faculty research awards from Google, JP Morgan Chase, and the Sloan Foundation, as well as a gift from Siemens AG. This research was sponsored in part by National Science Foundation Convergence Accelerator award 2040675 and the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system (Nystrom et al., 2015), which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

References

- Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks, 2017.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.
- Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.
- Bernstein, J., Vahdat, A., Yue, Y., and Liu, M.-Y. On the distance between two neural networks and the stability of learning. *arXiv preprint arXiv:2002.03432*, 2020.
- Brock, A., Lim, T., Ritchie, J. M., and Weston, N. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.
- Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.
- Fang, T., Schwing, A., and Sun, R. Precondition layer and its use for {gan}s, 2021. URL <https://openreview.net/forum?id=1yXhko8GZEE>.
- Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.
- Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gouk, H., Frank, E., Pfahringer, B., and Cree, M. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Jolicoeur-Martineau, A. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL <http://yann.lecun.com/exdb/mnist/>.
- LeCun, Y., Bottou, L., Orr, G. B., and Müller, K. R. *Efficient BackProp*, pp. 9–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-49430-0. doi: 10.1007/3-540-49430-8_2. URL https://doi.org/10.1007/3-540-49430-8_2.
- Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- Lin, Z., Thekumparampil, K. K., Fanti, G., and Oh, S. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *ICML*, 2020.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Miyato, T. and Koyama, M. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.
- Nystrom, N. A., Levine, M. J., Roskies, R. Z., and Scott, J. R. Bridges: A uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, XSEDE ’15*, pp. 30:1–30:8, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3720-5. doi: 10.1145/2792745.2792775. URL <http://doi.acm.org/10.1145/2792745.2792775>.
- Odena, A., Buckman, J., Olsson, C., Brown, T., Olah, C., Raffel, C., and Goodfellow, I. Is generator conditioning causally related to gan performance? In *International Conference on Machine Learning*, pp. 3849–3858, 2018.
- Pascanu, R., Mikolov, T., and Bengio, Y. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2:417, 2012.
- Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318, 2013.
- Pennington, J., Schoenholz, S. S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *arXiv preprint arXiv:1711.04735*, 2017.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pp. 901–909, 2016.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.
- Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pp. 2483–2493, 2018.
- Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.
- Seginer, Y. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.
- Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., and Wilkins-Diehr, N. Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74, 2014.
- Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6541–6550, 2018.
- Warde-Farley, D. and Bengio, Y. Improving generative adversarial networks with denoising feature matching. 2016.

Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

Yoshida, Y. and Miyato, T. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

A. Proof of Prop. 1

The proposition makes use of the following observation: For the discriminator defined in (1), the norm of gradient for w_t is upper bounded by

$$\|\nabla_{w_t} D_\theta(x)\|_F \leq \|x\| \cdot \prod_{i=1}^L \|a_i\|_{\text{Lip}} \cdot \prod_{i=1}^L \|w_i\|_{\text{sp}} / \|w_t\|_{\text{sp}} \quad \text{for } \forall t \in [1, L] \quad (4)$$

To prove this, for simplicity of notation, let $o_a^i = a_i \circ l_{w_i} \circ \dots \circ a_1 \circ l_{w_1}$, and $o_l^i = l_{w_i} \circ a_{i-1} \circ \dots \circ a_1 \circ l_{w_1}$.

It is straightforward to show that the norm of each internal output of discriminator is bounded by

$$\|o_a^t(x)\| \leq \|x\| \cdot \prod_{i=1}^t \|a_i\|_{\text{Lip}} \cdot \prod_{i=1}^t \|w_i\|_{\text{sp}} \quad (5)$$

and

$$\|o_l^t(x)\| \leq \|x\| \cdot \prod_{i=1}^{t-1} \|a_i\|_{\text{Lip}} \cdot \prod_{i=1}^t \|w_i\|_{\text{sp}}. \quad (6)$$

This holds because

$$\|o_a^t(x)\| = \|a_i(o_l^t(x))\| \leq \|a_i\|_{\text{Lip}} \cdot \|o_l^t(x)\|$$

and

$$\|o_l^t(x)\| = \|l_{w_i}(o_a^{t-1}(x))\| \leq \|w_t\|_{\text{sp}} \cdot \|o_a^{t-1}(x)\|,$$

from which we can show the desired inequalities by induction.

Next, we observe that the norm of each internal gradient is bounded by

$$\|\nabla_{o_a^t(x)} D_\theta(x)\| \leq \prod_{i=t+1}^L \|a_i\|_{\text{Lip}} \cdot \prod_{i=t+1}^L \|w_i\|_{\text{sp}} \quad (7)$$

and

$$\|\nabla_{o_l^t(x)} D_\theta(x)\| \leq \prod_{i=t}^L \|a_i\|_{\text{Lip}} \cdot \prod_{i=t+1}^L \|w_i\|_{\text{sp}}. \quad (8)$$

This holds because

$$\|\nabla_{o_a^t(x)} D_\theta(x)\| = \|w_{t+1}^T \nabla_{o_l^{t+1}(x)} D_\theta(x)\| \leq \|w_{t+1}\|_{\text{sp}} \|\nabla_{o_l^{t+1}(x)} D_\theta(x)\|$$

and

$$\|\nabla_{o_l^t(x)} D_\theta(x)\| = \left\| \langle \nabla_{o_a^t(x)} D_\theta(x), [a'_t(x)|_{x=o_l^t(x)}] \rangle \right\| \leq \|a_t\|_{\text{Lip}} \|\nabla_{o_a^t(x)} D_\theta(x)\|,$$

from which we can show inequalities Eqs. (7) and (8) by induction.

Now we have that

$$\begin{aligned} \|\nabla_{w_t} D_\theta(x)\|_F &= \left\| \nabla_{o_l^t(x)} D_\theta(x) \cdot (o_a^{t-1}(x))^T \right\|_F \\ &= \|\nabla_{o_l^t(x)} D_\theta(x)\| \cdot \|o_a^{t-1}(x)\| \\ &\leq \prod_{i=t}^L \|a_i\|_{\text{Lip}} \cdot \prod_{i=t+1}^L \|w_i\|_{\text{sp}} \cdot \|x\| \cdot \prod_{i=1}^{t-1} \|a_i\|_{\text{Lip}} \cdot \prod_{i=1}^{t-1} \|w_i\|_{\text{sp}} \\ &= \|x\| \cdot \prod_{i=1}^L \|a_i\|_{\text{Lip}} \cdot \prod_{i=1}^L \|w_i\|_{\text{sp}} / \|w_t\|_{\text{sp}} \end{aligned}$$

where we use Eqs. (5) to (8) at the inequality. The upper bound of gradient's Frobenius norm for spectrally-normalized discriminators follows directly.

B. Proof of Prop. 2

Proof. As $l_w(x)$ is a linear transformation, we have $l_{cw}(x) = c \cdot l_w(x)$, and $l_w(cx) = c \cdot l_w(x)$. Moreover, since ReLU and leaky ReLU is linear in \mathbb{R}^+ and \mathbb{R}^- region, we have $a_i(cx) = c \cdot a_i(x)$. Therefore, we have

$$\begin{aligned} D'_\theta(x) &= (a_L \circ l_{c_L \cdot w_L} \circ a_{L-1} \circ l_{c_{L-1} \cdot w_{L-1}} \circ \dots \circ a_1 \circ l_{c_1 \cdot w_1})(x) \\ &= \prod_{i=1}^L c_i \cdot (a_L \circ l_{w_L} \circ a_{L-1} \circ l_{w_{L-1}} \circ \dots \circ a_1 \circ l_{w_1})(x) \\ &= D_\theta(x) \end{aligned}$$

□

C. Additional Analysis of Gradient

In § 3, we discuss the gradients with respect to $w'_i = \frac{w_i}{u_i^T w_i v_i}$, where u_i, v_i are the singular vectors corresponding to the largest singular values. In this section we discuss the gradients with respect the actual parameter w_i . From Eq. (12) in (Miyato et al., 2018) we know

$$\nabla_{w_t} D_\theta(x) = \frac{1}{\|w_t\|_{sp}} \left(\nabla_{w'_t} D_\theta(x) - \left(\left(\nabla_{o_l^t(x)} D_\theta(x) \right)^T o_l^t(x) \right) \cdot u_t v_t^T \right)$$

From App. A, we know that $\|\nabla_{w'_t} D_\theta(x)\|_F$, $\|\nabla_{o_l^t(x)} D_\theta(x)\|$, and $\|o_l^t(x)\|$ have upper bounds. Furthermore, $\|u_t v_t^T\|_F = 1$. Therefore, $\left\| \nabla_{w'_t} D_\theta(x) - \left(\left(\nabla_{o_l^t(x)} D_\theta(x) \right)^T o_l^t(x) \right) \cdot u_t v_t^T \right\|_F$ has an upper bound. From Theorem 1.1 in (Seginer, 2000) we know that if w_t is initialized with i.i.d random variables from uniform or Gaussian distribution, $\mathbb{E}(\|w_t\|_{sp})$ is lower bounded away from zero at initialization. So $\|\nabla_{w_t} D_\theta(x)\|_F$ is upper bounded at initialization. Moreover, we observe empirically that $\|w_t\|_{sp}$ is usually increasing during training. Therefore, $\|\nabla_{w_t} D_\theta(x)\|_F$ is typically upper bounded during training as well.

D. Analysis of Hessian

The following proposition states that spectral normalization also gives an upper bound on $\|H_{w_i}(D_\theta)(x)\|_{sp}$ for networks with ReLU or leaky ReLU internal activations.

Proposition 3 (Upper bound of Hessian's spectral norm). *Consider the discriminator defined in Eq. (1). Let $H_{w_i}(D_\theta)(x)$ denote the Hessian of D_θ at x with respect with the vector form of w_i . If the internal activations are ReLU or leaky ReLU, the spectral norm of $H_{w_i}(D_\theta)(x)$ is upper bounded by*

$$\|H_{w_i}(D_\theta)(x)\|_{sp} \leq \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \|x\|^2 \cdot \prod_{i=1}^L \|w_i\|_{sp}^2 / \|w_t\|_{sp}^2$$

The proof is in App. D.1. Following Prop. 3, we can easily show the upper bound of Hessian's spectral norm for spectral normalized discriminators.

Corollary 1 (Upper bound of Hessian's spectral norm for spectral normalization). *If the internal activations are ReLU or leaky ReLU, and $\|w_i\|_{sp} \leq 1$ for all $i \in [1, L]$, then*

$$\|H_{w_i}(D_\theta)(x)\|_{sp} \leq \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \|x\|^2 .$$

Moreover, if the activation for the last layer is sigmoid (e.g., for vanilla GAN (Goodfellow et al., 2014)), we have

$$\|H_{w_i}(D_\theta)(x)\|_{sp} \leq 0.1 \|x\|^2 ;$$

if the activation function for the last layer is identity (e.g., for WGAN-GP (Gulrajani et al., 2017)), we have

$$\|H_\theta(D_\theta)(x)\|_{sp} = 0 .$$

D.1. Proof of Prop. 3

Lemma 1. *The spectral norm of each internal Hessian is bounded by*

$$\left\| H_{o_a^t(x)} D_\theta(x) \right\|_{sp} \leq \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \prod_{i=t+1}^L \|w_i\|_{sp}^2$$

and

$$\left\| H_{o_l^t(x)} D_\theta(x) \right\|_{sp} \leq \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \prod_{i=t+1}^L \|w_i\|_{sp}^2$$

Proof. We have

$$\begin{aligned} \left\| H_{o_a^t(x)} D_\theta(x) \right\|_{sp} &= \left\| w_{t+1}^T \cdot \nabla_{a_l^{t+1}(x)} D_\theta(x) \cdot w_{t+1} \right\|_{sp} \\ &\leq \left\| \nabla_{a_l^{t+1}(x)} D_\theta(x) \right\|_{sp} \|w_{t+1}\|_{sp}^2. \end{aligned}$$

We also have

$$\begin{aligned} \left\| H_{o_l^t(x)} D_\theta(x) \right\|_{sp} &= \left\| \text{diag}([a'_t(x)]_{x=o_a^t(x)}) \cdot H_{o_a^{t+1}(x)} D_\theta(x) \cdot \text{diag}([a'_t(x)]_{x=o_a^t(x)}) \right\|_{sp} \\ &\leq \left\| H_{o_a^{t+1}(x)} D_\theta(x) \right\|_{sp} \end{aligned}$$

where we use the property that ReLU or leaky ReLU is piece-wise linear. The desired inequalities then follow by induction. \square

Now let's come back to the proof for Prop. 3.

Proof. We have

$$\frac{\partial D_\theta}{\partial (w_t)_{ij} \partial (w_t)_{kl}} = \left(H_{o_l^t}(D_\theta)(x) \right)_{ik} \cdot (o_a^{t-1}(x))_j \cdot (o_a^{t-1}(x))_l.$$

Therefore,

$$\|H_{w_i}(D_\theta)(x)\|_{sp} \leq \left\| H_{o_l^t}(D_\theta)(x) \right\|_{sp} \|o_a^{t-1}(x)\|_\infty^2 \leq \left\| H_{o_l^t}(D_\theta)(x) \right\|_{sp} \|o_a^{t-1}(x)\|^2$$

Applying Eq. (5) and Lemma 1 we get

$$\begin{aligned} \|H_{w_i}(D_\theta)(x)\|_{sp} &\leq \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \prod_{i=t+1}^L \|w_i\|_{sp}^2 \cdot \|x\|^2 \cdot \prod_{i=1}^{t-1} \|w_i\|_{sp}^2 \\ &= \left\| H_{o_l^L(x)} D_\theta(x) \right\|_{sp} \cdot \|x\|^2 \cdot \prod_{i=1}^L \|w_i\|_{sp}^2 / \|w_t\|_{sp}^2 \end{aligned}$$

\square

E. Proof of Thm. 1

Proof. For any discriminator $D_\theta = a_L \circ l_{w_L} \circ a_{L-1} \circ l_{w_{L-1}} \circ \dots \circ a_1 \circ l_{w_1}$, consider $\theta' = \left\{ w'_t \triangleq c_t w_t \right\}_{t=1}^L$ with the constraint $\prod_{i=1}^L c_i = 1$ and $c_i \in \mathbb{R}^+$. Let $Q = \|\nabla_{w'_i} D_{\theta'}(x)\|_{\text{F}} \|w'_i\|_{\text{sp}}$. We have

$$\begin{aligned} \|\nabla_{\theta'} D_{\theta'}(x)\|_{\text{F}} &= \sqrt{\sum_{i=1}^L \|\nabla_{w'_i} D_{\theta'}(x)\|_{\text{F}}^2} \\ &= \sqrt{\sum_{i=1}^L \frac{Q^2}{c_i^2 \|w_i\|_{\text{sp}}^2}} \\ &\geq \sqrt{L \left(\prod_{i=1}^L \frac{Q^2}{c_i^2 \|w_i\|_{\text{sp}}^2} \right)^{1/L}} \\ &= \sqrt{L} \cdot Q^{1/L} \cdot \left(\prod_{i=1}^L \|w_i\|_{\text{sp}} \right)^{-1/L} \end{aligned}$$

and the equality is achieved iff $c_i^2 \|w_i\|_{\text{sp}}^2 = c_j^2 \|w_j\|_{\text{sp}}^2, \forall i, j \in [1, L]$ according to AM-GM inequality. When $c_i^2 \|w_i\|_{\text{sp}}^2 = c_j^2 \|w_j\|_{\text{sp}}^2, \forall i, j \in [1, L]$, we have $c_t = \prod_{i=1}^L \|w_i\|_{\text{sp}}^{1/L} / \|w_t\|_{\text{sp}}$. \square

F. Proof of Thm. 2

Proof. Since a_{ij} are symmetric random variables, we know $\mathbb{E} \left(\frac{a_{ij}}{\|A\|_{\text{sp}}} \right) = 0$. Further, by symmetry, we have that for any $(i, j) \neq (h, \ell)$, $\mathbb{E} \left(\frac{a_{ij}^2}{\|A\|_{\text{sp}}^2} \right) = \mathbb{E} \left(\frac{a_{h\ell}^2}{\|A\|_{\text{sp}}^2} \right)$. Therefore, we have

$$\text{Var} \left(\frac{a_{ij}}{\|A\|_{\text{sp}}} \right) = \mathbb{E} \left(\frac{a_{ij}^2}{\|A\|_{\text{sp}}^2} \right) = \frac{1}{mn} \cdot \mathbb{E} \left(\frac{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}{\|A\|_{\text{sp}}^2} \right) = \frac{1}{mn} \cdot \mathbb{E} \left(\frac{\|A\|_{\text{F}}^2}{\|A\|_{\text{sp}}^2} \right)$$

Our approach will be to upper and lower bound the quantity $\frac{1}{mn} \cdot \mathbb{E} \left(\frac{\|A\|_{\text{F}}^2}{\|A\|_{\text{sp}}^2} \right)$.

Upper bound Assume the singular values of A are $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min\{m,n\}}$. We have

$$\frac{1}{mn} \cdot \mathbb{E} \left(\frac{\|A\|_{\text{F}}^2}{\|A\|_{\text{sp}}^2} \right) = \frac{1}{mn} \cdot \mathbb{E} \left(\frac{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}{\sigma_1^2} \right) \leq \frac{\min\{m, n\}}{mn} = \frac{1}{\max\{m, n\}},$$

which gives the desired upper bound.

Lower bound Now for the lower bound, if a_{ij} are drawn from zero-mean Gaussian distribution and $\max\{m, n\} \geq 3$, we have

$$\frac{1}{mn} \cdot \mathbb{E} \left(\frac{\|A\|_{\text{F}}^2}{\|A\|_{\text{sp}}^2} \right) \tag{9}$$

$$\begin{aligned} &= \frac{1}{mn} \cdot \mathbb{E} \left(\frac{1}{\|A\|_{\text{sp}}^2 / \|A\|_{\text{F}}^2} \right) \\ &\geq \frac{1}{mn} \cdot \frac{1}{\mathbb{E} \left(\left(\frac{\|A\|_{\text{F}}}{\|A\|_{\text{sp}}} \right)^2 \right)} \\ &= \frac{1}{mn} \cdot \frac{1}{\mathbb{E} \left(\|B\|_{\text{sp}}^2 \right)} \tag{10} \end{aligned}$$

where $B \in R^{m \times n}$ is uniformly sampled from the sphere of $m \times n$ -dimension unit ball. We use the following lemma to lower bound (10).

Lemma 2 (Theorem 1.1 in (Seginer, 2000)). *Assume $A \in R^{m \times n}$ is uniformly sampled from the sphere of $m \times n$ -dimension unit ball. When $\max\{m, n\} \geq 3$, we have*

$$\mathbb{E}\left(\|A\|_{sp}^2\right) \leq K^2 \left(\mathbb{E}\left(\max_{1 \leq i \leq m} \|a_{i\bullet}\|^2\right) + \mathbb{E}\left(\max_{1 \leq j \leq n} \|a_{\bullet j}\|^2\right) \right),$$

where K is a constant which does not depend on m, n . Here $a_{i\bullet}$ denotes the i -th row of A , and $a_{\bullet j}$ denotes the j -th column of A .¹

We thus have that

$$\frac{1}{mn} \cdot \frac{1}{\mathbb{E}\left(\|B\|_{sp}^2\right)} \geq \frac{1}{mn} \cdot \frac{1}{K^2 \left(\mathbb{E}\left(\max_{1 \leq i \leq m} \|b_{i\bullet}\|^2\right) + \mathbb{E}\left(\max_{1 \leq j \leq n} \|b_{\bullet j}\|^2\right) \right)}.$$

Hence, we need to upper bound $\mathbb{E}\left(\max_{1 \leq i \leq m} \|b_{i\bullet}\|^2\right)$ and $\mathbb{E}\left(\max_{1 \leq j \leq n} \|b_{\bullet j}\|^2\right)$. Let $z \in \mathbb{R}^m$ be a vector uniformly sampled from the sphere of m -dimension unit ball. Observe that $z \stackrel{d}{=} [\|b_{1\bullet}\|, \dots, \|b_{m\bullet}\|]$. The following lemma upper bounds the square of the infinity norm of this vector.

Lemma 3. *Assume $z = [z_1, z_2, \dots, z_n]$ is uniformly sampled from the sphere of n -dimension unit ball, where $n \geq 2$. Then we have*

$$\mathbb{E}\left(\max_{1 \leq i \leq n} z_i^2\right) \leq \frac{4 \log(n)}{n-1}.$$

(Proof in App. F.1)

Hence, when $m, n \geq 2$, we have

$$\mathbb{E}\left(\max_{1 \leq i \leq m} \|b_{i\bullet}\|^2\right) \leq \frac{4 \log(m)}{m-1}$$

Similarly, we have

$$\mathbb{E}\left(\max_{1 \leq j \leq n} \|b_{\bullet j}\|^2\right) \leq \frac{4 \log(n)}{n-1}$$

Therefore,

$$\begin{aligned} & \text{Var}\left(\frac{a_{ij}}{\|A\|_{sp}}\right) \\ & \geq \frac{1}{mn} \cdot \frac{1}{K^2 \left(\frac{4 \log(m)}{m-1} + \frac{4 \log(n)}{n-1} \right)} \\ & \geq \frac{1}{8K^2} \cdot \frac{1}{n \log(m) + m \log(n)} \\ & \geq \frac{1}{16K^2} \cdot \frac{1}{\max\{m, n\} \log(\min\{m, n\})} \end{aligned}$$

which gives the result. □

¹Note that the original theorem in (Seginer, 2000) requires that the entries of A be i.i.d. symmetric random variables, whereas in our case the entries are not i.i.d., as we require $\|A\|_F = 1$. However, the i.i.d. assumption in their proof is only used to ensure that A , $S_{\sigma^{(1)}, \epsilon^{(1)}}(A)$, and $S_{\sigma^{(2)}, \epsilon^{(2)}}(A)$ have the same distribution, where $\sigma^{(t)}$ for $t = 0, 1$ are vectors of independent random permutations; $\epsilon^{(t)}$ for $t = 0, 1$ are matrices of i.i.d. random variables with equal probability of being ± 1 ; and $S_{\sigma^{(1)}, \epsilon^{(1)}}(A) = \left(\epsilon_{ij}^{(1)} \cdot a_{i, \sigma_i^{(1)}(j)} \right)_{i,j}$ and $S_{\sigma^{(2)}, \epsilon^{(2)}}(A) = \left(\epsilon_{ij}^{(2)} \cdot a_{\sigma_j^{(2)}(i), j} \right)_{i,j}$. Our matrix A satisfies this requirement, and therefore the same theorem holds.

F.1. Proof of Lemma 3

Proof.

$$\begin{aligned}
 & \mathbb{E} \left(\max_{1 \leq i \leq n} z_i^2 \right) \\
 &= \int_0^1 \mathbb{P} \left(\max_{1 \leq i \leq n} z_i^2 \geq \delta \right) d\delta \\
 &\leq \int_0^1 \min \left\{ 1, n \cdot \mathbb{P} (z_1^2 \geq \delta) \right\} d\delta
 \end{aligned} \tag{11}$$

where (11) follows from the union bound. Next, we use the following lemma to upper bound $\mathbb{P} (z_1^2 \geq \delta)$.

Lemma 4. Assume $z = [z_1, z_2, \dots, z_n]$ is uniformly sampled from the sphere of n -dimension unit ball, where $n \geq 2$. Then for $\frac{1}{n} \leq \delta < 1$ and $\forall i \in [1, n]$, we have

$$\mathbb{P} (z_i^2 \geq \delta) \leq e^{-\frac{n-1}{2} \cdot \delta + 1}.$$

(Proof in App. F.2). This in turn gives

$$\begin{aligned}
 \int_0^1 \min \left\{ 1, n \cdot \mathbb{P} (z_1^2 \geq \delta) \right\} d\delta &\leq \int_0^{\min \left\{ 1, \frac{2 \log(n)+2}{n-1} \right\}} 1 \cdot d\delta + \int_{\min \left\{ 1, \frac{2 \log(n)+2}{n-1} \right\}}^1 n \cdot e^{-\frac{n-1}{2} \cdot \delta + 1} d\delta \\
 &\leq \begin{cases} \frac{1}{2 \log(n)+2} - \frac{2n}{n-1} e^{-\frac{n-3}{2}} + \frac{2}{n-1} & (n \leq 6) \\ \frac{4 \log(n)}{n-1} & (n \geq 7) \end{cases} \\
 &\leq \frac{4 \log(n)}{n-1}
 \end{aligned} \tag{12}$$

where Eq. (12) follows from Lemma 4.

□

F.2. Proof of Lemma 4

Proof. Due to the symmetry of z_i , we only need to prove the inequality for $i = 1$ case. Let $x = [x_1, \dots, x_n] \sim \mathcal{N}(\mathbf{0}, I_n)$, where I_n is the identity matrix in n dimension. We know that $\frac{x_1^2}{\sum_{i=1}^n x_i^2} \stackrel{d}{=} z_1^2$. Therefore, we have

$$\mathbb{P} (z_1^2 \geq \delta) = \mathbb{P} \left(\frac{x_1^2}{\sum_{j=1}^n x_j^2} \geq \delta \right) = \mathbb{P} \left(\frac{x_1^2}{(\sum_{i=2}^n x_i^2)/(n-1)} \geq \frac{(n-1)\delta}{1-\delta} \right).$$

Note that x_1^2 and $\sum_{i=2}^n x_i^2$ are two independent chi-squared random variables, therefore, we know that $\frac{x_1^2}{(\sum_{i=2}^n x_i^2)/(n-1)} \sim F(1, n-1)$, where F denotes the central F-distribution. Therefore,

$$\begin{aligned}
 \mathbb{P} \left(\frac{x_1^2}{(\sum_{i=2}^n x_i^2)/(n-1)} \geq \frac{(n-1)\delta}{1-\delta} \right) &= 1 - I_\delta \left(\frac{1}{2}, \frac{n-1}{2} \right) \\
 &= I_{1-\delta} \left(\frac{n-1}{2}, \frac{1}{2} \right) \\
 &= \frac{B_{1-\delta} \left(\frac{n-1}{2}, \frac{1}{2} \right)}{B \left(\frac{n-1}{2}, \frac{1}{2} \right)},
 \end{aligned} \tag{13}$$

where $I_x(a, b)$ is the regularized incomplete beta function, $B_x(a, b)$ is the incomplete beta function, and $B(a, b)$ is beta function.

For the ease of computation, we take the log of Eq. (13). The numerator gives

$$\begin{aligned} & \log \left(B_{1-\delta} \left(\frac{n-1}{2}, \frac{1}{2} \right) \right) \\ &= \log \left(\frac{(1-\delta)^{(n-1)/2}}{(n-1)/2} {}_2F_1 \left(\frac{n-1}{2}, \frac{1}{2}; \frac{n+1}{2}; 1-\delta \right) \right) \\ &= \frac{n-1}{2} \log(1-\delta) - \log(n-1) + \log \left({}_2F_1 \left(\frac{n-1}{2}, \frac{1}{2}; \frac{n+1}{2}; 1-\delta \right) \right) + \log(2) , \end{aligned} \quad (14)$$

where ${}_2F_1(\cdot)$ is the hypergeometric function. Let $(q)_i = \begin{cases} 1 & (i=0) \\ q(q+1)\dots(q+i-1) & (i>0) \end{cases}$, we have

$$\begin{aligned} & {}_2F_1 \left(\frac{n-1}{2}, \frac{1}{2}; \frac{n+1}{2}; 1-\delta \right) \\ &= \sum_{i=0}^{\infty} \frac{\left(\frac{n-1}{2}\right)_i \left(\frac{1}{2}\right)_i (1-\delta)^i}{\left(\frac{n+1}{2}\right)_i \cdot i!} \\ &\leq \sum_{i=0}^{\infty} \frac{\left(\frac{1}{2}\right)_i (1-\delta)^i}{i!} \\ &= \delta^{-\frac{1}{2}} \end{aligned} \quad (15)$$

Substituting it into Eq. (14) gives

$$\log \left(B_{1-\delta} \left(\frac{n-1}{2}, \frac{1}{2} \right) \right) \leq \frac{n-1}{2} \log(1-\delta) - \log(n-1) - \frac{1}{2} \log(\delta) + \log(2) . \quad (16)$$

The log of the denominator of (13) is

$$\begin{aligned} & \log \left(B \left(\frac{n-1}{2}, \frac{1}{2} \right) \right) \\ &= \log \left(\frac{\Gamma(\frac{n-1}{2}) \Gamma(\frac{1}{2})}{\Gamma(\frac{n}{2})} \right) \\ &\geq \log \left(\sqrt{\pi} \cdot \left(\frac{n+1}{2} \right)^{-\frac{1}{2}} \right) \\ &= -\frac{1}{2} \log(n+1) + \frac{1}{2} \log(2) + \frac{1}{2} \log(\pi) . \end{aligned} \quad (17)$$

where Γ denotes the Gamma function and we use the Gautschi's inequality: $\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} < (x+1)^{\frac{1}{2}}$ for positive real number x .

Combining Eq. (13), Eq. (16), and Eq. (17) we get

$$\begin{aligned} & \log \left(\mathbb{P} \left(\frac{x_1^2}{(\sum_{i=2}^n x_i^2)/(n-1)} \geq \frac{(n-1)\delta}{1-\delta} \right) \right) \\ &\leq \frac{n-1}{2} \log(1-\delta) - \log(n-1) + \frac{1}{2} \log(n+1) - \frac{1}{2} \log(\delta) + \frac{1}{2} \log(2/\pi) \\ &\leq \frac{n-1}{2} \log(1-\delta) - \frac{1}{2} \log(n-1) - \frac{1}{2} \log(\delta) + \frac{1}{2} \log(6/\pi) \\ &\leq \frac{n-1}{2} \log(1-\delta) - \frac{1}{2} \log \left(\frac{n-1}{n} \right) + \frac{1}{2} \log(6/\pi) \\ &\leq \frac{n-1}{2} \log(1-\delta) + \frac{1}{2} \log \frac{12}{\pi} \\ &\leq -\frac{n-1}{2} \cdot \delta + 1 \end{aligned}$$

Therefore, we have

$$\mathbb{P}(z_1^2 \geq \delta) \leq e^{-\frac{n-1}{2} \cdot \delta + 1}$$

□

G. Proof of Thm. 3

Proof. Let $s_w = c_{in}c_{out}k_wk_h$. Since w_{ij} are symmetric random variables, we know $\mathbb{E}\left(\frac{w_{ij}}{\sigma_w}\right) = 0$. Therefore, we have

$$\text{Var}\left(\frac{w_{ij}}{\sigma_w}\right) = \mathbb{E}\left(\frac{w_{ij}^2}{\sigma_w^2}\right) = \frac{1}{s_w} \cdot \mathbb{E}\left(\frac{\sum_{i=1}^m \sum_{j=1}^n w_{ij}^2}{\sigma_w^2}\right) = \frac{1}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\sigma_w^2}\right)$$

Note that

$$\begin{aligned} \frac{1}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\sigma_w^2}\right) &\in \left[\frac{2}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\|w^{c_{out} \times (c_{in}k_wk_h)}\|_{sp}^2 + \|w^{c_{in} \times (c_{out}k_wk_h)}\|_{sp}^2}\right), \right. \\ &\quad \left. \frac{4}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\|w^{c_{out} \times (c_{in}k_wk_h)}\|_{sp}^2 + \|w^{c_{in} \times (c_{out}k_wk_h)}\|_{sp}^2}\right) \right]. \end{aligned}$$

Assume the singular values of $w^{c_{out} \times (c_{in}k_wk_h)}$ are $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{c_{out}}$, and the singular values of $w^{c_{in} \times (c_{out}k_wk_h)}$ are $\sigma'_1 \geq \sigma'_2 \geq \dots \geq \sigma'_{c_{in}}$. We have

$$\begin{aligned} &\frac{4}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\|w^{c_{out} \times (c_{in}k_wk_h)}\|_{sp}^2 + \|w^{c_{in} \times (c_{out}k_wk_h)}\|_{sp}^2}\right) \\ &= \frac{4}{s_w} \cdot \mathbb{E}\left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{c_{out}} \sigma_i^2}{\sigma_1^2} + \frac{1}{2} \cdot \frac{\sum_{i=1}^{c_{in}} \sigma'^2_i}{\sigma'_1^2}\right) \leq \frac{2(c_{out} + c_{in})}{s_w} = \frac{2}{c_{in}k_wk_h + c_{out}k_wk_h}, \end{aligned}$$

which gives the desired upper bound.

As for the lower bound, observe that

$$\begin{aligned} &\frac{2}{s_w} \cdot \mathbb{E}\left(\frac{\|w\|_F^2}{\|w^{c_{out} \times (c_{in}k_wk_h)}\|_{sp}^2 + \|w^{c_{in} \times (c_{out}k_wk_h)}\|_{sp}^2}\right) \\ &= \frac{2}{s_w} \cdot \mathbb{E}\left(\frac{1}{\left\|\frac{w^{c_{out} \times (c_{in}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2 + \left\|\frac{w^{c_{in} \times (c_{out}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2}\right) \\ &\geq \frac{2}{s_w} \cdot \frac{1}{\mathbb{E}\left(\left\|\frac{w^{c_{out} \times (c_{in}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2\right) + \mathbb{E}\left(\left\|\frac{w^{c_{in} \times (c_{out}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2\right)} \end{aligned}$$

Then we can follow the same approach in App. F for bounding $\mathbb{E}\left(\left\|\frac{w^{c_{out} \times (c_{in}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2\right)$ and $\mathbb{E}\left(\left\|\frac{w^{c_{in} \times (c_{out}k_wk_h)}}{\|w\|_F}\right\|_{sp}^2\right)$, which gives the desired lower bound. □

H. Datasets and Metrics

H.1. Datasets

MNIST (LeCun & Cortes, 2010) We use the training set for our experiments, which contains 60000 images of handwritten digits of shape $28 \times 28 \times 1$. The pixels values are normalized to $[0, 1]$ before feeding to the discriminators.

CIFAR10 (Krizhevsky et al., 2009) We use the training set for our experiments, which contains 50000 images of shape $32 \times 32 \times 3$. The pixels values are normalized to $[-1, 1]$ before feeding to the discriminators.

STL10 (Coates et al., 2011) We use the unlabeled set for our experiments, which contains 100000 images of shape $96 \times 96 \times 3$. Following (Miyato et al., 2018), we resize the images to $48 \times 48 \times 3$ for training. The pixels values are normalized to $[-1, 1]$ before feeding to the discriminators.

CelebA (Liu et al., 2015) This dataset contains 202599 images. For each image, we crop the center 128×128 , and resize it to $64 \times 64 \times 3$ for training. The pixels values are normalized to $[-1, 1]$ before feeding to the discriminators.

ImageNet (ILSVRC2012) (Russakovsky et al., 2015) The dataset contains 1281167 images. Following (Miyato et al., 2018), for each images, we crop the central square of the images according to $\min(\text{width}, \text{height})$, and then reshape it to $128 \times 128 \times 3$ for training. The pixels values are normalized to $[-1, 1]$ before feeding to the discriminators.

H.2. Metrics

Inception score (Salimans et al., 2016) Following (Miyato et al., 2018), we use 50000 generated images and split them into 10 sets for computing the score.

FID (Heusel et al., 2017) Following (Miyato et al., 2018), we use 5000 real images and 10000 generated images for computing the score.

I. Gradient Explosion and Vanishing in GANs

I.1. Results

To illustrate that gradient explosion and vanishing are closely related to the instability in GANs, we trained a WGAN (Gulrajani et al., 2017) on the CIFAR10 dataset with different hyper-parameters leading to stable training, exploding gradients, and vanishing gradients over 40,000 training iterations (more experimental details in App. I.2). Fig. 10 shows the resulting inception scores for each of these runs, and Fig. 11 shows the corresponding magnitudes of the gradients over the course of training. Note that the stable run has improved sample quality and stable gradients throughout training. This phenomenon has also been observed in prior literature (Arjovsky & Bottou, 2017; Brock et al., 2018). We will demonstrate that by controlling these gradients, SN (and SN_w in particular) is able to achieve more stable training and better sample quality.

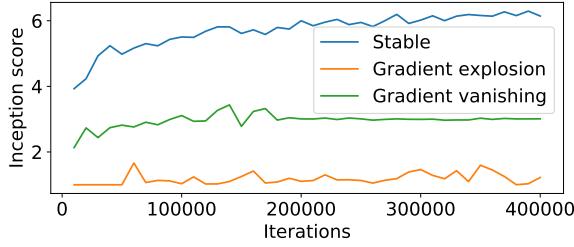


Figure 10. Inception score over the course of training. The “gradient vanishing” inception score plateaus as training is stalled.

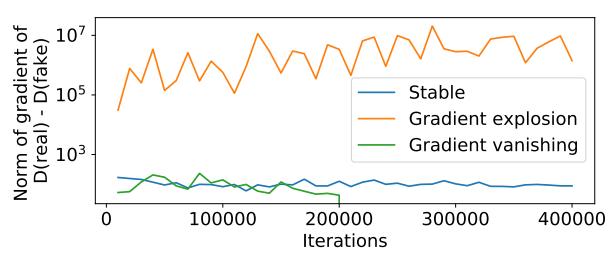


Figure 11. Norm of gradient with respect to parameters during training. The vanishing gradient collapses after 200k iterations.

I.2. Experimental Details

The network architectures are shown in Tables 2 and 3. The dataset is CIFAR10. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. Let λ be the WGAN’s gradient penalty weight (Gulrajani et al., 2017). For the stable run, $\alpha_g = 0.0001, \alpha_d = 0.0002, \beta_1 = 0.5, \beta_2 = 0.999, \lambda = 10, n_{dis} = 1$. For the gradient explosion run, $\alpha_g = 0.001, \alpha_d = 0.001, \beta_1 = 0.5, \beta_2 = 0.999, \lambda = 10, n_{dis} = 1$. For the gradient vanishing run,

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
Fully connected ($M_g \times M_g \times 512$). BN. ReLU.
Deconvolution ($c = 256, k = 4, s = 2$). BN. ReLU.
Deconvolution ($c = 128, k = 4, s = 2$). BN. ReLU.
Deconvolution ($c = 64, k = 4, s = 2$). BN. ReLU.
Deconvolution ($c = 3, k = 3, s = 1$). Tanh.

Table 2. Generator network architectures for CIFAR10, STL10, and CelebA experiments (from (Miyato et al., 2018)). For CIFAR10, $M_g = 4$. For STL10, $M_g = 6$. For CelebA, $M_g = 8$. BN stands for batch normalization. c stands for number of channels. k stands for kernel size. s stands for stride.

$x \in \mathbb{R}^{M \times M \times 3}$
Convolution ($c = 64, k = 3, s = 1$). Leaky ReLU (0.1).
Convolution ($c = 64, k = 4, s = 2$). Leaky ReLU (0.1).
Convolution ($c = 128, k = 3, s = 1$). Leaky ReLU (0.1).
Convolution ($c = 128, k = 4, s = 2$). Leaky ReLU (0.1).
Convolution ($c = 256, k = 3, s = 1$). Leaky ReLU (0.1).
Convolution ($c = 256, k = 4, s = 2$). Leaky ReLU (0.1).
Convolution ($c = 512, k = 3, s = 1$). Leaky ReLU (0.1).
Fully connected (1).

Table 3. Discriminator network architectures for CIFAR10, STL10, and CelebA experiments (from (Miyato et al., 2018)). For CIFAR10, $M = 32$. For STL10, $M = 48$. For CelebA, $M = 64$. c stands for number of channels. k stands for kernel size. s stands for stride.

$\alpha_g = 0.001, \alpha_d = 0.001, \beta_1 = 0.5, \beta_2 = 0.999, \lambda = 50, n_{dis} = 1$, and the activation functions in the discriminator are changed from leaky ReLU to ReLU.

J. Experimental Details and Additional Results on Gradient Norms

J.1. Experimental Details

For the MNIST experiment, the network architectures are shown in Tables 4 and 5. All experiments are run for 100 epochs. Batch size is 64. The optimizer is Adam. $\alpha_g = 0.001, \alpha_d = 0.001, \beta_1 = 0.5, \beta_2 = 0.999, n_{dis} = 1$.

For the CIFAR10 experiment, , the network architectures are shown in Tables 2 and 3. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. $\alpha_g = 0.0001, \alpha_d = 0.0001, \beta_1 = 0.5, \beta_2 = 0.999, n_{dis} = 1$.

Let λ be the WGAN’s gradient penalty weight (Gulrajani et al., 2017). For the runs without SN, $\lambda = 10$. For the runs with SN, we use the strict SN implementation (Farnia et al., 2018) in order to verifying the theoretical results (the popular SN implementation (Miyato et al., 2018) only gives a loose bound on the actual spectral norm of layers, see § 4). Since it already ensures that the Lipschitz constant of the discriminator is no more than 1, we discard the gradient penalty loss from training.

For all the results, the gradient norm only considers the weights and excludes the biases (if exist), so as to be consistent with the theoretical analysis.

$z \in \mathbb{R}^{100} \sim \text{Uniform}(-1, 1)$
Fully connected ($7 \times 7 \times 128$). Leaky ReLU (0.2). BN.
Deconvolution ($c = 64, k = 5, s = 2$). Leaky ReLU (0.2). BN.
Deconvolution ($c = 1, k = 5, s = 2$). Sigmoid.

Table 4. Generator network architectures for MNIST experiments. BN stands for batch normalization. c stands for number of channels. k stands for kernel size. s stands for stride.

$x \in \mathbb{R}^{28 \times 28 \times 1}$
Convolution ($c = 64, k = 5, s = 2$, no bias). Leaky ReLU (0.2).
Convolution ($c = 128, k = 5, s = 2$, no bias). Leaky ReLU (0.2).
Convolution ($c = 256, k = 5, s = 2$, no bias). Leaky ReLU (0.2).
Fully connected (1, no bias).

Table 5. Discriminator network architectures for MNIST experiments. c stands for number of channels. k stands for kernel size. s stands for stride.

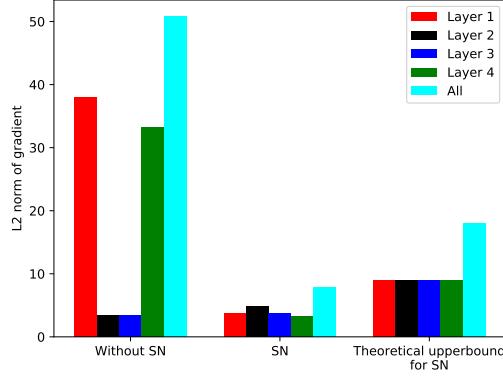


Figure 12. Gradient norms of each discriminator layer in MNIST at epoch 50.

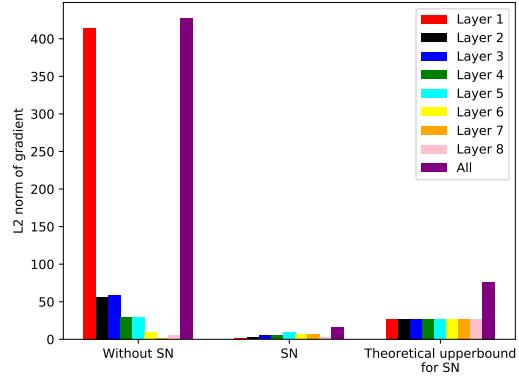


Figure 13. Gradient norms of each discriminator layer in CIFAR10 at iteration 10000.

J.2. Additional Results

Figs. 12 and 13 show the gradient norms of each discriminator layer in MNIST and CIFAR10. Despite the difference on the network architecture and dataset, we see the similar phenomenon: when training without SN, some layers have extremely large gradient norms, which causes the overall gradient norm to be large; when training with SN, the gradient norms are much smaller and are similar across different layers.

K. Experimental Details and Additional Results for Confirming Eq. (2)

K.1. Experimental Details

For the MNIST experiment, the network architectures are shown in Tables 4 and 5. All experiments are run for 100 epochs. Batch size is 64. The optimizer is Adam. $\alpha_g = 0.001, \alpha_d = 0.001, \beta_1 = 0.5, \beta_2 = 0.999, n_{dis} = 1$. We use WGAN loss with the strict SN implementation (Farnia et al., 2018). Since it already ensures that the Lipschitz constant of the discriminator is no more than 1, we discard the gradient penalty loss from training. The random scaling are selected in a way the geometric mean of spectral norms of all layers equals 1.

For the CIFAR10 and STL10 experiments , the network architectures are shown in Tables 2 and 3. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. $\alpha_g = 0.0001, \alpha_d = 0.0001, \beta_1 = 0.5, \beta_2 = 0.999, n_{dis} = 1$. We use hinge loss (Miyato et al., 2018) with the strict SN implementation (Farnia et al., 2018). The random scaling are selected in a way the geometric mean of spectral norms of all layers equals 1.75, which avoids the gradient vanishing problem as seen in § 4.

K.2. Additional Results

Figs. 14 and 15 show the ratios of the gradient norms at each layer and the inverse ratios of the spectral norms in CIFAR10 and STL10. Generally, we see that the most of the points are near the diagonal line, which means that the assumption in Eq. (2) is reasonably true in practice. However, we note that the last layer (layer 8) somehow has slightly smaller gradient,

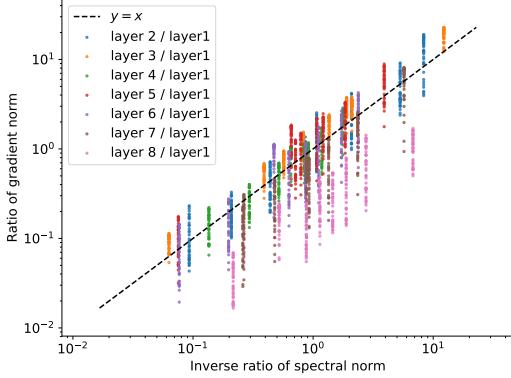


Figure 14. Ratio of gradient norm v.s. inverse ratio of spectral norm in CIFAR10.

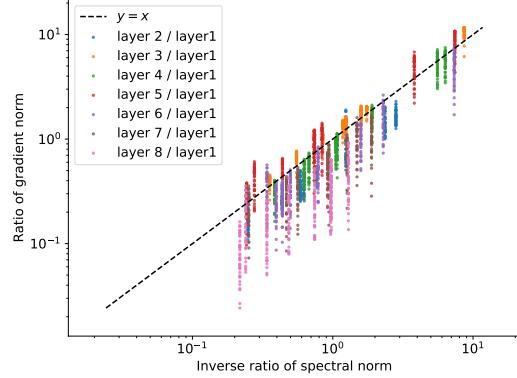


Figure 15. Ratio of gradient norm v.s. inverse ratio of spectral norm in STL10.

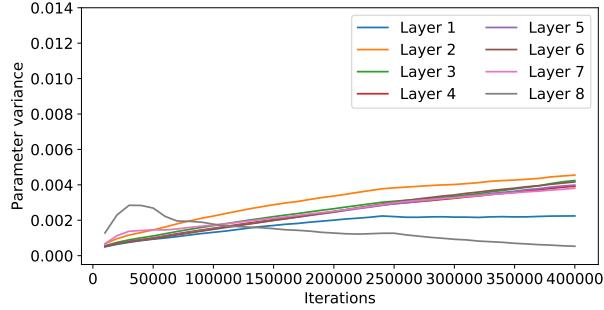


Figure 16. Parameter variance without SN in CIFAR10.

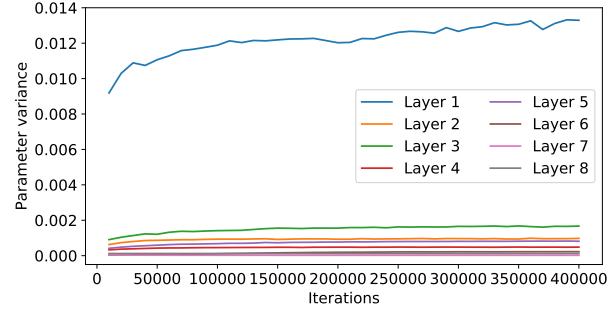


Figure 17. Parameter variance with SN in CIFAR10.

as the points of ‘‘layer 8 / layer 1’’ are slightly lower than the diagonal line. This could result from the fact that layer 8 is a fully connected layer whereas all other layers are convolutional layers. We defer the more detailed analysis of this phenomenon to future work.

L. Experimental Details and Additional Results on Vanishing Gradient

L.1. Experimental Details

The network architectures are shown in Tables 2 and 3. The dataset is CIFAR10. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $n_{dis} = 1$. We use hinge loss (Miyato et al., 2018).

L.2. Parameter Variance With and Without SN

Figs. 16 and 17 show the parameter variance of each layer without and with SN. Note that Fig. 17 is just collecting the empirical lines in Fig. 7 for the ease of comparison here. Figs. 18 and 19 show the gradient norm and inception score.

We can see that when training with SN, the parameter variance is stable throughout training (Fig. 17), and the magnitude of gradient is also stable (Fig. 18). However, when training without SN, the parameter variance tends to increase throughout training (Fig. 16), which causes a quick decrease in the magnitude of gradient in the beginning of training (Fig. 18) because of the saturation of hinge loss (§ 4). Because SN promotes the stability of the variance and gradient throughout training, we see that SN improves the sample quality significantly (Fig. 19).

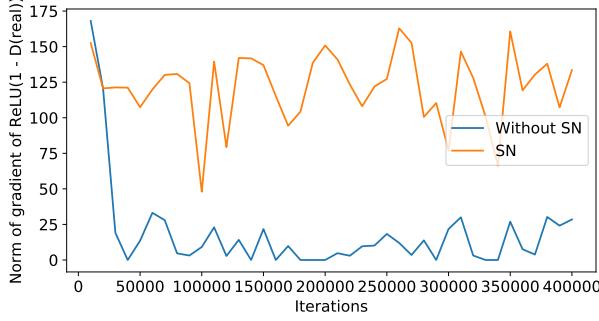


Figure 18. Gradient norm with and without SN in CIFAR10.

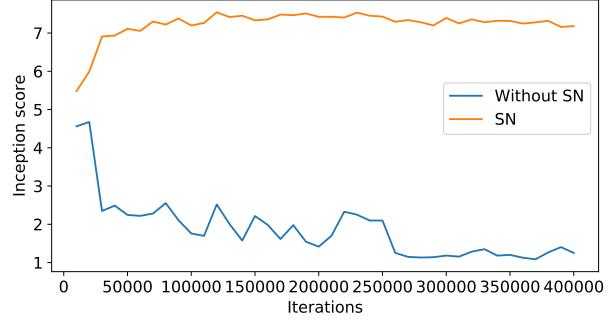


Figure 19. Inception score with and without SN in CIFAR10.

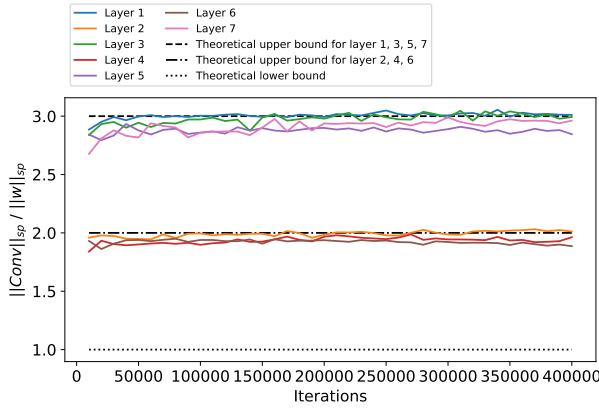


Figure 20. The ratio of two spectral norms throughout the training of the popular SN (Miyato et al., 2018) in CIFAR10.

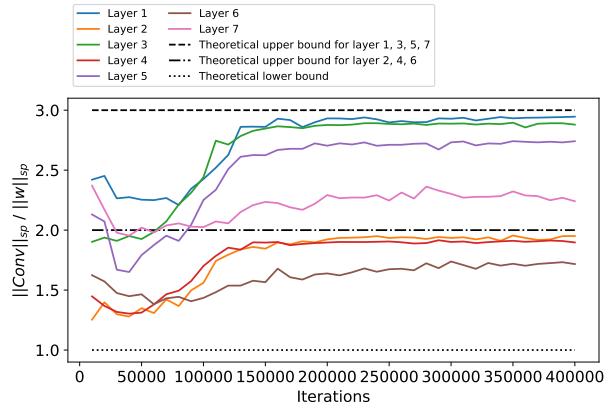


Figure 21. The ratio of two spectral norms throughout the training of the strict SN (Farnia et al., 2018) in CIFAR10.

L.3. Comparing Two Variants Spectral Norms

Figs. 20 and 21 show the ratio between two versions of spectral norm (Miyato et al., 2018; Farnia et al., 2018) throughout the training of the popular SN (Miyato et al., 2018) and the strict SN (Farnia et al., 2018). $\|Conv\|_{sp}$ denotes the spectral norm of the expanded matrix $\|\tilde{w}\|_{sp}$ used in (Farnia et al., 2018). $\|w\|_{sp}$ denotes the spectral norm of reshaped matrix $\|\hat{w}\|_{sp}$ used in (Miyato et al., 2018). The theoretical lower and upper bound are calculated according to Corollary 1 in (Tsuzuku et al., 2018). We can see that no matter in which architecture, $\|\tilde{w}\|_{sp}$ is usually strictly larger than $\|\hat{w}\|_{sp}$. Note that the reason why in some cases the ratio exceeds the upper bound in Fig. 20 is because the spectral norms are calculated using power iteration (Miyato et al., 2018; Farnia et al., 2018) which has approximation error.

L.4. Parameter Variance of Scaled SN

Figure Fig. 22 shows the parameter variance of scaled SN for both SN versions (Miyato et al., 2018; Farnia et al., 2018). We can see that when scale=1.75, the product of parameter variances for SN_{Conv} (Farnia et al., 2018) is similar to the one of SN_w (Miyato et al., 2018). Moreover, by comparing Fig. 22 and Fig. 6 we can see that when the products of variances of two SN variants are similar, the sample quality is also similar. This confirms the intuition from LeCun initialization (LeCun et al., 1998) that the magnitude of variance plays an important role on the performance of neural network, and it should not be too large nor too small.

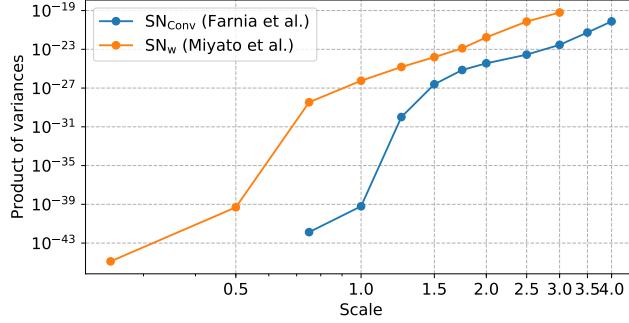


Figure 22. The parameter variance of scaled SN in CIFAR10.

M. Experimental Details and Additional Results on Scaling (§ 5.2)

M.1. Experimental Details

The network architectures are shown in Tables 2 and 3. SN models are run for 400k iterations. LeCun initialization models are run till the sample quality converges or starts dropping (usually within 400k iterations). Batch size is 64. The optimizer is Adam. $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $n_{dis} = 1$. We use hinge loss (Miyato et al., 2018).

Since LeCun initialization is unstable when the scaling is not proper, in Fig. 8, we plot the best score during training instead of the score at the end of training.

M.2. Additional Results

Although the good scaling modes for SN and LeCun initialization seem to be very different in Fig. 8, there indeed exists a (perhaps coincidental) correspondence in terms of parameter variances. In Fig. 23, we show the inception score v.s. parameter variances for SN and LeCun initialization. We can see that the first good mode occurs when log of the product of parameter variances is around -70 to -60, and the second mode is around -50 to -40.

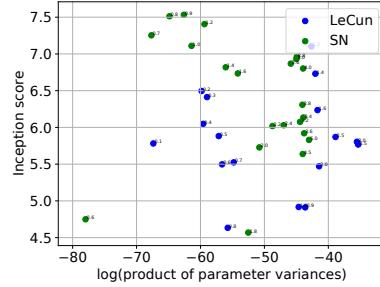


Figure 23. Inception score v.s. parameter variances of scaled SN and scaled LeCun initialization in CIFAR10. Each point corresponds to one run, at the point when the score is the best during training. The numbers near each point indicate the scaling.

N. Results with Different Hyper-parameters and SN Variants

In addition to SN (Miyato et al., 2018), we compare against two variants of SN proposed in the appendix of (Miyato et al., 2018), which we denote “same γ ” and “diff. γ ” (details in App. O). These two variants are reported to be worse than SN in (Miyato et al., 2018) and are not used in practice, but we include them here for reference. We run experiments on CIFAR10, STL10, CelebA, and ImageNet, with two widely-used metrics for sample quality: inception score (Salimans et al., 2016) and Frechet Inception Distance (FID) (Heusel et al., 2017) (details in App. H).

We use the network architecture from SN (Miyato et al., 2018). We controlled five hyper-parameters (Table 7, App. P): α_g and α_d , the generator/discriminator learning rates, β_1, β_2 , Adam momentum parameters (Kingma & Ba, 2014), and n_{dis} ,

	CIFAR10		STL10		CelebA
	Inception score \uparrow	FID \downarrow	Inception score \uparrow	FID \downarrow	FID \downarrow
Real data	11.26	9.70	26.70	10.17	4.44
SN (same γ)	6.46 ± 0.06	42.35 ± 0.74	8.86 ± 0.03	54.61 ± 0.51	7.74 ± 0.11
BSN (same γ)	6.69 ± 0.05	39.62 ± 0.40	8.76 ± 0.03	55.04 ± 0.48	7.83 ± 0.09
SN (diff. γ)	6.53 ± 0.01	41.88 ± 0.50	8.79 ± 0.03	56.76 ± 0.44	7.54 ± 0.08
BSN (diff. γ)	6.72 ± 0.05	38.15 ± 0.72	8.80 ± 0.03	53.99 ± 0.33	7.67 ± 0.04
SN	7.22 ± 0.09	31.43 ± 0.90	9.16 ± 0.03	42.89 ± 0.54	9.09 ± 0.32
BSN	7.58 ± 0.04	26.62 ± 0.21	9.25 ± 0.01	42.98 ± 0.54	8.54 ± 0.20

Table 6. Inception scores and FIDs on CIFAR10, STL10, and CelebA. Each experiment is conducted with 5 random seeds, with mean and standard error reported. We follow the common practice of excluding Inception Score in CelebA as the inception network is pretrained on ImageNet, which is very different from CelebA. The bold font marks the best numbers between SN and BSN using the same variant. The red color marks the best numbers among all runs. The “same γ ” and “diff. γ ” variants are not used in practice and are reported to have bad performance in (Miyato et al., 2018).

the number of discriminator updates per generator update. Three hyper-parameter settings are from (Miyato et al., 2018), with equal discriminator and generator learning rates; the final two test *unequal* learning rates for showing a more thorough comparison. More details are in Apps. P and Q.

As in (Miyato et al., 2018), we report the metrics from the best hyper-parameter for each algorithm in Table 6. *BSN outperforms the standard SN in all sample quality metrics except FID score on STL10, where their metrics are within standard error of each other.* Regarding the SN variants with γ , in CIFAR10 and STL10, they have worse performance than SN and BSN, same as reported in (Miyato et al., 2018). In CelebA, the SN variants have better performance for the best hyper-parameter setting. But in general, these SN variants are very sensitive to hyper-parameters (Apps. P to R), therefore they are not adopted in practice (Miyato et al., 2018). Nevertheless, BSN is still able to improve or have similar performance on those two variants in most of the settings.

More importantly, the superiority of BSN is stable across hyper-parameters. Figs. 24 and 25 show the inception scores of all the hyper-parameters we tested on CIFAR10 and STL10. BSN has the best or competitive performance in most of the settings. The only exception is $n_{dis} = 5$ setting in STL10, where we observe that the performance from both SN and BSN have larger variance across different random seeds, and the SN variants with γ perform better. On CelebA, BSN also outperforms SN in FID across all hyper-parameters (App. R), and it outperforms all SN variants in every hyper-parameter setting except one (Fig. 55).

More results (generated images, training curves, FID plots) are in Apps. P to R.

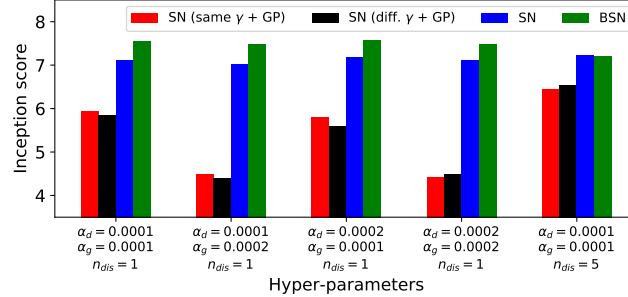


Figure 24. Inception score in CIFAR10. The results are averaged over 5 random seeds.

O. Details on SN Variants

In Appendix E of (Miyato et al., 2018), a variant of SN is introduced. Instead of strictly setting the spectral norm of each layer, the idea of this approach is to release the constraint by multiplying each spectral normalized weights with a trainable

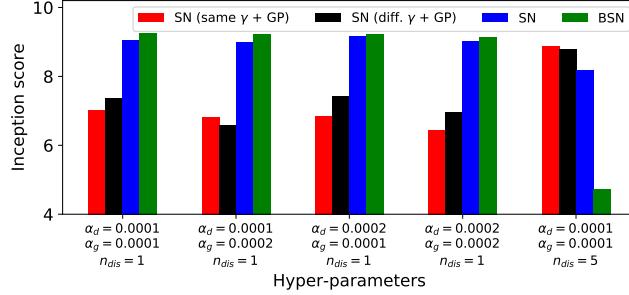


Figure 25. Inception score in STL10. The results are averaged over 5 random seeds.

α_g	α_d	β_1	β_2	n_{dis}
0.0001	0.0001	0.5	0.9	5
0.0001	0.0001	0.5	0.999	1
0.0002	0.0002	0.5	0.999	1
0.0001	0.0002	0.5	0.999	1
0.0002	0.0001	0.5	0.999	1

Table 7. Hyper-parameters tested in CIFAR10 and STL10 experiments. The first three settings are from (Miyato et al., 2018; Gulrajani et al., 2017; Warde-Farley & Bengio, 2016; Radford et al., 2015). α_g and α_d : learning rates for generator and discriminator. β_1, β_2 : momentum parameters in Adam. n_{dis} : number of discriminator updates per generator update.

parameter γ . However, this would make the gradient of discriminator arbitrarily large, which violates the original motivation of SN. Therefore, the approach incorporates gradient penalty (Gulrajani et al., 2017) for setting the Lipschitz constant of discriminator to 1. The gradient penalty weights are set to 10 in all experiments.

However, from the description in (Miyato et al., 2018), it is unclear if all layers have the same or separated γ . Therefore, we try both versions in our experiments. “Same γ ” denotes that version where all layers share the same γ . “Diff. γ ” denotes the version where each layer has a separate γ .

P. Experimental Details and Additional Results on CIFAR10

P.1. Experimental Details

The network architectures are shown in Tables 2 and 3. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. We use the five hyper-parameter settings listed in Table 7. (In Table 1 we only show the results from the first hyper-parameter setting.) We use hinge loss with the popular SN implementation (Miyato et al., 2018).

For SSN in Table 1, we ran following scales: [0.7, 0.8, 0.9, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 4.5, 5.0, 5.5, 6.0, 7.0, 8.0, 9.0, 10.0], and present the results from best one for each metric. For BSSN in Table 1, we ran the following scales: [0.7, 0.8, 0.9, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0], and present the results from the best one for each metric.

P.2. FID Plot

Fig. 26 shows the FID score in CIFAR10 dataset. We can see that BSN has the best performance in all 5 hyper-parameter settings.

P.3. Training Curves

From App. N we can see that SN (no γ) and BSN generally have the best performance. Therefore, in this section, we focus on comparing these two algorithms with the training curves. Figs. 9 and 27 to 35 show the inception score and FID of these two algorithms during training. Generally, we see that BSN converges slower than SN *at the beginning of training*. However,

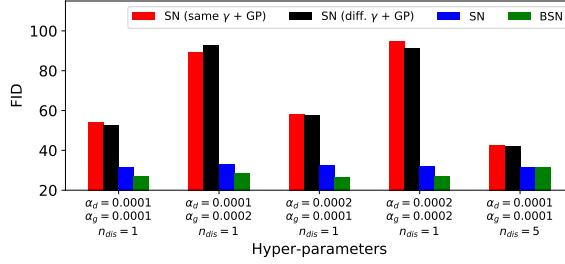


Figure 26. FID in CIFAR10. The results are averaged over 5 random seeds.

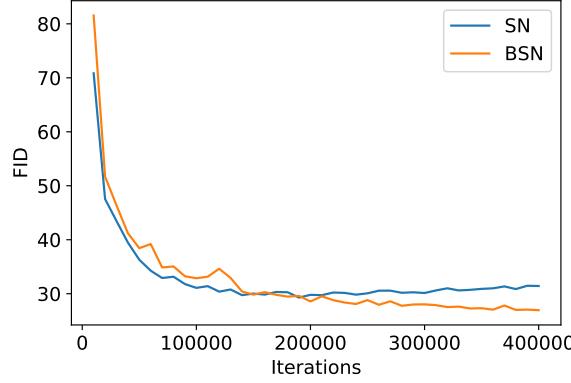


Figure 27. FID in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

as training proceeds, the sample quality of SN often drops (e.g. Figs. 9 and 27 to 33), whereas the sample quality of BSN always increases and then stabilizes at the high level. In most cases, BSN not only outperforms SN at the end of training, but also outperforms the peak sample quality of SN during training (e.g. Figs. 9 and 27 to 33). From these results, we can conclude that BSN improves both the sample quality and training stability over SN.

P.4. Generated Images

Figs. 36 to 39 show the generated images from the run with the best inception score for each algorithm.

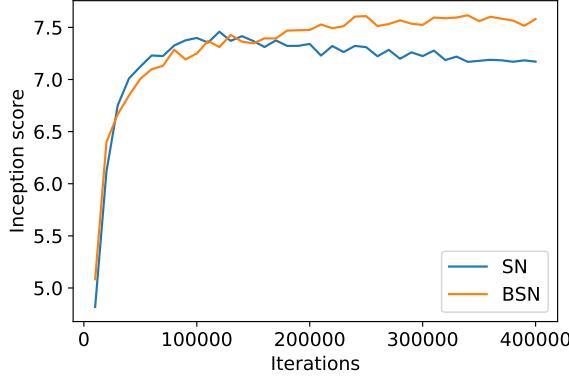


Figure 28. Inception score in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

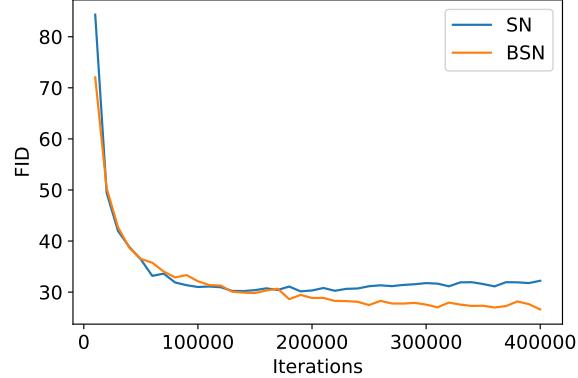


Figure 29. FID in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

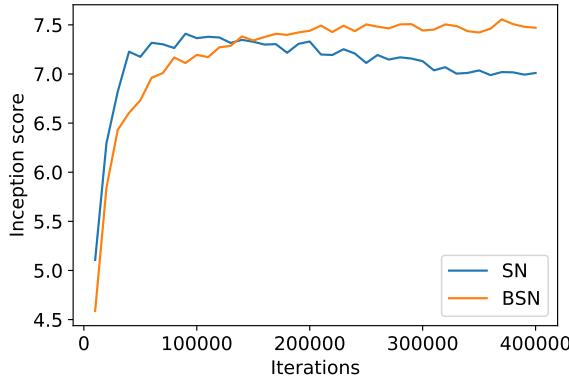


Figure 30. Inception score in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

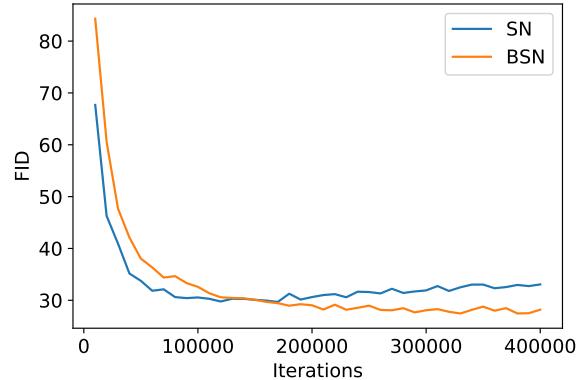


Figure 31. FID in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

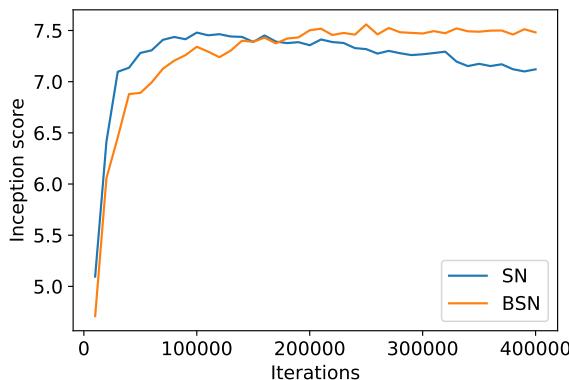


Figure 32. Inception score in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

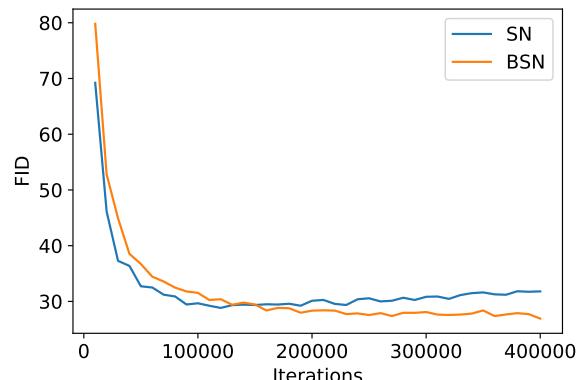


Figure 33. FID in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

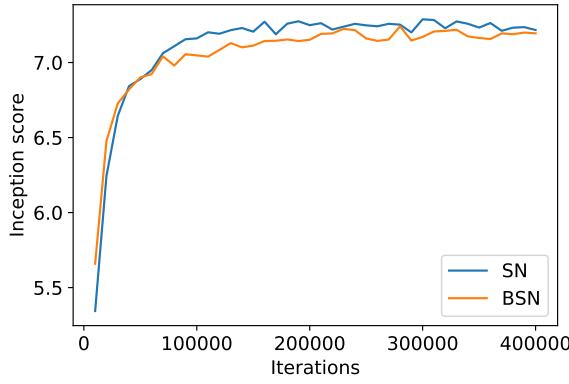


Figure 34. Inception score in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$.

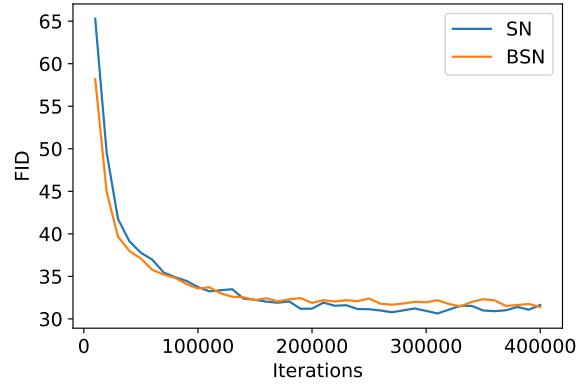


Figure 35. FID in CIFAR10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$.

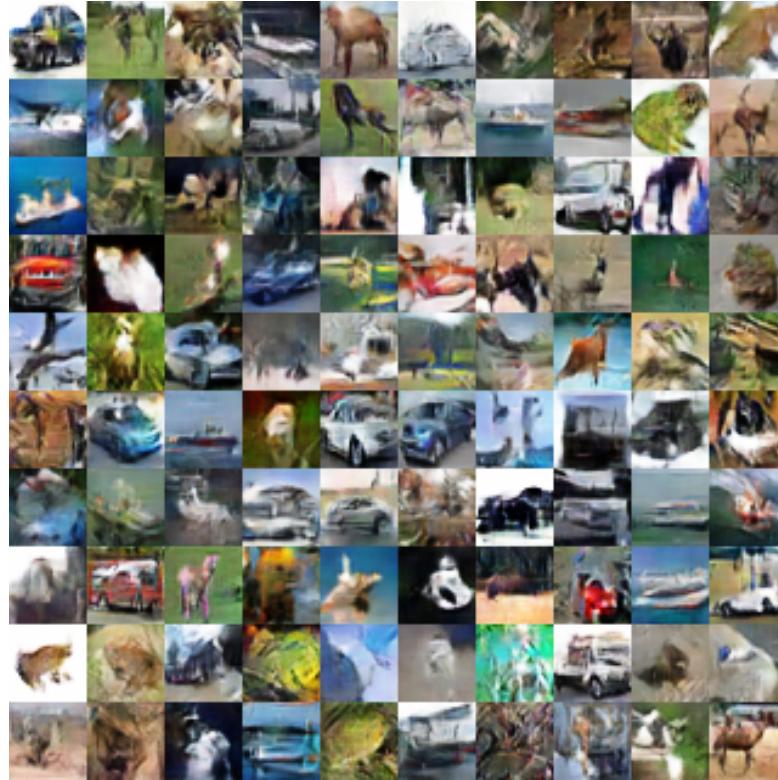


Figure 36. Generated samples from the best run of SN (same γ) in CIFAR10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. Inception score is 6.64. FID is 41.01.

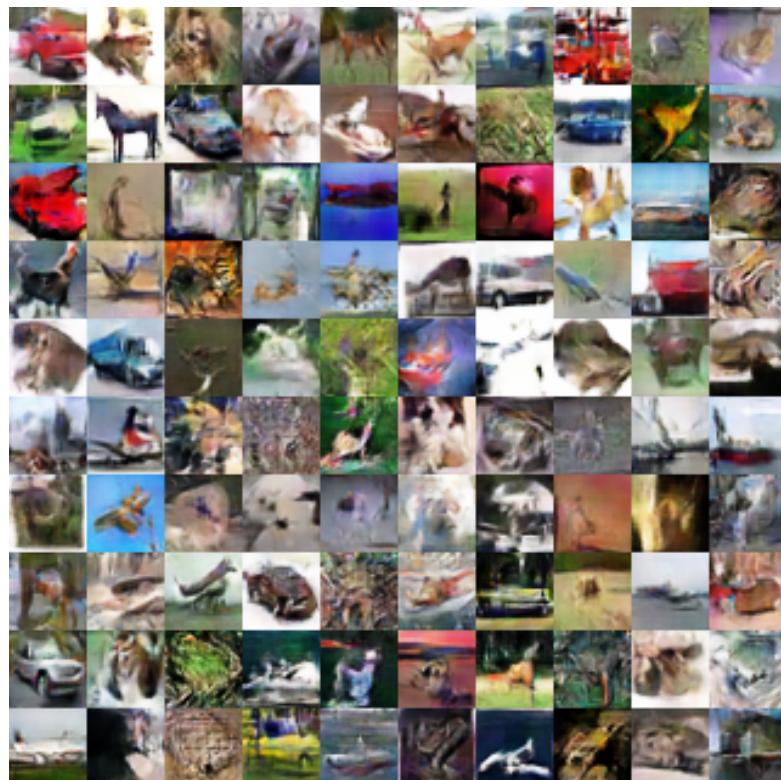


Figure 37. Generated samples from the best run of SN (diff. γ) in CIFAR10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. Inception score is 6.55. FID is 41.18.



Figure 38. Generated samples from the best run of SN in CIFAR10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$. Inception score is 7.56. FID is 28.64.

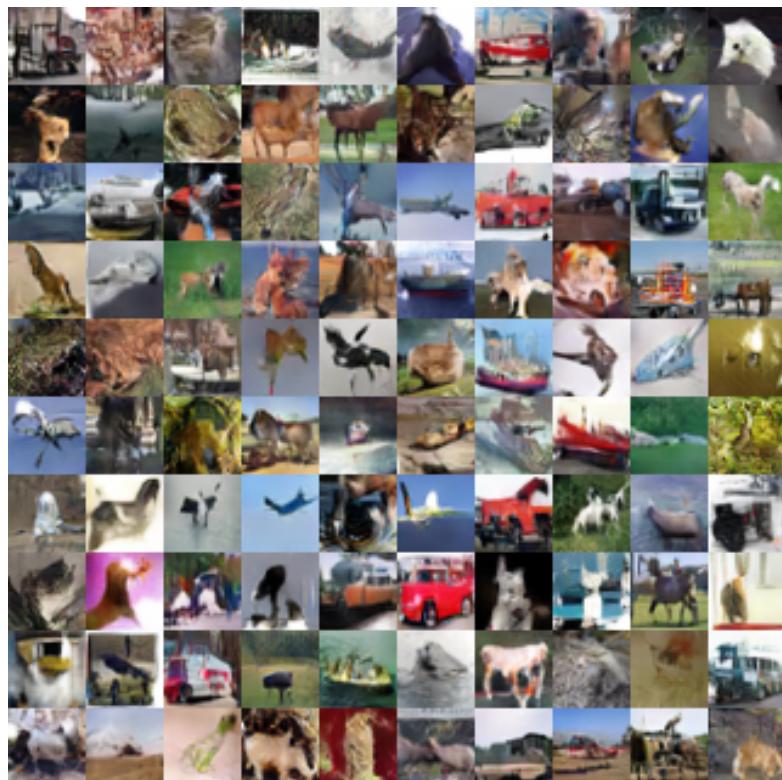


Figure 39. Generated samples from the best run of BSN in CIFAR10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$. Inception score is 7.70. FID is 25.96.

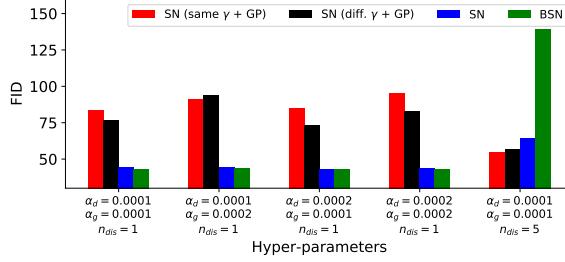


Figure 40. FID in STL10. The results are averaged over 5 random seeds.

Q. Experimental Details and Additional Results on STL10

Q.1. Experimental Details

The network architectures are shown in Tables 2 and 3. Batch size is 64. The optimizer is Adam. We use the five hyper-parameter settings listed in Table 7. (In Table 1 we only show the results from the first hyper-parameter setting.) We use hinge loss with the popular SN implementation (Miyato et al., 2018).

SN (no γ) and BSN under $n_{dis} = 1$ settings are run for 800k iterations as we observe that they need longer time to converge. All other experiments are run for 400k iterations.

For SSN and BSSN in Table 1, we ran following scales: [0.7, 0.8, 0.9, 1.0, 1.2, 1.4, 1.6], and present the results from best one for each metric.

Q.2. FID Plot

Fig. 40 shows the FID score in STL10 dataset. We can see that BSN has the best or competitive performance in most of the hyper-parameter settings. Again, the only exception is $n_{dis} = 5$ setting.

Q.3. Training Curves

From App. N we can see that SN (no γ) and BSN generally have the best performance. Therefore, in this section, we focus on comparing these two algorithms with the training curves. Figs. 41 to 50 show the inception score and FID of these two algorithms during training. Generally, we see that BSN converges slower than SN at the beginning of training. However, as training proceeds, BSN finally has better metrics in most cases. Note that unlike CIFAR10, SN seems to be more stable in STL10 as its sample quality does not drop in most hyper-parameters. But the key conclusion is the same: in most cases, BSN not only outperforms SN at the end of training, but also outperforms the peak sample quality of SN during training (e.g. Figs. 41 to 48). The only exception is the $n_{dis} = 5$ setting, where both SN and BSN has instability issue: the sample quality first improves and then significantly drops. The problem with BSN seems to be severer. We discussed about this problem in App. N.

Q.4. Generated Images

Figs. 51 to 54 show the generated images from the run with the best inception score for each algorithm.

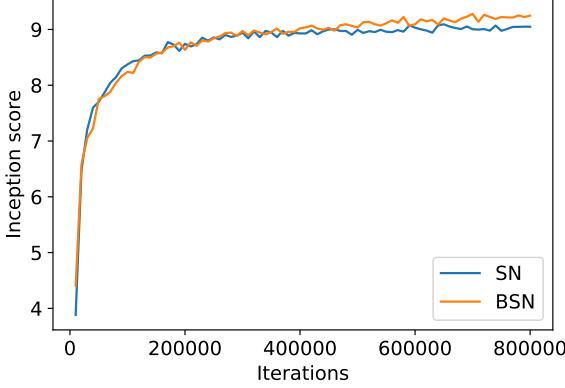


Figure 41. Inception score in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

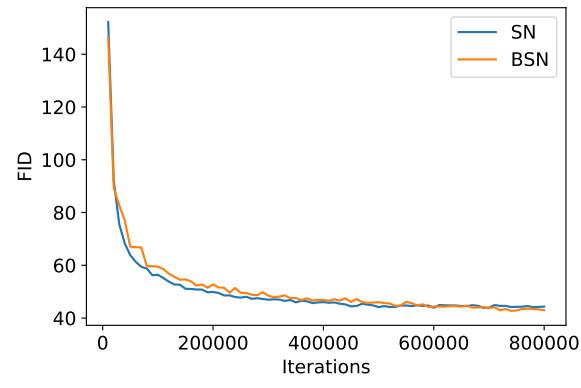


Figure 42. FID in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

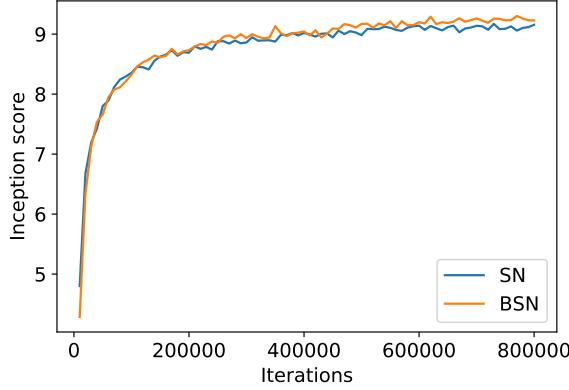


Figure 43. Inception score in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

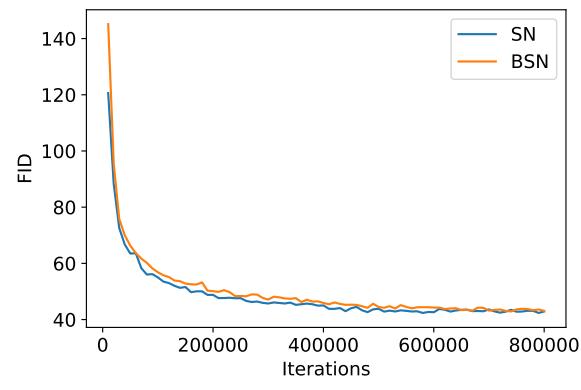


Figure 44. FID in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

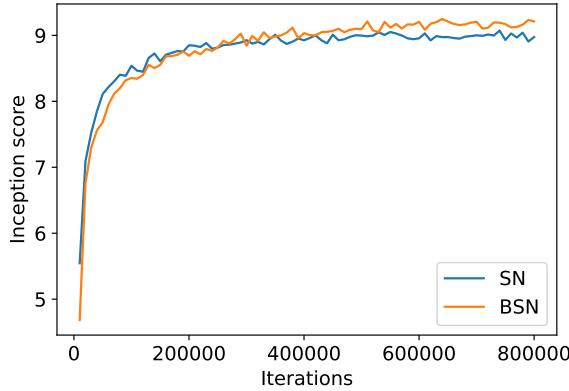


Figure 45. Inception score in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

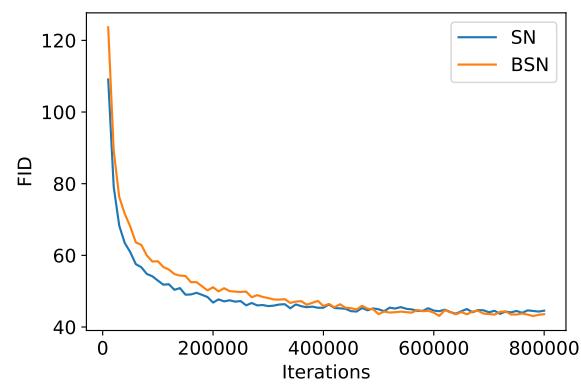


Figure 46. FID in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

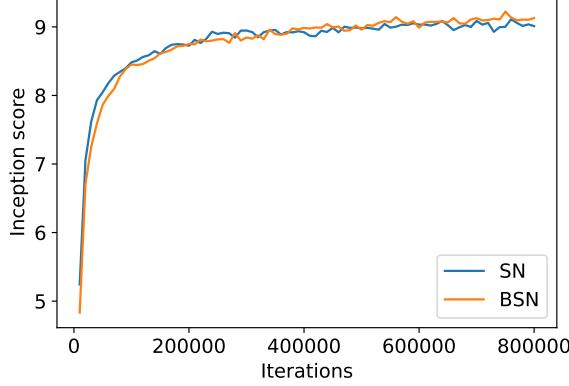


Figure 47. Inception score in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

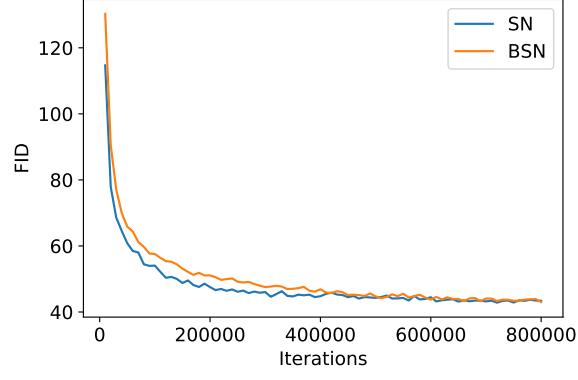


Figure 48. FID in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

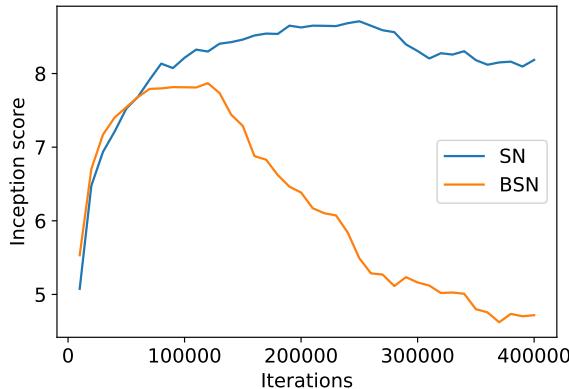


Figure 49. Inception score in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$.

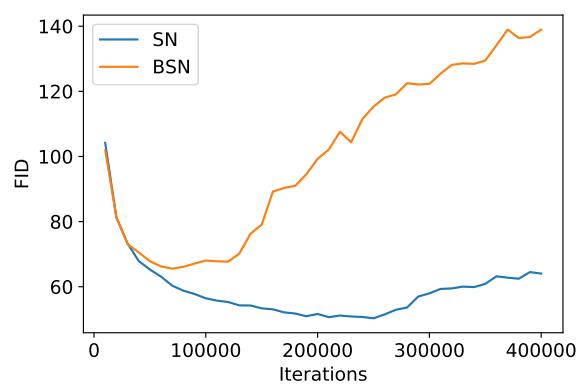


Figure 50. FID in STL10. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$.

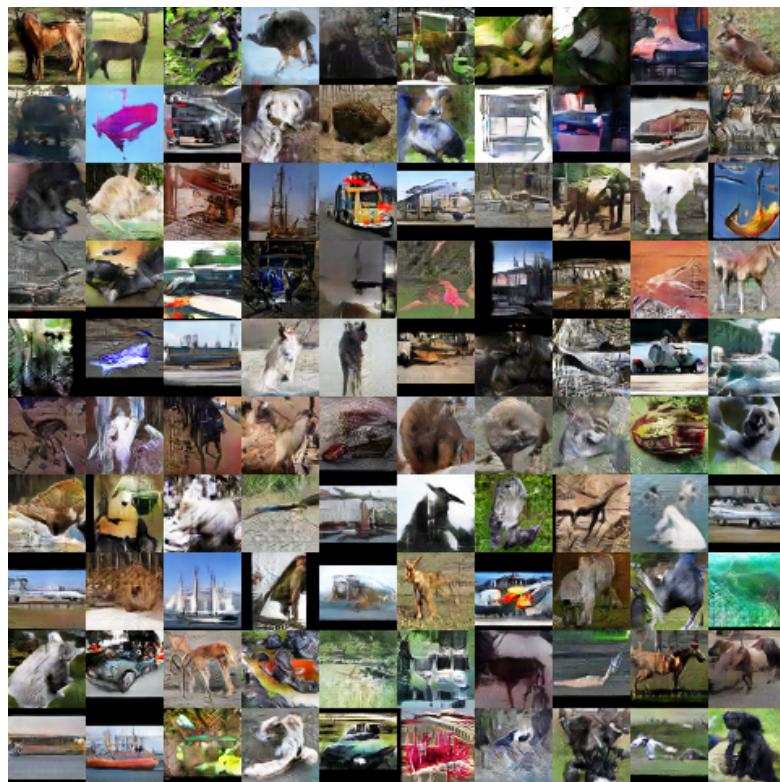


Figure 51. Generated samples from the best run of SN (same γ) in STL10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. Inception score is 8.96. FID is 53.94.

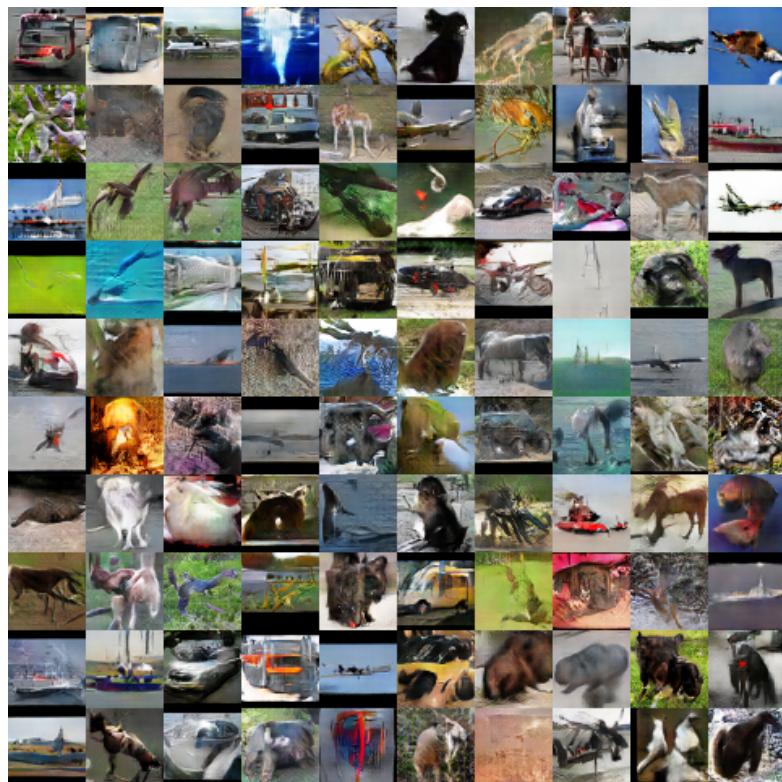


Figure 52. Generated samples from the best run of SN (diff. γ) in STL10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. Inception score is 8.88. FID is 56.14.

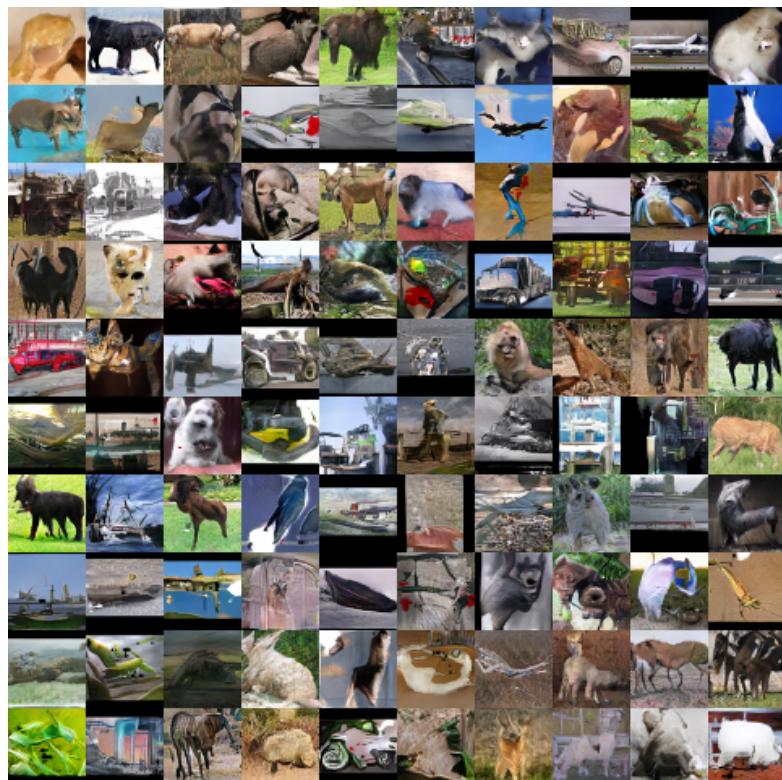


Figure 53. Generated samples from the best run of SN in STL10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$. Inception score is 9.26. FID is 44.38.

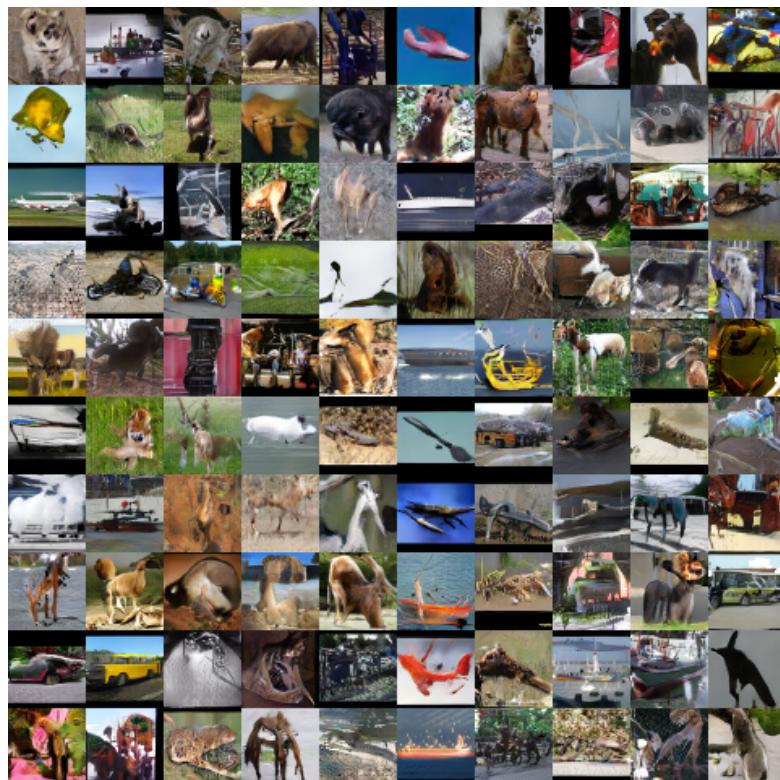


Figure 54. Generated samples from the best run of BSN in STL10. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$. Inception score is 9.46. FID is 42.78.

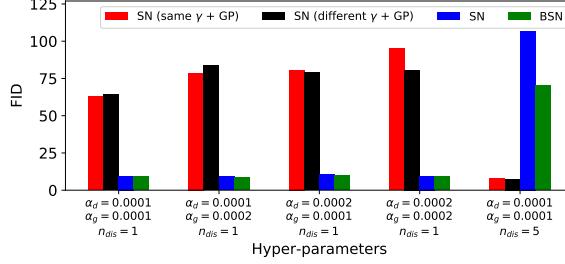


Figure 55. FID in CelebA. The results are averaged over 5 random seeds.

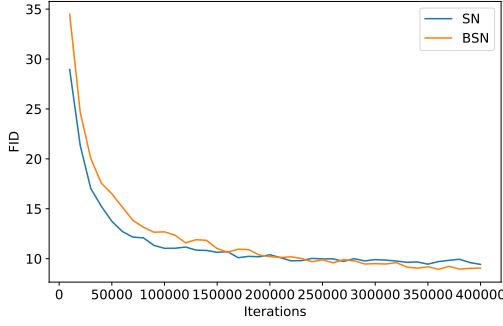


Figure 56. FID in CelebA. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

R. Experimental Details and Additional Results on CelebA

R.1. Experimental Details

The network architectures are shown in Tables 2 and 3. All experiments are run for 400k iterations. Batch size is 64. The optimizer is Adam. We use the five hyper-parameter settings listed in Table 7. (In Table 1 we only show the results from the first hyper-parameter setting.) We use hinge loss with the popular SN implementation (Miyato et al., 2018).

For SSN and BSSN in Table 1, we ran following scales: [0.7, 0.8, 0.9, 1.0, 1.2, 1.4, 1.6], and present the results from best one for each metric.

R.2. FID Plot

Fig. 55 shows the FID score in CelebA dataset. We can see that BSN outperforms the standard SN in all 5 hyper-parameter settings.

R.3. Training Curves

From App. N we can see that SN (no γ) and BSN generally have the best performance. Therefore, in this section, we focus on comparing these two algorithms with the training curves. Figs. 56 to 60 show the FID of these two algorithms during training. Generally, we see that BSN converges slower than SN *at the beginning of training*. However, as training proceeds, BSN finally has better metrics in all cases. Note that unlike CIFAR10, SN seems to be more stable in CelebA as its sample quality does not drop in most hyper-parameters. But the key conclusion is the same: in most cases, BSN not only outperforms SN at the end of training, but also outperforms the peak sample quality of SN during training (e.g. Figs. 56 to 59). The only exception is the $n_{dis} = 5$ setting, where both SN and BSN has instability issue: the sample quality first improves and then significantly drops. But even in this case, BSN has better final performance than the standard SN.

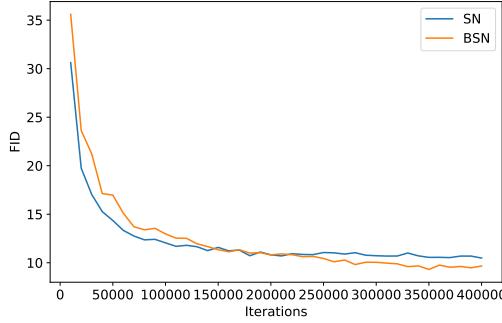


Figure 57. FID in CelebA. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

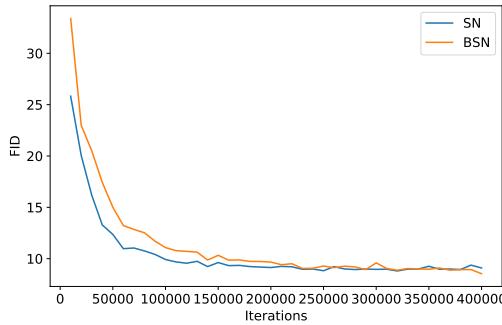


Figure 58. FID in CelebA. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$.

R.4. Generated Images

Figs. 61 to 64 show the generated images from the run with the best FID for each algorithm.

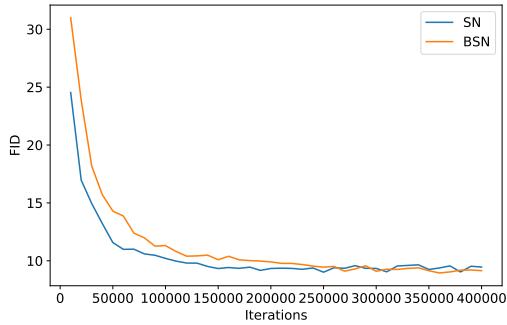


Figure 59. FID in CelebA. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0002$, $n_{dis} = 1$.

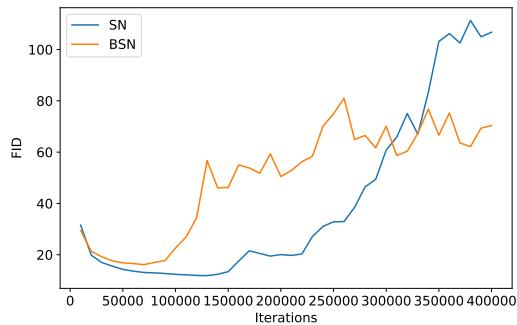


Figure 60. FID in CelebA. The results are averaged over 5 random seeds. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$.



Figure 61. Generated samples from the best run of SN (same γ) in CelebA. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. FID is 7.40.

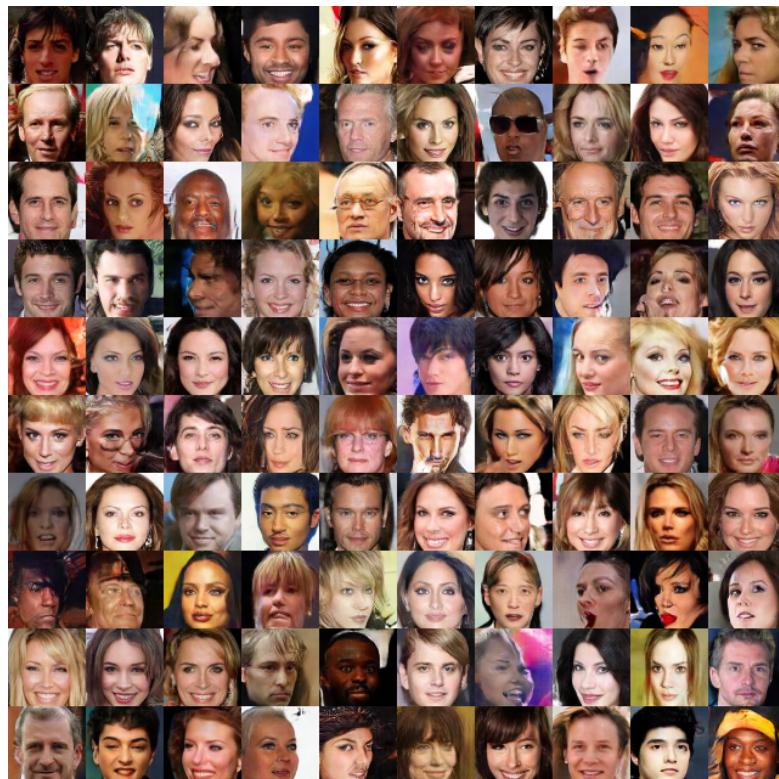


Figure 62. Generated samples from the best run of SN (diff. γ) in CelebA. The hyper-parameters are: $\alpha_g = 0.0001$, $\alpha_d = 0.0001$, $n_{dis} = 5$. FID is 7.29.

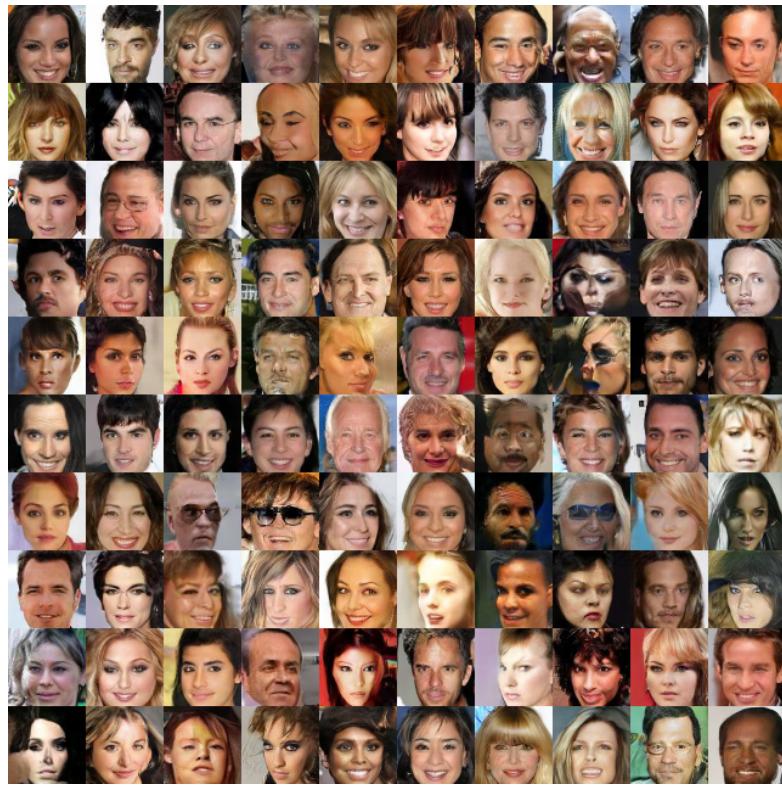


Figure 63. Generated samples from the best run of SN in CelebA. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$. FID is 8.34.



Figure 64. Generated samples from the best run of BSN in CelebA. The hyper-parameters are: $\alpha_g = 0.0002$, $\alpha_d = 0.0001$, $n_{dis} = 1$. FID is 8.06.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
Fully connected ($4 \times 4 \times 1024$).
ResNet-up ($c = 1024$).
ResNet-up ($c = 512$).
ResNet-up ($c = 256$).
ResNet-up ($c = 128$).
ResNet-up ($c = 64$).
BN. ReLU. Convolution ($c = 3, k = 3, s = 1$). Tanh

Table 8. Generator network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). BN stands for batch normalization. c stands for number of channels. k stands for kernel size. s stands for stride.

Direct connection
BN. ReLU. Unpooling(2). Convolution ($k = 3, s = 1$).
BN. ReLU. Convolution ($k = 3, s = 1$).
Shortcut connection
Unpooling(2). Convolution ($k = 1, s = 1$).

Table 9. ResNet-up network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). BN stands for batch normalization. k stands for kernel size. s stands for stride.

S. Experimental Details and Additional Results on ILSVRC2012

S.1. Experimental Details

The network architectures are shown in Tables 8 to 13. All experiments are run for 500k iterations. Discriminator batch size is 16. Generator batch size is 32. The optimizer is Adam. $\alpha_g = 0.002, \alpha_d = 0.002, \beta_1 = 0.0, \beta_2 = 0.9, n_{dis} = 5$ We use hinge loss with the popular SN implementation (Miyato et al., 2018).

S.2. Training Curves

Figs. 65 and 66 show the inception score and FID of SN and BSN during training.

For SN, we can see that the runs with scale=1.0/1.2/1.4 have similar performance throughout training. When scale=1.6, the performance is much worse.

For BSN, the runs with scale=1.2/1.4 perform better than SN runs throughout the training. When scale=1.6, BSN has similar performance as SN at the early stage of training, and is slightly better at the end. When scale=1.0, the performance is very bad as there is gradient vanishing problem.

$x \in \mathbb{R}^{128 \times 128 \times 3}$
ResNet-first ($c = 64$).
ResNet-down ($c = 128$).
ResNet-down ($c = 256$).
ResNet-down ($c = 512$).
ResNet-down ($c = 1024$).
ResNet ($c = 1024$).
ReLU. Global pooling. Fully connected (1).

Table 10. Discriminator network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). BN stands for batch normalization. c stands for number of channels. k stands for kernel size. s stands for stride.

Direct connection
ReLU. Convolution ($k = 3, s = 1$).
ReLU. Convolution ($k = 3, s = 1$). Average pooling(2).
Shortcut connection
Convolution ($k = 1, s = 1$). Average pooling(2).

Table 11. ResNet-down network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). k stands for kernel size. s stands for stride.

Direct connection
Convolution ($k = 3, s = 1$).
ReLU. Convolution ($k = 3, s = 1$). Average pooling(2).
Shortcut connection
Average pooling(2). Convolution ($k = 1, s = 1$).

Table 12. ResNet-first network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). k stands for kernel size. s stands for stride.

Direct connection
ReLU. Convolution ($k = 3, s = 1$).
ReLU. Convolution ($k = 3, s = 1$).
Shortcut connection
Convolution ($k = 1, s = 1$).

Table 13. ResNet network architectures for ILSVRC2012 experiments (from (Miyato et al., 2018)). k stands for kernel size. s stands for stride.

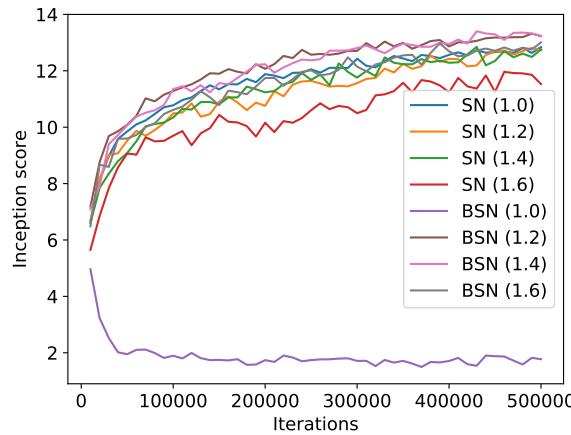


Figure 65. Inception score in ILSVRC2012. The results are averaged over 5 random seeds.

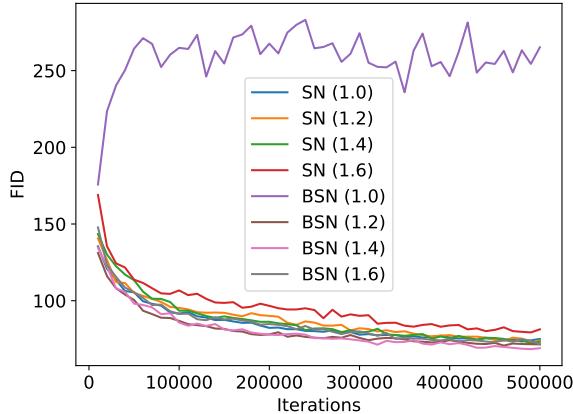


Figure 66. FID in ILSVRC2012. The results are averaged over 5 random seeds.

S.3. Generated Images

Figs. 67 to 74 show the generated images from the run with the best inception score for SN and BSN with different scale parameters.

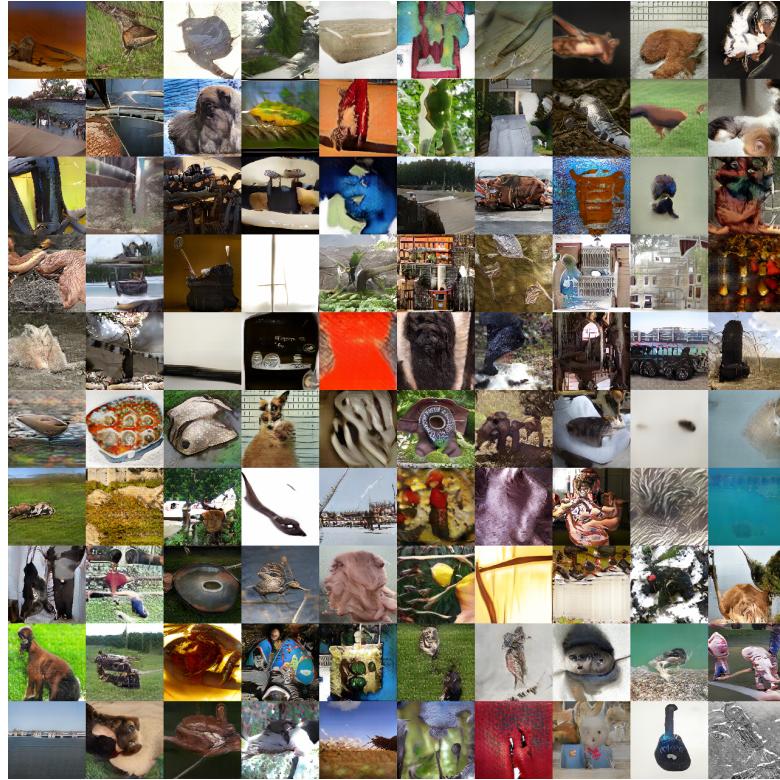


Figure 67. Generated samples from the best run of SN (scale=1.0) in ILSVRC2012. Inception score is 13.50. FID is 72.18.



Figure 68. Generated samples from the best run of SN (scale=1.2) in ILSVRC2012. Inception score is 13.04. FID is 72.51.

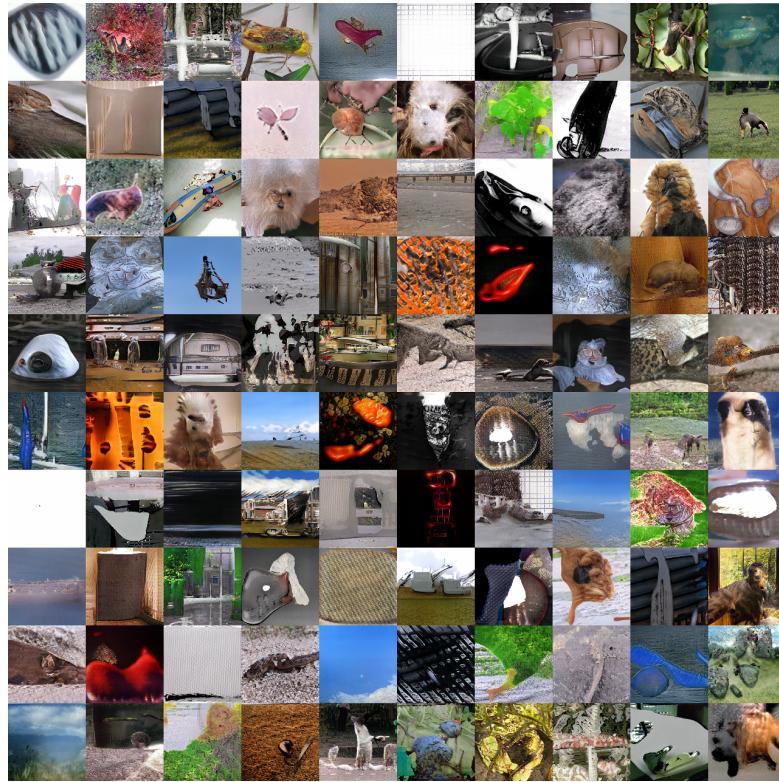


Figure 69. Generated samples from the best run of SN (scale=1.4) in ILSVRC2012. Inception score is 13.04. FID is 69.12.



Figure 70. Generated samples from the best run of SN (scale=1.6) in ILSVRC2012. Inception score is 12.62. FID is 70.36.

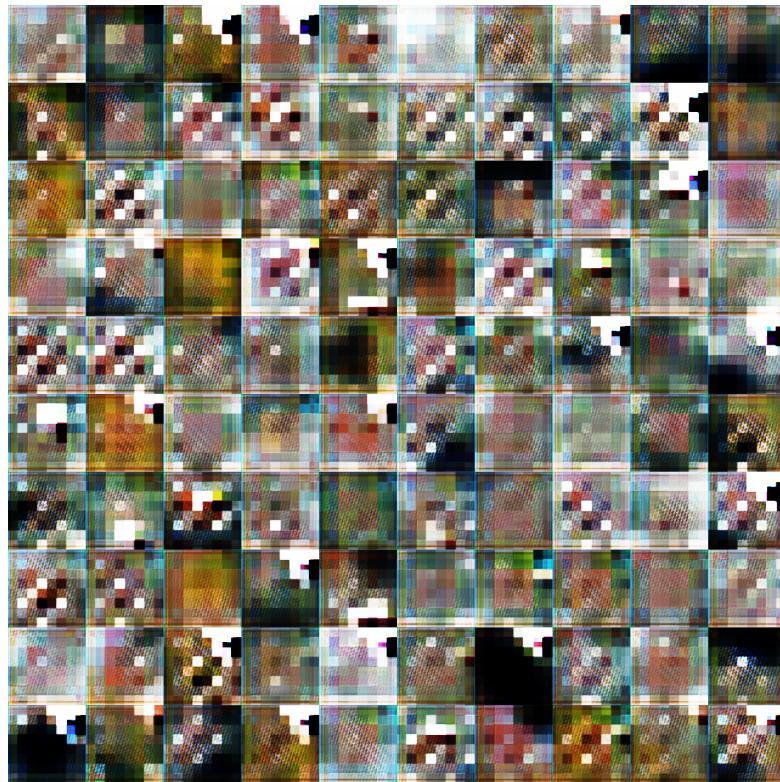


Figure 71. Generated samples from the best run of BSN (scale=1.0) in ILSVRC2012. Inception score is 2.07. FID is 242.51.

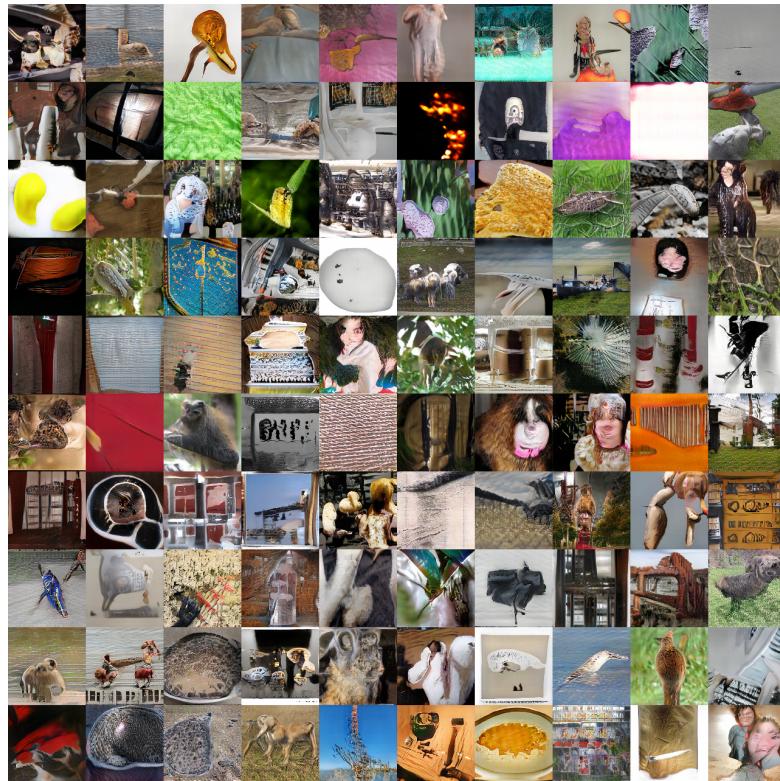


Figure 72. Generated samples from the best run of BSN (scale=1.2) in ILSVRC2012. Inception score is 13.55. FID is 71.30.

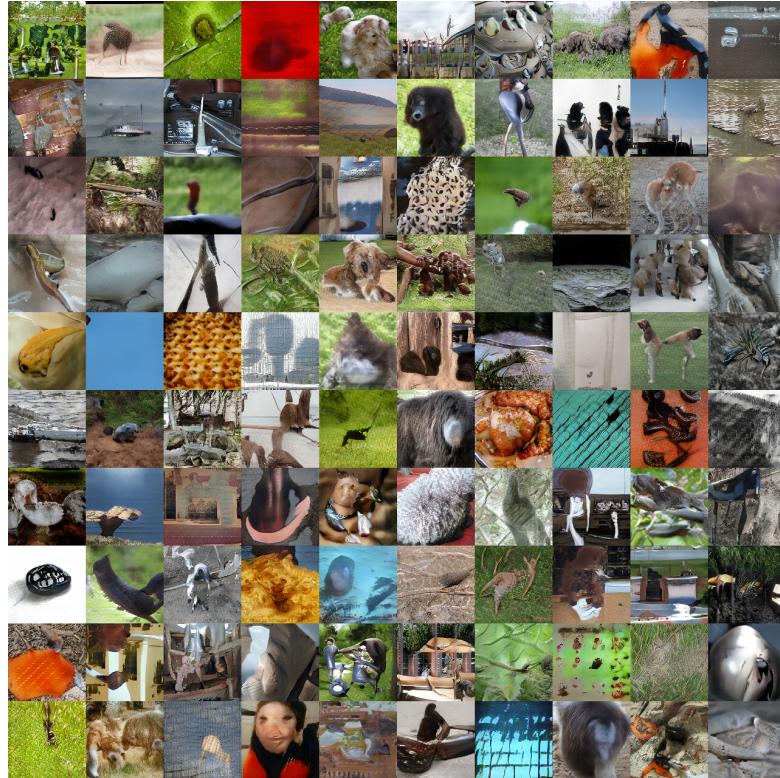


Figure 73. Generated samples from the best run of BSN (scale=1.4) in ILSVRC2012. Inception score is 13.63. FID is 70.88.



Figure 74. Generated samples from the best run of BSN (scale=1.6) in ILSVRC2012. Inception score is 13.24. FID is 69.06.