Spectral Regularization for Combating Mode Collapse in GANs

Kanglin Liu[1,2,3], Wenming Tang[1,2,3], Fei Zhou[1,2,3], Guoping Qiu[1,2,3,4]
[1] Shenzhen University, Shenzhen, China
[2] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, China
[3] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[4] University of Nottingham, Nottingham, United Kingdom
max.liu.426@gmail.com, guoping.qiu@nottingham.ac.uk

Abstract

Despite excellent progress in recent years, mode collapse remains a major unsolved problem in generative adversarial networks (GANs). In this paper, we present spectral regularization for GANs (SR-GANs), a new and robust method for combating the mode collapse problem in GANs. Theoretical analysis shows that the optimal solution to the discriminator has a strong relationship to the spectral distributions of the weight matrix. Therefore, we monitor the spectral distribution in the discriminator of spectral normalized GANs (SN-GANs), and discover a phenomenon which we refer to as spectral collapse, where a large number of singular values of the weight matrices drop dramatically when mode collapse occurs. We show that there are strong evidence linking mode collapse to spectral collapse; and based on this link, we set out to tackle spectral collapse as a surrogate of mode collapse. We have developed a spectral regularization method where we compensate the spectral distributions of the weight matrices to prevent them from collapsing, which in turn successfully prevents mode collapse in GANs. We provide theoretical explanations for why SR-GANs are more stable and can provide better performances than SN-GANs. We also present extensive experimental results and analysis to show that SR-GANs not only always outperform SN-GANs but also always succeed in combating mode collapse where SN-GANs fail.

1. Introduction

Generative Adversarial Networks (GANs) [5] are one of the most significant developments in machine learning research of the past decade. Since their first introduction, GANs have attracted intensive interest in the machine learning community not only for their ability to learn highly structured probability distributions but also for their theoretically implications [5, 13, 2, 17]. Essentially, GANs are constructed around two functions [3, 9]: the generator $\mathbf{G}$, which maps a sample z to the data distribution, and the discriminator $\mathbf{D}$, which is trained to distinguish real samples of a dataset from fake samples produced by the generator. With the goal of reducing the difference between the distributions of generated and real samples, a GAN training algorithm trains $\mathbf{G}$ and $\mathbf{D}$ in tandem.

GAN training is dynamic and sensitive to nearly every aspect of its setup, from optimization parameters to model architecture [1]. Training instability, or mode collapse, is one of the major obstacles in developing applications. Despite excellent progresses in recent years [6, 12, 10, 15, 7], the mode collapse problem still persists. For example, one of the most impressive works to emerge recently is BigGANs [1], which is the largest published GAN system based on the state of the art Spectral Normalization (SN-GAN)[10]. However, BigGANs can still suffer from the training instability problem, especially when the batch size is scaled up. Although implementing training stabilization measures such as employing R_1 zero-centred gradient penalty term [1] in the loss metric of the discriminator to prevent spectral noise can improve stability, this can cause severe degradation in performance, resulting in a 45% reduction in Inception Score.

In this paper, we present Spectral Regularization, a robust method for combating the mode collapse problem in GANs. Theoretically, we analyze the optimal solution to a linear discriminator function constrained by 1-Lipschitz continuity, and find the optimal solution is taken when all singular values of weight matrix are 1. Even though, in the implementation of GAN models, $\mathbf{D}$ is non-linear, we reason that the spectral distributions in $\mathbf{D}$ may also have a strong relation to its performance. Through comprehensive analysis of spectral distributions in a large number of GAN models trained with the state of the art SN-GAN algorithm, we discover that when mode collapse occurs to a model, spectral distributions of $\overline{W}_{\mathrm{SN}}(W)$ in $\mathbf{D}$ also collapse, where $\overline{W}_{\mathrm{SN}}(W)$ is spectral normalized weight ma-

trix. Specifically, we observe that when a model performs well and no mode collapse occurs, there are a large number of singular values of $\overline{W}_{\mathrm{SN}}(W)$ in $\mathbf{D}$ very close to 1, and that when mode collapse occurs to a model, singular values of $\overline{W}_{\mathrm{SN}}(W)$ in $\mathbf{D}$ will drop dramatically. We refer to the phenomenon where a large number of singular values drop significantly as **spectral collapse**.

In all GAN models of various sizes and trained with a variety of parameter settings on datasets extensively used in the literature, we observe that mode collapse and spectral collapse always go side by side. This fact leads us to reason that mode collapse in SN-GANs is caused by spectral collapse in $\mathbf{D}'\mathbf{s}$ weight matrices. Based on such insight into spectral distributions of $\overline{W}_{\mathrm{SN}}(W)$, we propose a new and robust method called spectral regularization to prevent GANs from mode collapse. In addition to normalizing the weight matrices, spectral regularization imposes constraints on $\mathbf{D}'\mathbf{s}$ weight matrices by compensating their spectral distributions to avoid spectral collapse. Theoretical analysis shows that spectral regularization is better than spectral normalization at preventing weight matrix from concentrating into one particular direction. We show that SN-GANs are a special case of spectral regularization, and in a series of extensive experiments we demonstrate that spectral regularization not only provides superior performances to spectral normalization but also can always avoid mode collapse in cases where spectral normalization failed.

Our contributions can be summarized as follows:

(1) Through theoretical analysis and extensive experimental observations, we provide an insight into the likely causes of mode collapse in a state of the art GAN normalization technique, spectral normalization (SN-GANs). We introduce the concept of **spectral collapse** and provide strong evidence to link spectral collapse with mode collapse in SN-GANs.

(2) Based on above insight, we have developed a new robust regularization method, **Spectral Regularization**, where we compensate the spectral distributions of the weight matrices in $\mathbf{D}$ to prevent spectral collapse, thus preventing mode collapse in GANs. Extensive experimental results show that spectral regularization not only can always prevent mode collapse but also can consistently provide improved performances over SN-GANs.

2. Analysis of Mode Collapse in SN-GANs

2.1. A Brief Summary of SN-GANs

For easy discussion, we first briefly recap the essential ideas of the spectral normalization technique for training GANs [10]. As far we are aware, this is currently one of the best methods in the literature and has been successfully used to construct large systems such as BigGANs [1] . For convenience, we largely follow the notation convention of [10]. Considering a simple discriminator of a neural network of the following form:

$$f(x,\theta) = W^{L+1}(a_L \cdot W^L \cdot a_{L-1} \cdot W^{L-1} \cdots a_1 W^1 x) \quad (1)$$

where $\theta := \{W^1, \cdots, W^L, W^{L+1}\}$ is the learning parameters set, $W^l \in \mathbb{R}^{d_l \times d_{l-1}}$, $W^{L+1} \in \mathbb{R}^{1 \times d_L}$, and a_l is an element-wise non-linear activation function. We omit the bias term of each layer for simplicity. The final output of the discriminator is given by

$$D(x,\theta) = \mathcal{A}(f(x,\theta)) \quad (2)$$

where $\mathcal{A}$ is an activation function corresponding to the divergence of a distance measure of users' choice.

The standard formulation of GANs is given by [10, 13]:

$$\min_G \max_D V(G,D) \quad (3)$$

where min and max of G and D are taken over the set of the generator and discriminator functions respectively. The conventional form of $V(G, D)$ is given by $E_{x\sim q_{data}}[\log D(x)] + E_{x'\sim q_G}[\log(1 - D(x'))]$ [10], where q_{data} is the data distribution and q_G is the model (generator) distribution.

To guarantee Lipschitz continuity, spectral normalization [10] controls the Lipschitz constant of the discriminator function by literally constraining the spectral norm of each layer:

$$\overline{W}_{\mathrm{SN}}(W) := W/\sigma(W) \quad (4)$$

where σ(W) is the spectral norm of the weight matrix W in the discriminator network, which is equivalent to the largest singular value of W.

The authors of SN-GANs [10] and those of BigGANs [1] have demonstrated the superiority of spectral normalization over other normalization or regularization techniques, e.g., gradient penalty [6], weight normalization[15] and orthonormal regularization [4]. However, as a state of the art GAN model, BigGANs (based on spectral normalization) can still suffer from mode collapse. Therefore, mode collapse remains an unsolved open problem, seeking better and more robust solution is very important for advancing GANs.

2.2. Theoretical Analysis

In order to unearth the likely causes of mode collapse, we start by analyzing the optimal solution to 1-Lipschitz constrained discriminator.

To be specific, Proposition 1 in [6] has proven that the optimal solution to 1-Lipschitz discriminator function f^* has gradient norm 1 almost everywhere. Assuming the discriminator f is a linear function, we find that the optimal solution is obtained only when all the singular values are 1. This can be verified by Corollary 1 (see proof in Appendix).

Corollary 1. Let P_r and P_g be two distributions in $\mathcal{X}$, a compact metric space. A linear and 1-Lipschitz constrained function $f^* = Wx$, is the optimal solution of

$max_{\|f\|_{Lip}\leqslant 1} E_{x\sim P_r}[f(x)] - E_{x\sim P_g}[f(x)]$. Then all the singular values of the weight matrix W are 1.

We can see that, for a linear f, the spectral distribution is strongly related to the performance of $\mathbf{D}$. For discriminators in GANs, f is nonlinear. However, we reason that their spectral distributions may also have a strong relation to the performance of discriminator. As a result, we can monitor the spectral distribution to investigate the mode collapse problem.

2.3. Mode Collapse vs Spectral Collapse

In order to find the link between mode collapse and spectral distributions, we have conducted a series of experiments for unconditional image generation on CIFAR-10 [16] and STL-10 [8] datasets. Our implementation is based on the SN-GANs architecture of [10], which uses the hinge loss as the discriminator objective and is given by:

$$\begin{aligned} L_D =& E_{x\sim q_{data}}[\min(0, -1 + D(x))] \\ &+ E_{x\sim q_G}[\min(0, -1 - D(x))] \end{aligned} \tag{5}$$

The optimization settings follow literature [10, 11]. Previous authors have shown that increasing batch size or decreasing discriminator capacity could potentially lead to mode collapse [1]. We therefore conduct experiments for various combinations of batch and channel sizes as listed in Table 1. We follow the practices in the literature of using Inception Score (IS) [14] and Fréchet Inception Distance (FID) [8] as approximate measures of sample quality, and results are shown in Table 2 where we also identify all settings where mode collapse has occurred to SN-GANs. Through monitoring Inception Scores, Fréchet Inception Distance and synthetic images during training, mode collapse is observed in 10 settings including $B_{64-64}, B_{128-64}, B_{256-64}$, $C_{8-32}, C_{16-32}, C_{32-32}, C_{64-32}, C_{128-32}$, E_{256-64} and E_{256-32}. In other 16 settings, mode collapse has not happened.

Mode collapse is a persistent problem in GAN training and is also a major issue in SN-GANs as has been shown in BigGANs[1] and in Table 2. Here, we monitor the entire spectral distributions of SN-GANs, i.e., all singular values of $\overline{W}_{\text{SN}}(W)$ in the discriminator network during training.

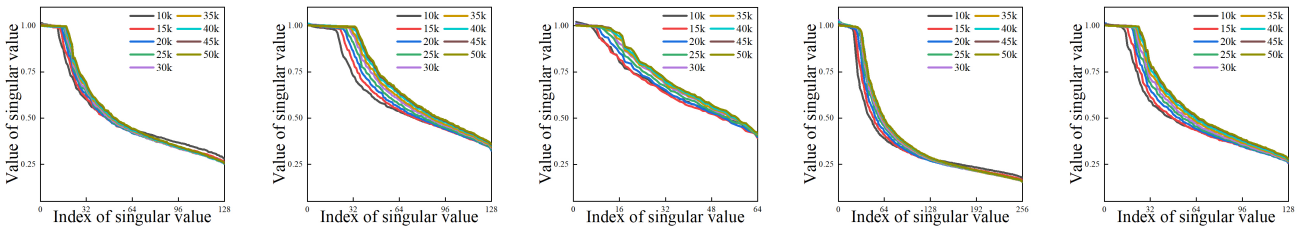
The discriminator network in our implementation uses the same architecture as that in the original SN-GANs[10] and has 10 convolutional layers, please see Appendix for the setting details. In order to discover the likely causes of mode collapse, we plot the spectral distributions of every layer (except skip connection layers) of the discriminator for all 26 settings. In the following, we present some typical examples and readers are referred to the Appendix for all other plots.

Figure 1 shows the spectral distributions of $layer\_9$ of 5 settings where mode collapse does not happen. Figure 2 shows the spectral distributions of $layer\_9$ of all 10 settings where mode collapse has occurred. Through analyzing the spectral distribution plots in Figure 1 and Figure 2, we notice a very interesting pattern. In the cases where no mode collapse happens, the shapes of the spectral distribution curves do not change significantly with the number of training iteration. On the other hand, for those settings where mode collapse has occurred, the shapes of the spectral distribution curves change significantly as training progresses. In particular, a large number of singular values become very small when training passes a certain number of iterations. This is as if the curves have "collapsed", and we refer to this phenomenon as **spectral collapse**.

The phenomenon of spectral collapse is also observed across different settings. Figure 3 plots the spectral distributions of the 5 groups of experimental settings in Table 1. It is seen that in groups A and D, the spectral distributions across different settings are very similar and no spectral collapse is observed. Very interestingly, no mode collapse is observed either. In group B, the spectral distributions of B_{64-64}, B_{128-64} and B_{256-64} have collapsed, not surprisingly, mode collapse also happens to these 3 settings. In group C, the spectral distributions of all settings have collapsed, i.e., most singular values are very small (except for the first one which is forced to be 1 by spectral normalization). Again as expected, mode collapse happens to all settings in this group. In group E, it is seen that the two settings E_{256-64} and E_{256-32} have suffered from spectral collapse. Again, mode collapse is observed for these two settings.

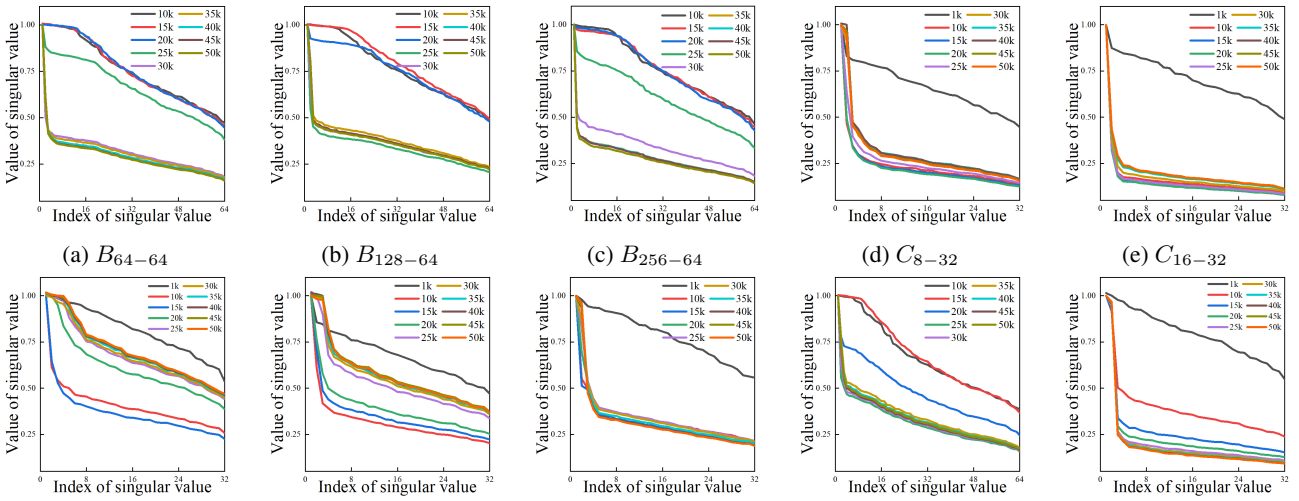
Setting	Batch	CH	Dataset	Setting	Batch	CH	Dataset
A_{16-128}	16	128	CIFAR-10	C_{8-32}	8	32	CIFAR-10
A_{32-128}	32	128	CIFAR-10	C_{16-32}	16	32	CIFAR-10
A_{64-128}	64	128	CIFAR-10	C_{32-32}	32	32	CIFAR-10
$A_{128-128}$	128	128	CIFAR-10	C_{64-32}	64	32	CIFAR-10
$A_{256-128}$	256	128	CIFAR-10	C_{128-32}	128	32	CIFAR-10
$A_{512-128}$	512	128	CIFAR-10	$D_{128-256}$	128	256	CIFAR-10
$A_{1024-128}$	1024	128	CIFAR-10	$D_{256-256}$	256	256	CIFAR-10
B_{8-64}	8	64	CIFAR-10	$D_{512-256}$	512	256	CIFAR-10
B_{16-64}	16	64	CIFAR-10	E_{16-128}	16	128	STL-10
B_{32-64}	32	64	CIFAR-10	E_{64-128}	64	128	STL-10
B_{64-64}	64	64	CIFAR-10	$E_{256-128}$	256	128	STL-10
B_{128-64}	128	64	CIFAR-10	E_{256-64}	256	64	STL-10
B_{256-64}	256	64	CIFAR-10	E_{256-32}	256	32	STL-10

Table 1. Experiment settings. The experiments are divided into 5 groups A, B, C, D and E. Within each group, the models share exactly the same network architecture but differ in batch size. For groups $A - D$, we vary the batch sizes inside each group to study how batch sizes relate to mode collapse, and we change the channel sizes between groups to investigate how discriminator capacity affects mode collapse. Group E is experiments applied to a different data set. The purpose is to evaluate how different data affect mode collapse. Batch represents the batch size. CH is the channel size of the discriminator. The subscript of each group name annotates the batch and channel setting of that experiment, e.g., A_{a-b} represents setting with a batch size a and a CH size b.



(a) A_{64-128} (b) $A_{256-128}$ (c) B_{64-64} (d) $D_{128-256}$ (e) $E_{256-128}$

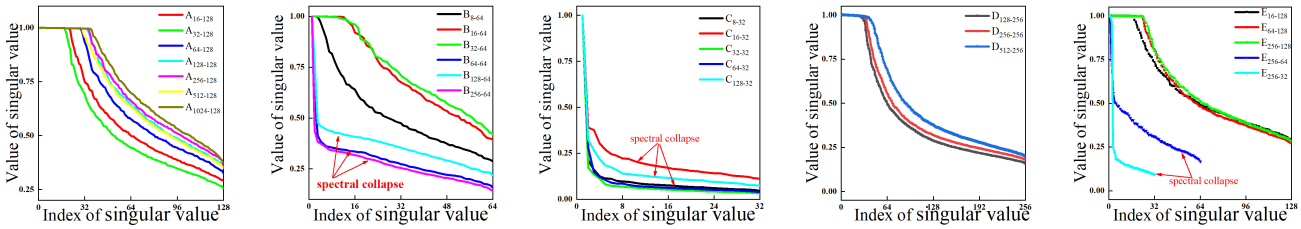
Figure 1. Spectral distributions in $layer\_9$ for Good GANs (no mode collapse) at different number of iterations. The curves represent the spectral distributions after $10k$ iterations, $15k$ iterations, ..., and $50k$ iterations.



(a) B_{64-64} (b) B_{128-64} (c) B_{256-64} (d) C_{8-32} (e) C_{16-32}

(f) C_{32-32} (g) C_{64-32} (h) C_{128-32} (i) E_{256-64} (j) E_{256-32}

Figure 2. Spectral distributions in $layer\_9$ for settings where mode collapse occurs. The curves represent the spectral distributions after $1k$ iterations, $10k$ iterations, ..., and $50k$ iterations.



(a) group A (b) group B (c) group C (d) group D (e) group E

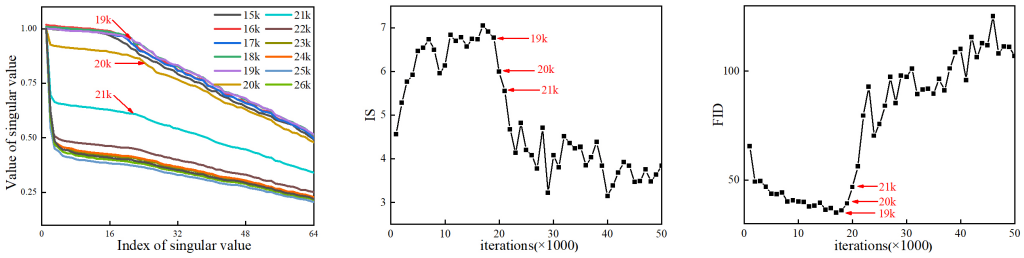
Figure 3. Spectral distributions (after $50k$ iterations) in $layer\_9$ for different settings.

In order to understand what has happened when spectral collapse occurs, Figure 4 shows how a typical spectral distribution relates to Inception Score and Fréchet Inception Distance during training. It is seen that up to 19k iterations both IS and FID are showing good performances, and the corresponding spectral distribution has a large number of large singular values. At 20k iterations, IS and FID performances start to drop, correspondingly, the spectral distribution starts to fall. At 21k iterations, the IS and FID performances have dropped significantly and mode collapse has started, and very importantly, the spectral distribution has dropped dramatically - starting to collapse.

The association of mode collapse with spectral collapse is observed for all the layers and on all settings (readers are referred to the Appendix for more examples). We therefore believe that mode collapse and spectral collapse happen at the same time, and spectral collapse is the likely cause of mode collapse. In the following section, we will introduce spectral regularization to prevent spectral collapse thus avoiding mode collapse.

3. Spectral Regularization

We have now established that spectral collapse is closely linked to mode collapse in SN-GANs. In this section, we introduce spectral regularization, a technique for preventing spectral collapse. We show that preventing spectral collapse



(a) spectral distributions (b) Inception Score (c) Fréchet Inception Distance

Figure 4. An example showing how spectral distributions relate to Inception Score and Fréchet Inception Distance. Here the setting is B_{128-64} and the spectral distributions correspond to those of $layer\_9$.

can indeed solve the mode collapse problem, thus demonstrating that spectral collapse is the cause of mode collapse rather than a mere symptom.

Performing singular value decomposition, the weight matrix W can be expressed as:

$$W = U \cdot \Sigma \cdot V^T \tag{6}$$

where both U and V are orthogonal matrix, the columns of U, $[u_1, u_2, \cdots, u_m]$, are called left singular vectors of W, the columns of V, $[v_1, v_2, \cdots, v_n]$, are called right singular vectors of W, and Σ can be expressed as:

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{7}$$

where $D = diag\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ represents the spectral distribution of W.

When mode collapse occurs, spectral distributions concentrate on the first singular value, and the rest singular values drop dramatically (spectral collapse). To avoid spectral collapse, we first apply ΔD to compensate D, where ΔD is given by $diag\{\sigma_1 - \sigma_1, \sigma_1 - \sigma_2, \cdots, \sigma_1 - \sigma_i, 0, \cdots, 0\}$, and i is a hyperparameter ($1 \leq i \leq r$). Spectral regularization turns D into D' as follows: $D' = D + \Delta D = diag[\sigma_1, \cdots, \sigma_1, \sigma_{i+1}, \cdots, \sigma_r]$. Correspondingly, W turns to W': $W' = W + \Delta W$, where ΔW is given by:

$$\Delta W = U \cdot \begin{bmatrix} \Delta D & 0 \\ 0 & 0 \end{bmatrix} \cdot V^T = \sum_{k=2}^{i} (\sigma_1 - \sigma_k) u_k v_k^T \tag{8}$$

Finally, we apply spectral normalization to guarantee Lipschitz continuity, and obtain our spectral regularized $\overline{W}_{\mathrm{SR}}(W)$:

$$\overline{W}_{\mathrm{SR}}(W) = \frac{W + \Delta W}{\sigma(W)} = \overline{W}_{\mathrm{SN}}(W) + \Delta W / \sigma(W) \tag{9}$$

Clearly, spectral normalization is a special case of spectral regularization (when $i = 1$).

3.1. Gradient Analysis of Spectral Regularization

We perform gradient analysis to show that spectral regularization provides a more effective way over spectral normalization in preventing W from concentrating into one particular direction during training and thus avoiding spectral collapse.

From equation (9), we can write the gradient of $\overline{W}_{\mathrm{SR}}(W)$ with respect to W_{ab} as:

$$\begin{aligned} \frac{\partial \overline{W}_{\mathrm{SR}}(W)}{\partial W_{ab}} &= \frac{1}{\sigma(W)} \{E_{ab} - \overline{W}_{\mathrm{SN}}[u_1 v_1^T]_{ab} \\ &- \frac{\Delta W}{\sigma(W)} [u_1 v_1^T]_{ab} + \sum_{k=2}^{i} [u_1 v_1^T - u_k v_k^T]_{ab} \cdot u_k v_k^T \} \end{aligned} \tag{10}$$

where $[\cdot]_{ab}$ represents the (a, b)-th entry of corresponding matrix, E_{ab} is the matrix whose (a, b)-th entry is 1 and zero everywhere else.

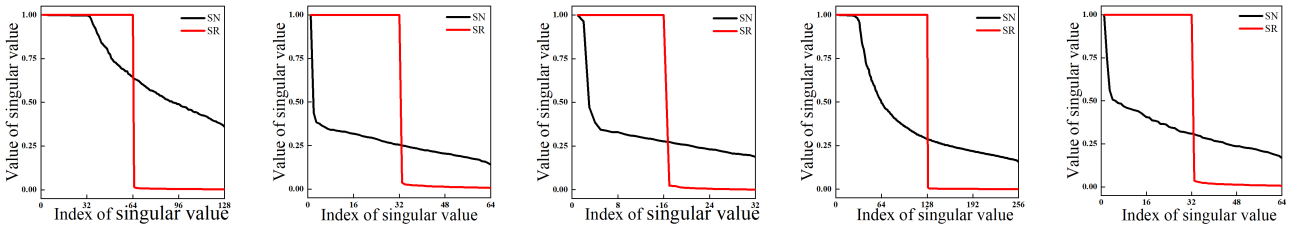
We would like to comment on the implication of equation (10). The first two terms, $E_{ab} - \overline{W}_{\mathrm{SN}}[u_1 v_1^T]_{ab}$, are the gradient of spectral normalization $\frac{\partial \overline{W}_{\mathrm{SN}}(W)}{\partial W_{ab}}$ [10], this is very easy to see from equation (9). As explained in [10], the second term can be regarded as being able to prevent the columns space of W from concentrating into one particular direction in the course of training. In other words, spectral normalization prevents the transformation of each layer from becoming sensitive only in one direction. However, as we have seen (e.g. Figure 2), despite performing spectral normalization, the spectral distributions of $\overline{W}_{\mathrm{SN}}(W)$ can still concentrate on the first singular value thus causing spectral collapse. This shows the limited ability of spectral normalization in preventing W from spectral collapse.

In addition to the first two terms of spectral normalization, spectral regularization introduces the third and fourth terms in equation (10). It can be seen that the third term enhances the effect of the second term, through which W is much less likely to concentrate into one particular direction. Furthermore, the fourth term can be seen as the regularization term, encouraging W to move along all i directions pointed to by $u_k v_k^T$, for $k = 1, 2, ..., i$, each weighted by the adaptive regularization coefficient $[u_1 v_1^T - u_k v_k^T]_{ab}$. This encourages W to make full use of the directions pointed to by $u_j v_j^T$, thus preventing W from being concentrated on only 1 direction, which in turn stabilizes the training process.

From above analysis, it is clear that as compared to spectral normalization, spectral regularization of equation (10) encourages W of the discriminator to move in a variety

cExperimentSetting	IS		FID		MC	SC	cExperimentSetting	IS		FID		MC	SC
	SN	SR	SN	SR				SN	SR	SN	SR		
A_{16-128}	8.15±.09	8.35±.09	22.31±.28	24.67±.28	×	×	C_{8-32}	4.21±.18	4.93±.20	80.00±1.12	66.05±2.12	**SN**	**SN**
A_{32-128}	8.38±.07	8.45±.10	25.96±.42	22.00±.17	×	×	C_{16-32}	4.05±.15	4.78±.23	79.69±.21	59.25±.43	**SN**	**SN**
A_{64-128}	8.39±.15	8.65±.12	21.15±.15	20.31±.18	×	×	C_{32-32}	4.29±.08	4.70±.15	78.39±.17	62.10±.24	**SN**	**SN**
$A_{128-128}$	8.61±.12	8.72±.08	21.01±.23	19.98±.19	×	×	C_{64-32}	4.30±.14	5.00±.14	85.15±1.20	56.11±.54	**SN**	**SN**
$A_{256-128}$	8.45±.14	8.48±.03	20.87±.25	19.87±.21	×	×	C_{128-32}	4.87±.14	5.30±.07	71.10±.89	54.39±.41	**SN**	**SN**
$A_{512-128}$	8.34±.09	8.53±.04	21.85±.14	20.13±.12	×	×	$D_{128-256}$	**8.14±.06**	**8.92±.18**	**24.43±.41**	**18.95±.23**	×	×
$A_{1024-128}$	8.31±.21	8.52±.16	21.68±.35	20.34±.13	×	×	$D_{256-256}$	8.29±.12	8.83±.14	22.54±.29	19.56±.11	×	×
B_{8-64}	6.67±.05	7.42±.06	45.19±.89	35.78±.11	×	×	$D_{512-256}$	8.33±.09	8.36±.12	22.58±.16	21.82±.29	×	×
B_{16-64}	7.34±.06	7.59±.08	31.73±.49	29.42±.22	×	×	E_{16-128}	8.63±.15	8.69±.16	44.24±.56	43.19±.33	×	×
B_{32-64}	7.18±.03	7.48±.09	33.76±.35	28.60±.25	×	×	E_{64-128}	8.98±.20	9.14±.18	42.40±.56	39.89±.89	×	×
B_{64-64}	6.96±.11	7.52±.11	36.65±.29	28.40±.36	**SN**	**SN**	$E_{256-128}$	9.10±.13	9.11±.17	40.11±.89	40.08±.29	×	×
B_{128-64}	7.10±.14	7.13±.05	35.99±.48	31.41±.56	**SN**	**SN**	E_{256-64}	7.38±.14	7.67±.06	74.50±1.52	69.20±.83	**SN**	**SN**
B_{256-64}	6.85±.08	7.58±.03	35.88±.42	27.68±.23	**SN**	**SN**	E_{256-32}	4.04±.11	4.38±.07	98.50±1.34	89.17±1.23	**SN**	**SN**

Table 2. IS and FID results for different settings, where IS is Inception Score and FID is Fréchet Inception Distance. For IS, higher is better, while lower is better for FID. SN, SR represent Spectral normalization and Spectral Regularization, respectively. MC stands for mode collapse, and SC stands for spectral collapse, × represents that no mode collapse or spectral collapse occurs. **SN** in the MC column or SC column represents that mode collapse or spectral collapse occurred to spectral normalization. Note that neither mode collapse nor spectral collapse happen to spectral regularization for all settings.



(a) $A_{128-128}$ (b) B_{256-64} (c) C_{128-32} (d) $D_{128-256}$ (e) E_{256-64}

Figure 5. The effect of SN-GANs and SR-GANs algorithms on spectral distributions. The plots show the spectral distributions of the weight matrix in $layer\_9$. Spectral collapse and mode collapse have happened to SN-GANs in (b), (c), and (e). In all cases, there is no spectral collapse and mode collapse in SR-GANs.

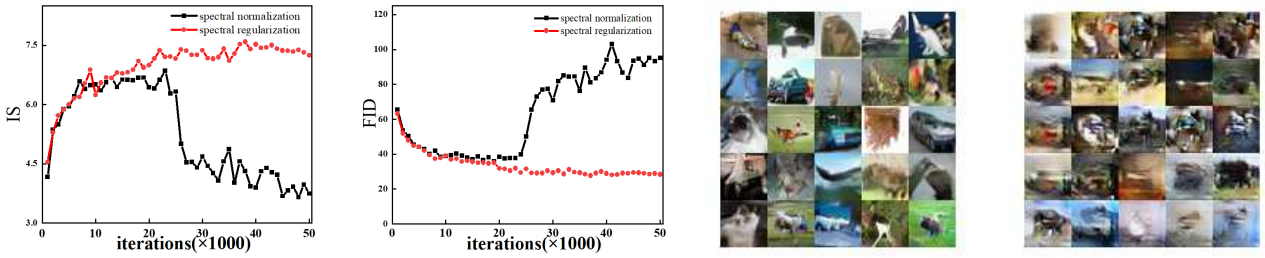
of directions thus preventing it from concentrating only on one direction, which in turn prevents spectral collapse. We will show in the experimental section that performing spectral regularization can indeed prevent mode collapse where spectral normalization has failed.

4. Experiments

For all settings listed in Table 1, we have conducted experiments using SN-GANs and the newly introduced spectral regularization algorithm (we use the abbreviation: SR-GANs for the spectral regularized GANs). All procedures and settings for SN-GANs and SR-GANs are identical, except that for SR-GANs the last discriminator update implements spectral regularization (equation 9) and SN-GANs implement spectral normalization (equation 4). The default value of the hyperparameter i in SR-GANs is empirically set as $i = 0.5N$, where N is the number of singular values in the corresponding weight matrix. Readers are referred to Appendix for the details of the network architecture settings.

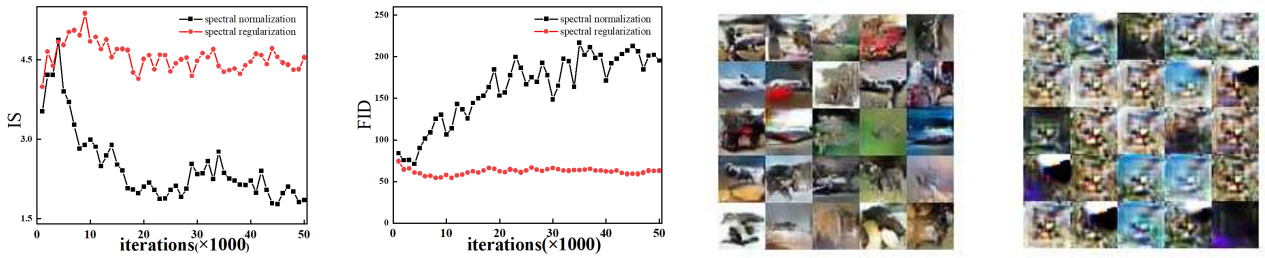
The Inception Score (IS) and Fréchet Inception Distance (FID) performances are shown in Table 2. Please note that in the cases where mode collapse have happened, IS and FID are the best results before mode collapse. It is clearly seen that in all cases, SR-GANs outperforms SN-GANs. In particular, for the setting of $D_{128-256}$, SR-GAN has improved IS by 9.5% and FID by 22.4%. On average, SR-GANs have improved the IS by 8.9% and FID by 18.9% over SN-GANs. Very importantly, in all 10 settings where mode collapse has occurred to SN-GANs, none has happened to SR-GANs. In fact, we have not yet observed mode collapses in an extensive set of experiments. We therefore have demonstrated that the new SR-GANs is superior to SN-GANs in both quality and stability.

Figure 5 shows an example of how the spectral distributions of the weights of the discriminator are affected by SN-GANs and SR-GANs. It is seen that SN-GANs normalize the largest singular value. However, in some cases, it cannot stop other singular values to drop significantly thus causing spectral collapse which in turn results in mode collapse. In contrast, SR-GANs ensures that the first i singular values are 1 in all cases, thus ensuring that spectral collapse would not happen hence preventing mode collapse. Similar effects



(a) Inception Score (b) Fréchet Inception Distance (c) synthetic images with SR (d) synthetic images with SN

Figure 6. Inception Score, Fréchet Inception Distance and synthetic images of SN-GANs and SR-GANs for the setting B_{256-64}



(a) Inception Score (b) Fréchet Inception Distance (c) synthetic images with SR (d) synthetic images with SN

Figure 7. Inception Score, Fréchet Inception Distance and synthetic images of SN-GAN and SR-GAN for the setting C_{128-32}

are observed in all layers and for all settings. This illustrates that SR-GANs can indeed prevent spectral collapse which in turn avoid mode collapse.

A combination of large batch and small channel sizes can easily cause SN-GANs to suffer from mode collapse. An example is B_{256-64} in our experiment. Figure 6 (a) and Figure 6 (b) show the changes of IS and FID measures of this setting during training. It is seen that after about 20k iterations, the performance of SN-GAN has started to drop and eventually lead to mode collapse. In contrast, the performance of SR-GAN is improved steadily as training progresses. Importantly, no mode collapse has occurred. Figure 6 (c) and Figure 6 (d) show some example images generated by SN-GAN and SR-GAN of this setting. It is clearly seen from Figure 6 (d) that mode collapse has indeed occurred to SN-GAN.

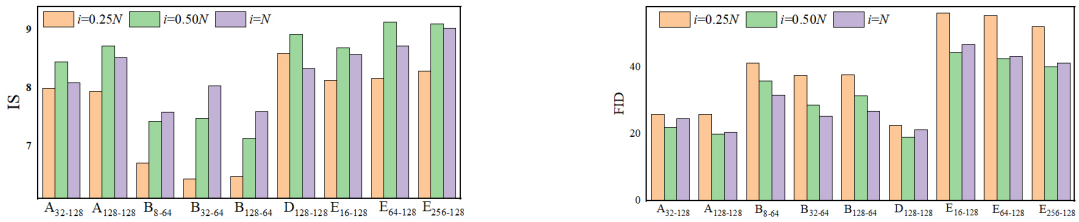
When channel size is small, mode collapse will happen to SN-GAN regardless of batch size as shown in our group C experiments. Figure 7 shows the training history of SN-GAN and SR-GAN for the setting C_{128-32}. It is seen that for SN-GAN, mode collapse has happened almost at the start of the training process and performance continues to deteriorate until eventually lead to mode collapse. In contrast, the performance of SR-GAN improves steadily and eventually converges (no mode collapse). Examples of generated images by the two training methods for this setting are also shown in the Figure. It is again clearly seen that mode collapse has indeed happened to SN-GAN while the images generated by SR-GAN are of better quality and more varieties.

In Section 2, we show that mode collapse is strongly linked to spectral collapse. By introducing spectral regularization to adjust the singular values of the weight matrices to prevent them from dropping to small values thus preventing spectral collapse, we have successfully introduced a new method for combating mode collapse. From the results presented here in this section, we have shown that regularizing the spectral distributions of the weight matrices to ensure a large number of their singular values not drop to small values can indeed prevent spectral collapse, which in turn has successfully prevented mode collapse.

4.1. The Hyperparameter i in SR-GANs

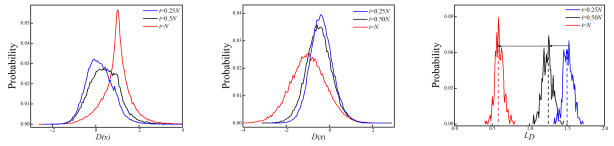
SR-GAN has a single hyperparameter i and its value will affect performances. In the experiments above, i in SR-GANs is set to $i = 0.5N$, where N is the number of singular values. Clearly, when $i = 1$, SR-GAN is the same as SN-GAN, therefore SN-GAN is a special case of SR-GAN. To investigate the effect of i, we gradually increase i, and observe its influence on model performance. In Figure 8, we show the Inception Scores and Fréchet Inception Distances for different values of i. For experiment groups A, D and E, increasing i from $0.25N$ to $0.5N$, the performances are improved. However, continuously increasing i from $0.5N$ to N, the performances deteriorate. For experiments in group B, performances increase steadily with i.

To understand why i affects performances in this way, we feed the discriminator function with the generated data and real data from both the training and testing sets, and then record the statistics of $D(x)$ in equation (2) and the



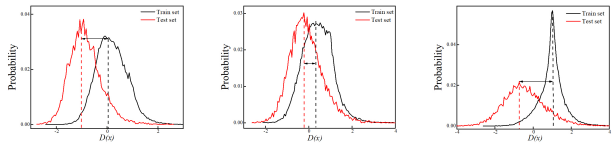
(a) Inception Score (b) Fréchet Inception Distance

Figure 8. The effect of i on model performance. N represents the number of singular values in corresponding weight matrix.



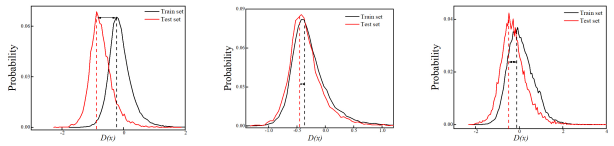
(a) $x \sim q_{train}$ (b) $x \sim q_G$ (c) L_D

Figure 9. Statistics of $D(x)$ and L_D.



(a) i=0.25N (b) i=0.50N (c) i=N

Figure 10. Statistics of $D(x)$ with setting $A_{128-128}$.



(a) i=0.25N (b) i=0.50N (c) i=N

Figure 11. Statistics of $D(x)$ with setting B_{128-64}.

discriminator objective L_D in equation (5). For explanation convenience, some typical results are illustrated here and more data can be found in the Appendix.

The probability distributions of $D(x)$ for the generated data $D(x)|_{x\sim q_G}$ and that for the training data $D(x)|_{x\sim q_{train}}$ for the setting of $A_{128-128}$ and different i values are shown in Figure 9 (a) and Figure 9 (b), respectively. Here q_{train} represents training set, and q_G represents generated set. The probability distributions of L_D is shown in Figure 9 (c).

When increasing i from $0.25N$ to N, the distributions of $D(x)|_{x\sim q_{train}}$ have a tendency of moving to the right, and at the same time the distributions of $D(x)|_{x\sim q_G}$ have a tendency of moving to the left. This means that the discriminator can better discriminate between the real and generated samples. This is also verified by the distributions of L_D as can be clearly seen in Figure 9 (c).

To investigate discriminator's performance on the testing set, we show the probability distributions of $D(x)|_{x\sim q_{train}}$ and $D(x)|_{x\sim q_{test}}$ for the setting $A_{128-128}$ in Figure 10, where q_{test} represents test set. It is seen that for $i = 0.25N$ and $i = 0.5N$, the two distributions are more similar to each other than that of $i = N$. In the case of $i = N$, the discriminator behaves significantly differently between the training data and testing data, this means that overfitting has occurred and results in a drop in performances. In summary, Figure 9 and Figure 10 explain the performance drop for setting $i = N$ in experiment groups A, D and E.

Furthermore, we monitor the statistics of $D(x)$ for the settings in group B to explain why i affects the behaviors of SR-GANs as in Figure 8. The probability distributions of $D(x)$ for the setting B_{128-64} are shown in Figure 11. We can see that for all the i values, the probability distributions of the discriminator output for the training and testing data agree well with each other, indicating no overfitting has occurred.

Although there is no systematic method for determining the best i value for different settings, our experiences is that setting $i = 0.5N$ seems to work well. In a series of extensive experiments we conducted, setting $i = 0.5N$, SR-GANs always outperform SN-GANs and very importantly, we have not yet observed mode collapse.

5. Conclusions

In this paper, we monitor spectral distributions of the discriminator's weight matrices in SN-GANs. We discover that when mode collapse occurs to a SN-GAN, a large number of its weight matrices singular values will drop to very small values, and we introduce the concept of spectral collapse to describe this phenomenon. We have provided strong evidence to link mode collapse with spectral collapse. Based on such link, we have successfully developed a spectral regularization technique for training GANs. We show that by compensating the spectral distributions of the weight matrices, we can successfully prevent spectral collapse which in turn can successfully prevent mode collapse. In a series of extensive experiments, we have successfully demonstrated that preventing spectral collapse can not only avoid mode collapse but also can improve GANs performances.

References

[1] Brock Andrew, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv*, 2018.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv*, page 1701.07875, 2017.

[3] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversar-ial networks. *arXiv preprint arXiv*, page 1703.10717, 2017.

[4] Andrew Brock, Theodore Lim, and James M. Ritchie. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv*, page 1609.07093, 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, and Mehdi Mirza. Generative adversarial nets. *Advances in neural information processing sys-tems*, pages 2672–2680, 2014.

[6] Ishaan Gulrajani, Faruk Ahmed, and Martin Arjovsky. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[7] Juha Heinonen. Lectures on lipschitz analysis. *University of Jyvaskyla*, 2005.

[8] Martin Heusel, Hubert Ramsauer, and Thomas Unterthiner. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv*, page 1706.08500, 2017.

[9] Xudong Mao, Qing Li, and Haoran Xie. Least squares generative ad-versarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017.

[10] Takeru Miyato, Toshiki Kataoka, and Masanori Koyama. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv*, page 1802.05957, 2018.

[11] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv*, page 1808.05637, 2018.

[12] Guojun Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv*, page 1701.06264, 2017.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv*, page 1511.06434, 2015.

[14] Tim Salimans, Ian Goodfellow, and Wojciech Zaremba. Improved techniques for training gans. *Advances in Neural Information Processing Sys-tems*, pages 2234–2242, 2016.

[15] Tim Salimans and Durk Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems*, pages 901–909, 2016.

[16] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Ma-chine Intelligence*, 30(11):901–909, 2008.

[17] Jiqing Wu, Zhiwu Huang, and Janine Thoma. Energy-relaxed wassertein gans (energywgan): Towards more stable and high resolution image generation. *arXiv preprint arXiv*, page 1712.01026, 2017.