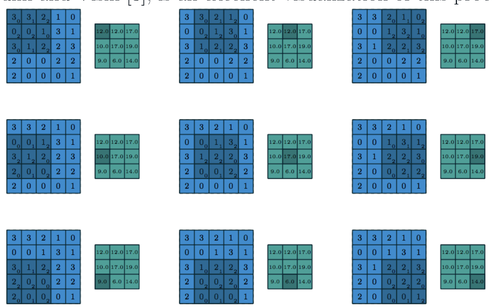
# Lipschitz Constraints in Wasserstein GANs

Kyle Sargent

Department of Computer Science, Harvard University

## Lipschitz Constant of a Convolution: Approximations and Bounds

For our purposes, convolution refers to the convolving of a 2-dimensional square kernel over a 2-dimensional input pane. In a general convolution, the kernel discretely steps over the input pane; at each step, the dot product between the current area of the input pane and the kernel is computed and added to the output. The figure below, due to Dumoulin and Visin [3], is an excellent visualization of this process.



In fact, convolution of an $n \times n$ input pane by a $k \times k$ kernel can be thought of as a linear transformation by a matrix. For example, let $K$ be a 2x2 kernel

$$K = \begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \end{bmatrix}$$

And consider a 3x3 input pane $p$

$$P = \begin{bmatrix} p_1 & p_2 & p_3 \\ p_4 & p_5 & p_6 \\ p_7 & p_8 & p_9 \end{bmatrix}$$

Convolution of $P$ by $K$ would yield a $2x2$ output pane, without padding. But we could also think of this as a map $C_K : \mathbb{R}^{3\times3} \to \mathbb{R}^{2\times2} \simeq T_K : \mathbb{R}^9 \to \mathbb{R}^4$. The matrix for this transformation would be given by the so-called "Toeplitz" matrix

$$T_K = \begin{bmatrix} k_1 & k_2 & 0 & k_3 & k_4 & 0 & 0 & 0 & 0 \\ 0 & k_1 & k_2 & 0 & k_3 & k_4 & 0 & 0 & 0 \\ 0 & 0 & k_1 & k_2 & 0 & k_3 & k_4 & 0 & 0 \\ 0 & 0 & 0 & k_1 & k_2 & 0 & k_3 & k_4 & 0 \end{bmatrix}$$

Thus convolution is a linear map, so its Lipschitz constant is just its spectral norm. We consider two methods for approximating the spectral norm of a convolution.

**Method 1:** Let $T_K$ be the block-Toeplitz matrix of a general convolution, and $R_K$ be the reshaped kernel matrix from above. Then

$$\sigma(T_K) \leq \sqrt{n}\sigma(R_k)$$

where $n$ is some constant (consult Tsuzuku, Sato, and Sugiyama [6] for details) and $R_k$ is the reshaped matrix of kernels, and the spectral norm $\sigma$ of it is computed via power iteration.

**Method 2 (Toeplitz Normalization):** For a given convolution kernel, the transpose of the implicit Toeplitz matrix is given by convolution via the original kernel, rotated 180 degrees. We can thus perform a power iteration of the form

$$(1)\ v_i \leftarrow \frac{C_{\tilde{k}}^T(u_{i-1})}{\|C_{\tilde{k}}^T(u_{i-1})\|_2}$$

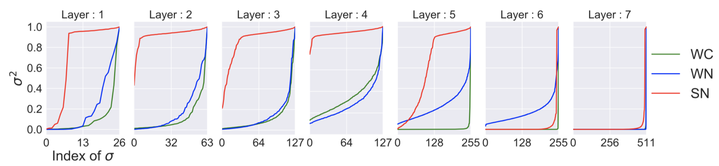
$$(2)\ u_i \leftarrow \frac{C_k^T(v_i)}{\|C_k(v_i)\|_2}$$

## Mode Collapse and Weight Spectra

Some methods for enforcing a Lipschitz constraint can be too restrictive. Consider the following normalizations

1. Weight clipping (WC) to a fixed interval $[-c, c]$ as in Arjovsky, Chintala, and Bottou [1]
2. The weight normalization (WN) of Salimans and Kingma [5], setting the $l_2$ norm of the row of each weight matrix to 1
3. Spectral normalization (SN)

We make the following observation: all three normalizations are sufficient to enforce Lipschitz continuity of the discriminator. But not all normalizations which enforce Lipschitz continuity are created equally; some introduce undesirable side effects such as spectral collapse. The following figure is due to Miyato et al. [4]:




## Spectral Gradient Clipping

Brock, Donahue, and Simonyan [2] experimented with spectral clamping: setting $\sigma_0 = \min(\sigma_0, r \cdot \sigma_1)$ where $r > 1$ is some fixed ratio. Spectral gradient clipping follows essentially the same principle as spectral clipping, just applied to weight gradients.

$$(1)\ \sigma_0 \leftarrow u_0^T W' v_0$$

$$(2)\ v_1 \leftarrow \frac{(W' - \sigma_0 uv^T)u_1}{\|(W' - \sigma_0 uv^T)u_1\|_2}$$

$$(3)\ u_1 \leftarrow \frac{(W' - \sigma_0 uv^T)^T v_1}{\|(W' - \sigma_0 uv^T)^T v_1\|_2}$$

$$(4)\ \sigma_1 \leftarrow u_1^T (W' - \sigma_0 uv^T) v_1$$

After each backward pass, we simply clamp the gradients so that $\sigma_0$ is at most $r \cdot \sigma_1$, where $r$ is some fixed hyperparameter greater than 1.

$$W' \leftarrow W' - \min(0, \sigma_0 - r \cdot \sigma_1) u_0 v_0^T$$

## Results: Toeplitz-Normalized Discriminator

We identify a trade-off between model capacity and $K-$Lipschitz continuity of the discriminator.

| Results (Toeplitz-normalized Discriminator) | |
|---|---|
| $k$ | Inception Score |
| 1 | 2.62 |
| 2 | 7.34 |
| 4 | 5.77 |
| 8 | 3.05 |

## Results: Spectral Gradient Clipping

The values of $r$ which collapsed training were vastly different depending on whether we applied clipping in the generator or discriminator. In early training, we found the discriminator to regularly update the gradients of its weights with extremely low rank updates — with $\sigma_0/\sigma_1$ on the order of 30. Later, $\sigma_0/\sigma_1$ stabilized to around 3 for most updates. Thus, for small values of $r$, spectral clipping in the discriminator proved catastrophic. By contrast, the generator is considerably more amenable to spectral clamping.

| Results (Spectral gradient clamping with ratio $r$ in Generator) | |
|---|---|
| $r$ | Inception Score |
| 1.05 | 3.782 |
| 1.5 | 7.82 |
| 2 | 7.94 |

| Results (Spectral gradient clamping with ratio $r$ in Discriminator) | |
|---|---|
| $r$ | Inception Score |
| 4 | 3.84 |
| 15 | 7.92 |
| 30 | 7.95 |

## Acknowledgements

Thanks to Professor Alexander Rush for advising the thesis on which this poster is based.

## References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN". In: (2017). URL: https://arxiv.org/pdf/1701.07875.pdf.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *CoRR* abs/1809.11096 (2018). arXiv: 1809.11096. URL: http://arxiv.org/abs/1809.11096.

[3] Vincent Dumoulin and Francesco Visin. "A guide to convolution arithmetic for deep learning". In: (2016).

[4] Takeru Miyato et al. "Spectral Normalization for Generative Adversarial Networks". In: *CoRR* abs/1802.05957 (2018). arXiv: 1802.05957. URL: http://arxiv.org/abs/1802.05957.

[5] Tim Salimans and Diederik P. Kingma. "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *CoRR* abs/1602.07868 (2016). arXiv: 1602.07868. URL: http://arxiv.org/abs/1602.07868.

[6] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. "Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks". In: *CoRR* abs/1802.04034 (2018). arXiv: 1802.04034. URL: http://arxiv.org/abs/1802.04034.