Arjovsky et al. (2017) point out that $D_{JS}$ does not account [much] for the metric structure of the space.

\[
D_{JS}(\mu, \mu') = \min(|x|) \left( \frac{1}{\delta} \log \left( 1 + \frac{1}{2\delta} \right) - \left( 1 + \frac{1}{\delta} \right) \log \left( 1 + \frac{1}{\delta} \right) \right)
\]

Hence all $|x|$ greater than $\delta$ are seen the same.
An alternative choice is the “earth moving distance”, which intuitively is the minimum mass displacement to transform one distribution into the other.

\[
\mu = \frac{1}{4} \mathbf{1}_{[1,2]} + \frac{1}{4} \mathbf{1}_{[3,4]} + \frac{1}{2} \mathbf{1}_{[9,10]} \quad \mu' = \frac{1}{2} \mathbf{1}_{[5,7]}
\]

\[
W(\mu, \mu') = 4 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{2} = 3
\]

This distance is also known as the **Wasserstein distance**, defined as

\[
W(\mu, \mu') = \min_{q \in \Pi(\mu, \mu')} \mathbb{E}_{(X, X') \sim q} \left[ \|X - X'\| \right],
\]

where \(\Pi(\mu, \mu')\) is the set of distributions over \(\mathcal{X}^2\) whose marginals are \(\mu\) and \(\mu'\).
Intuitively, it increases monotonically with the distance between modes

$$\delta$$

$$\mu$$

$$\mu'$$

$$W(\mu, \mu') = \frac{1}{2} |x|$$

So it would make a lot of sense to look for a generator matching the density for this metric, that is

$$G^* = \underset{G}{\text{argmin}} \ W(\mu, \mu_G).$$

Unfortunately, the definition of \( W \) does not provide an operational way of estimating it.
A duality theorem from Kantorovich and Rubinstein implies

\[ W(\mu, \mu') = \max_{\|f\|_L \leq 1} \mathbb{E}_{X \sim \mu} [f(X)] - \mathbb{E}_{X \sim \mu'} [f(X)] \]

where

\[ \|f\|_L = \max_{x,x'} \frac{\|f(x) - f(x')\|}{\|x - x'\|} \]

is the Lipschitz seminorm.

\[ \mu = \frac{1}{4} 1_{[1,2]} + \frac{1}{4} 1_{[3,4]} + \frac{1}{2} 1_{[9,10]} \]

\[ \mu' = \frac{1}{2} 1_{[5,7]} \]

\[ W(\mu, \mu') = \left( 3 \times \frac{1}{4} + 1 \times \frac{1}{4} + 2 \times \frac{1}{2} \right) - \left( -1 \times \frac{1}{2} - 1 \times \frac{1}{2} \right) = 3 \]
Using this result, we are looking for a generator

\[ G^* = \arg\min_G W(\mu, \mu_G) \]

\[ = \arg\min_G \max_{\|D\|_L \leq 1} \left( E_{X \sim \mu} [D(X)] - E_{X \sim \mu_G} [D(X)] \right), \]

where the max is now an optimized predictor.

This is very similar to the original GAN formulation, except that the value of \( D \) is not interpreted through a log-loss, and there is a strong regularization on \( D \).

The main issue in this formulation is to optimize the network \( D \) under a constraint on its Lipschitz seminorm

\[ \|D\|_L \leq 1. \]

Arjovsky et al. achieve this by clipping \( D \)'s weights.
The two main benefits observed by Arjovsky et al. are

- A greater stability of the learning process, both in principle and in their experiments: they do not witness “mode collapse”.
- A greater interpretability of the loss, which is a better indicator of the quality of the samples.

Figure 2: Optimal discriminator and critic when learning to differentiate two Gaussians.

As we can see, the traditional GAN discriminator saturates and results in vanishing gradients. Our WGAN critic provides very clean gradients on all parts of the space.

(Arjovsky et al., 2017)
Figure 4: JS estimates for an MLP generator (upper left) and a DCGAN generator (upper right) trained with the standard GAN procedure. Both had a DCGAN discriminator. Both curves have increasing error. Samples get better for the DCGAN but the JS estimate increases or stays constant, pointing towards no significant correlation between sample quality and loss. Bottom: MLP with both generator and discriminator. The curve goes up and down regardless of sample quality. All training curves were passed through the same median filter as in Figure 3.

However, we do not claim that this is a new method to quantitatively evaluate generative models yet. The constant scaling factor that depends on the critic’s architecture means it’s hard to compare models with different critics. Even more, in practice the fact that the critic doesn’t have infinite capacity makes it hard to know just how close to the EM distance our estimate really is. This being said, we have successfully used the loss metric to validate our experiments repeatedly and without failure, and we see this as a huge improvement in training GANs which previously had no such facility.

In contrast, Figure 4 plots the evolution of the GAN estimate of the JS distance during GAN training. More precisely, during GAN training, the discriminator is trained to maximize

\[
L(D,g_{\theta}) = \mathbb{E}_{x \sim P_r}[\log D(x)] + \mathbb{E}_{x \sim P_{\theta}}[\log(1-D(x))]
\]

which is a lower bound of \(2 \text{JS}(P_r,P_{\theta}) - 2 \log 2\). In the figure, we plot the quantity \(\frac{1}{2}L(D,g_{\theta}) + \log 2\), which is a lower bound of the JS distance.

This quantity clearly correlates poorly with sample quality. Note also that the

\begin{equation}
\text{(Arjovsky et al., 2017)}
\end{equation}
However, as Arjovsky et al. wrote:

“Weight clipping is a clearly terrible way to enforce a Lipschitz constraint. If the clipping parameter is large, then it can take a long time for any weights to reach their limit, thereby making it harder to train the critic till optimality. If the clipping is small, this can easily lead to vanishing gradients when the number of layers is big, or batch normalization is not used (such as in RNNs).”

(Arjovsky et al., 2017)

In some way, the resulting Wasserstein GAN (WGAN) trades the difficulty to train $\mathbf{G}$ for the difficulty to train $\mathbf{D}$.

In practice, this weakness results in extremely long convergence time.

Gulrajani et al. (2017) proposed the improved Wasserstein GAN in which the constraint on the Lipschitz seminorm is replaced with a smooth penalty term. They state that if

$$D^* = \arg\max_{\|D\|_L \leq 1} \left( \mathbb{E}_{X \sim \mu} \left[ D(X) \right] - \mathbb{E}_{X \sim \mu_G} \left[ D(X) \right] \right)$$

then, with probability one under $\mu$ and $\mu_G$

$$\| \nabla D^*(X) \| = 1.$$

This implies that adding a regularization that pushes the gradient norm to one should not exclude [any of] the optimal discriminator[s].
So instead of looking for

$$\argmax_{\|D\|_L \leq 1} E_{X \sim \mu} [D(X)] - E_{X \sim \mu_G} [D(X)],$$

Gulrajani et al. propose to solve

$$\argmax_D E_{X \sim \mu} [D(X)] - E_{X \sim \mu_G} [D(X)] - \lambda E_{X \sim \mu_P} [(\|\nabla D(X)\| - 1)^2]$$

where $\mu_P$ is the distribution of a point $B$ sampled uniformly between a real sample $X$ and a fake sample $G(Z)$, that is $B = UX + (1 - U)X'$ where $X \sim \mu$, $X' \sim \mu_G$, and $U \sim \mathcal{U}[0, 1]$.

Note that this loss involves second-order derivatives.

Experiments show that this scheme is more stable than WGAN under many different conditions.

References
