Appendix

A. Training objective as an upper bound

Claim 1 Let $\mathbb{P}_d$ and $\mathbb{P}_f$ be two distributions. Suppose that $\hat{\mathbb{P}}_d$ and $\hat{\mathbb{P}}_f$ are empirical measures of $\mathbb{P}_d$ and $\mathbb{P}_f$, induced by random sets (of $n$ i.i.d samples) $D$ and $F$. Then

$$\hat{W}^2_2(\mathbb{P}_d, \mathbb{P}_f) \leq 16\mathbb{E}[\hat{W}^2_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)].$$  \hspace{1cm} (16)

Proof: Using the triangle inequality for the sliced Wasserstein distance, we have

$$\hat{W}^2_2(\mathbb{P}_d, \mathbb{P}_f) \leq 2\hat{W}^2_2(\mathbb{P}_d, \hat{\mathbb{P}}_d) + 2\hat{W}^2_2(\mathbb{P}_f, \hat{\mathbb{P}}_f).$$  \hspace{1cm} (17)

Using it again, we get

$$\hat{W}^2_2(\mathbb{P}_d, \mathbb{P}_f) \leq 2\hat{W}^2_2(\mathbb{P}_d, \hat{\mathbb{P}}_d) + 4\hat{W}^2_2(\mathbb{P}_f, \hat{\mathbb{P}}_f) + 4\hat{W}^2_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f).$$  \hspace{1cm} (18)

In the following we find upper bounds for $\hat{W}^2_2(\mathbb{P}_f, \hat{\mathbb{P}}_f)$ in terms of $\hat{W}^2_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)$. In order to do this, we must deconstruct the sliced Wasserstein distance. By definition, we have

$$\hat{W}^2_2(\mathbb{P}_f, \hat{\mathbb{P}}_f) = \int_{\omega \in \Omega} W^2_2(\mathbb{P}^\omega, \hat{\mathbb{P}}^\omega_\omega) \, d\omega.$$  \hspace{1cm} (19)

Consider any one projection $\omega$. We have a 1-d distribution $\mathbb{P}^\omega$, and its empirical measure $\hat{\mathbb{P}}^\omega$. Using Theorem 4.3 in [5]:

$$\mathbb{E}[W^2_2(\mathbb{P}^\omega, \hat{\mathbb{P}}^\omega_\omega)] \leq \mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)],$$  \hspace{1cm} (20)

where $\hat{\mathbb{P}}^\omega$ is an independent copy of $\hat{\mathbb{P}}^\omega$.

To bound $\mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)]$ in Eq. (20), we first see how the expected Wasserstein distance between two 1-d empirical measures $\hat{\mathbb{P}}^\omega$ and $\hat{\mathbb{P}}^\omega$ can be written in terms of the sets of samples $D^\omega$ and $F^\omega$ that they represent (i.e. are induced by). Note that $D^\omega$ and $F^\omega$ are obtained by simply projecting a the sets $D$ and $F$ onto the direction $\omega$. If $D^\omega_{\sigma_D(i)}$ and $F^\omega_{\sigma_F(i)}$ denote the $i$-th smallest sample in $D^\omega$ and $F^\omega$,

$$\mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(D^\omega_{\sigma_D(i)} - F^\omega_{\sigma_F(i)})^2].$$  \hspace{1cm} (21)

$D^\omega_{\sigma_D(i)}$ and $F^\omega_{\sigma_F(i)}$ are in fact the sample order statistics of $\mathbb{P}^\omega$ and $\mathbb{P}^\omega$. For $\hat{\mathbb{P}}^\omega$ and $\hat{\mathbb{P}}^\omega$, we can write this as

$$\mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)] = \frac{2}{n} \sum_{i=1}^{n} \text{Var}[\mathbb{P}^\omega_{\sigma_F(i)}].$$  \hspace{1cm} (22)

The RHS of Eq. (21) can be decomposed as

$$\mathbb{E}[(D^\omega_{\sigma_D(i)} - F^\omega_{\sigma_F(i)})^2] = \mathbb{E}[(D^\omega_{\sigma_D(i)} - \mathbb{E}[F^\omega_{\sigma_F(i)}] + \mathbb{E}[F^\omega_{\sigma_F(i)}] - F^\omega_{\sigma_F(i)})^2]$$

$$\geq \text{Var}[F^\omega_{\sigma_F(i)}],$$

hence

$$\frac{1}{n} \sum_{i=1}^{n} \text{Var}[\mathbb{P}^\omega_{\sigma_F(i)}] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(D^\omega_{\sigma_D(i)} - F^\omega_{\sigma_F(i)})^2].$$

Combining this result with Eq. (21) and Eq. (22) yields

$$\mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)] \leq 2\mathbb{E}[W^2_2(\hat{\mathbb{P}}^\omega, \hat{\mathbb{P}}^\omega_\omega)].$$
which, when combined with Eq. (20), results in

\[ E[W_2^2(P_f, \hat{P}_f)] \leq 2E[W_2^2(\hat{P}_d, \hat{P}_f)]. \] (23)

Applying the expectation operator on Eq. (19) and using Eq. (23),

\[ E[\hat{W}_2^2(P_f, \hat{P}_f)] \leq 2 \int_{\omega \in \Omega} E[W_2^2(\hat{P}_d^\omega, \hat{P}_f^\omega)] d\omega = 2E[\hat{W}_2^2(\hat{P}_d, \hat{P}_f)]. \] (24)

The same bound holds for \( E[\hat{W}_2^2(P_d, \hat{P}_d)]. \)

Substituting from Eq. (24) in Eq. (18) and applying the expectation operator, we get

\[ \hat{W}_2^2(P_d, P_f) \leq 16E[\hat{W}_2(\hat{P}_d, \hat{P}_f)], \] (25)

which completes the proof.

B. Bounds for generated distribution

**Corollary 1** Let \( P_d \) and \( P_f \) be two distributions. Suppose that \( \hat{P}_d \) and \( \hat{P}_f \) are \((n\text{-sample})\) empirical measures of \( P_d \) and \( P_f \), and let \( \hat{P}_f^\omega \) be an independent copy of \( \hat{P}_d \). For \( P_f^* \) defined by \( P_f^* = \arg\min_{P_f} E[\hat{W}_2^2(P_d, P_f)] \), the following holds:

\[ \hat{W}_2(P_d, P_f^*) \leq 14E[\hat{W}_2(\hat{P}_d, \hat{P}_d^\omega)]. \] (26)

**Proof:** This follows easily from Claim 1. Using Eq. (20), we can show that

\[ E[\hat{W}_2^2(P_f, \hat{P}_d)] \leq E[\hat{W}_2^2(P_f, \hat{P}_d)], \] (27)

and therefore we can rewrite (18) as:

\[ \hat{W}_2(P_d, P_f) \leq 2E[\hat{W}_2^2(\hat{P}_d, \hat{P}_d^\omega)] + 12E[\hat{W}_2(\hat{P}_d, \hat{P}_f)]. \] (28)

Since \( P_f^* \) minimizes \( E[\hat{W}_2^2(\hat{P}_d, \hat{P}_f)] \) over all \( P_f \),

\[ E[\hat{W}_2(\hat{P}_d, \hat{P}_f^\omega)] \leq E[\hat{W}_2(\hat{P}_d, \hat{P}_d^\omega)]. \] (29)

Therefore,

\[ \hat{W}_2(P_d, P_f^*) \leq 14E[\hat{W}_2(\hat{P}_d, \hat{P}_d^\omega)]. \] (30)

C. Discriminator update frequency experiments

We tested different discriminator update schemes (i.e., number of generator updates per discriminator updates, and number of iterations of discriminator updates). In Tab. 4 we show samples after 40 epochs of training on the LSUN dataset with these different schemes for two discriminator configurations. The generator architecture for both is the DCGAN.
The SWG is robust to different discriminator update schemes. Tested for two discriminator architectures (columns). Sample size = 64, learning rate = 0.0005, Adam optimizer, 40 epochs.
D. Network architectures for experiments on MNIST

Here we summarize the different network architectures used for experiments with the MNIST dataset presented in Sec. 4.2.

<table>
<thead>
<tr>
<th>Generator (Fully Connected)</th>
<th>Generator (Conv &amp; Deconv)</th>
<th>Discriminator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>output</strong>: 784-d sample</td>
<td><strong>output</strong>: 784-d sample</td>
<td><strong>output</strong>: scalar</td>
</tr>
<tr>
<td>fc-784, sigmoid</td>
<td>conv2d-1-3-1, sigmoid</td>
<td>2× fc-256, relu</td>
</tr>
<tr>
<td>7× fc-512, relu</td>
<td>deconv2d-16-3-2, (bn), relu</td>
<td></td>
</tr>
<tr>
<td><strong>input</strong>: 32-d random noise</td>
<td>conv2d-32-3-1, (bn), relu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>deconv2d-32-3-2, (bn), relu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv2d-64-3-1, (bn), relu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>deconv2d-64-3-2, (bn), relu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>fc-1024</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Generator and discriminator for MNIST. “fc-\(n\)” means applying a fully connected layer with \(n\) output units. Both “conv2d-\(c\)-\(k\)-\(s\)” and “deconv2d-\(c\)-\(k\)-\(s\)” mean applying \(c\) convolutional filters of size \(k\) by \(k\) with stride \(s\) by \(s\). “bn” means batch normalization.