Appendix

A. Proof of Proposition 3

Since $g_{\theta\sharp}S$ and T are normal currents we know $\mathbb{F}_\lambda(g_{\theta\sharp}S, T) < \infty$ for all $\theta \in \Theta$.

We now directly show Lipschitz continuity. First notice that

$$\mathbb{F}_\lambda(g_{\theta\sharp}S - T) = \mathbb{F}_\lambda(g_{\theta\sharp}S + g_{\theta'\sharp}S - g_{\theta'\sharp}S - T) \quad (36)$$
$$\leq \mathbb{F}_\lambda(g_{\theta\sharp}S - g_{\theta'\sharp}S) + \mathbb{F}_\lambda(g_{\theta'\sharp}S - T), \quad (37)$$

yields the following bound:

$$|\mathbb{F}_\lambda(g_{\theta\sharp}S - T) - \mathbb{F}_\lambda(g_{\theta'\sharp}S - T)| \leq \mathbb{F}_\lambda(g_{\theta\sharp}S - g_{\theta'\sharp}S). \quad (38)$$

Due to Prop. 1 we have that

$$\mathbb{F}_\lambda(g_{\theta\sharp}S - g_{\theta'\sharp}S) \leq \max\{1, \lambda\} \cdot \mathbb{F}(g_{\theta\sharp}S - g_{\theta'\sharp}S). \quad (39)$$

Now define the compact set $C \subset \mathbf{R}^d$ as

$$C = \big\{(1-t)g_\theta(z) + t g_{\theta'}(z) : \ z \in \operatorname{spt} S, \ 0 \leq t \leq 1\big\}, \quad (40)$$

and as in §4.1.12 in Federer (1969) for compact $K \subset \mathbf{R}^d$ the "stronger" flat norm

$$\mathbb{F}_K(T) = \sup\big\{T(\omega) \mid \omega \in \mathcal{D}^k(\mathbf{R}^d), \text{ with } \|\omega(x)\|^* \leq 1, \|d\omega(x)\|^* \leq 1 \text{ for all } x \in K\big\}. \quad (41)$$

Since the constraint in the supremum in (41) is less restrictive than in the definition of the flat norm (20), we have

$$\mathbb{F}(g_{\theta\sharp}S - g_{\theta'\sharp}S) \leq \mathbb{F}_C(g_{\theta\sharp}S - g_{\theta'\sharp}S). \quad (42)$$

Then, the inequality after §4.1.13 in Federer (1969) bounds the right side of (42) for $k > 0$ by

$$\mathbb{F}_C(g_{\theta\sharp}S - g_{\theta'\sharp}S) \leq \|S\|(|g_\theta - g_{\theta'}|\rho^k) + \|\partial S\|(|g_\theta - g_{\theta'}|\rho^{k-1}), \quad (43)$$

where $\rho(z) = \max\{\|\nabla_z g(z,\theta)\|, \|\nabla_z g(z,\theta')\|\} < \infty$ due to Assumption 1 and we write $\|S\|(f) = \int f(z)\,\mathrm{d}\|S\|(z)$, where $\|S\|$ is defined in the sense of (19). For $k = 0$, a similar bound can be derived without the term $\|\partial S\|$.

For $k > 0$, by setting $\mu_S = \|\partial S\| + \|S\|$ we can further bound the term in (43) by

$$\|S\|(|g_\theta - g_{\theta'}|\rho^k) + \|\partial S\|(|g_\theta - g_{\theta'}|\rho^{k-1}) \leq c_1 \cdot \int \|g_\theta(z) - g_{\theta'}(z)\| \mathrm{d}\mu_S(z), \quad (44)$$

where $c_1 = \sup_z \max\{\rho^k(z), \rho^{k-1}(z)\}$. For $k = 0$, the bound is derived analogously.

Now since $g(z,\cdot)$ is locally Lipschitz and $\Theta \subset \mathbf{R}^n$ is compact, $g(z,\cdot)$ is Lipschitz and we denote the constant as $\mathrm{Lip}(g)$, leading to the bound

$$\int \|g_\theta(z) - g_{\theta'}(z)\| \mathrm{d}\mu_S(z) \leq \mu_S(\mathcal{Z}) \,\mathrm{Lip}(g) \cdot \|\theta - \theta'\|. \quad (45)$$

Since $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ is a normal current, $\mu_S(\mathcal{Z}) < \infty$. Thus by combining (38), (39), (42), (43), (44) and (45) there is a finite $c_2 = \max\{1, \lambda\} \cdot c_1 \cdot \mu_S(\mathcal{Z}) \cdot \mathrm{Lip}(g) < \infty$ such that

$$|\mathbb{F}_\lambda(g_{\theta\sharp}S - T) - \mathbb{F}_\lambda(g_{\theta'\sharp}S - T)| \leq c_2\|\theta - \theta'\|. \quad (46)$$

Therefore, the cost $\mathbb{F}_\lambda(g_{\theta\sharp}S, T)$ in (27) is Lipschitz in θ and by Rademacher's theorem, §3.1.6 in Federer (1969), also differentiable almost everywhere.

B. Parameters and Network Architectures

For all experiments we use Adam optimizer (Kingma & Ba, 2014), with step size 10^{-4} and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.9$. The batch size is set to 50 in all experiments except the first one (which runs full batch with batch size 5). We always set $\lambda = 1$.

B.1. Illustrative 2D Example

We pick the same parameters for $k \in \{0, 1\}$. We set the penalty to $\rho = 10$ and use 5 discriminator updates per generator update as in (Gulrajani et al., 2017). The generator is a $5 - 6 - 250 - 250 - 250 - 2$ fully connected network with leaky ReLU activations. The first layer ensures that the latent coordinate z_1 has the topology of a circle, i.e., it is implemented as $(\cos(z_1), \sin(z_1), z_2, z_3, z_4, z_5)$. The discriminators ω^0 and $\omega^{1,1}$ are $2 - 100 - 100 - 100 - 1$ respectively $2 - 100 - 100 - 2$ nets with leaky ReLUs. The distribution on the latent is a uniform $z_1 \sim U([-\pi, \pi])$ and $z_i \sim \mathcal{N}(0, 1)$ for the remaining 4 latent codes.

B.2. MNIST

For the remaining experiments, we use only 1 discriminator update per iteration. The digits are resized to 32×32. For generator we use DCGAN architecture (Radford et al., 2015) without batch norm and with ELU activations, see Table 1. The discriminators are given by the architectures in Table 2, with leaky ReLUs between the layers.

Before computing $\langle \omega^{1,1}(x) \wedge \omega^{1,2}(x), v_1 \wedge v_2 \rangle$, the tangent images $v_1, v_2 \in \mathbf{R}^{32\cdot 32}$ are convolved with a Gaussian with a standard deviation of 2 and downsampled to 8×8 using average pooling. The distributions on the latent space are given by $z_1 \sim U([-7.5, 7.5])$, $z_2 \sim U([-0.5, 0.5])$ and $z_i \sim \mathcal{N}(0, 1)$ for the remaining 126 latent variables. The tangent vectors at each sample are computed by a 2 degree rotation and a dilation with radius one.

layer name	output size	filters
Reshape	$128 \times 1 \times 1$	–
Conv2DTranspose	$32F \times 4 \times 4$	$128 \to 32F$
Conv2DTranspose	$16F \times 8 \times 8$	$32F \to 16F$
Conv2DTranspose	$4F \times 16 \times 16$	$16F \to 4F$
Conv2DTranspose	$1 \times 32 \times 32$	$4F \to 1$

Table 1. Generator architecture for MNIST experiment, $F = 32$.

layer name	output size	filters
Reshape	$1 \times 32 \times 32$	–
Conv2D	$2F \times 16 \times 16$	$1 \to 2F$
Conv2D	$4F \times 8 \times 8$	$2F \to 4F$
Conv2D	$32F \times 4 \times 4$	$4F \to 32F$
Conv2D	$1 \times 1 \times 1$	$32F \to 1$
Conv2DTranspose	$1 \times 8 \times 8$	$32F \to 1$

Table 2. The discriminator ω^0 has $F = 32$ and red last layer. The discriminators $\omega^{1,1}$, $\omega^{1,2}$ have $F = 8$ and last layer in blue.

B.3. SmallNORB

We downsample the smallNORB images to 48×48. The architectures and parameters are chosen similar to the previous MNIST example, see Table 3 and Table 4.

layer name	output size	filters
Reshape	$128 \times 1 \times 1$	–
Conv2DTranspose	$32F \times 4 \times 4$	$128 \to 32F$
Conv2DTranspose	$16F \times 8 \times 8$	$32F \to 16F$
Conv2DTranspose	$16F \times 12 \times 12$	$16F \to 16F$
Conv2DTranspose	$4F \times 24 \times 24$	$16F \to 4F$
Conv2DTranspose	$1 \times 48 \times 48$	$4F \to 1$

Table 3. Generator for smallNORB experiment, $F = 24$.

layer name	output size	filters
Reshape	$1 \times 48 \times 48$	–
Conv2D	$2F \times 24 \times 24$	$1 \to 2F$
Conv2D	$4F \times 12 \times 12$	$2F \to 4F$
Conv2D	$32F \times 6 \times 6$	$4F \to 32F$
Conv2D	$1 \times 1 \times 1$	$32F \to 1$
Conv2DTranspose	$1 \times 12 \times 12$	$32F \to 1$

Table 4. SmallNORB discriminator ω^0, $F = 32$, last layer in shown in red, and tangent discriminators $\omega^{1,1}$, $\omega^{1,2}$, $\omega^{1,3}$ where $F = 8$ and last layer is highlighted in blue.

B.4. Tinyvideos

The architectures for the tinyvideo experiment are borrowed from the recent work Mescheder et al. (2018), see Table 5 and Table 6.

layer name	output size	filters
Fully Connected	8192	–
Reshape	$512 \times 4 \times 4$	–
ResNet-Block	$512 \times 4 \times 4$	$512 \to 512 \to 512$
NN-Upsampling	$512 \times 8 \times 8$	–
ResNet-Block	$256 \times 8 \times 8$	$512 \to 256 \to 256$
NN-Upsampling	$256 \times 16 \times 16$	–
ResNet-Block	$128 \times 16 \times 16$	$256 \to 128 \to 128$
NN-Upsampling	$128 \times 32 \times 32$	–
ResNet-Block	$64 \times 32 \times 32$	$128 \to 64 \to 64$
NN-Upsampling	$64 \times 64 \times 64$	–
ResNet-Block	$64 \times 64 \times 64$	$64 \to 64 \to 64$
Conv2D	$3 \times 64 \times 64$	$64 \to 3$

Table 5. Generator architecture for tinyvideos experiment.

layer name	output size	filters
Conv2D	$64 \times 64 \times 64$	$3 \to 64$
ResNet-Block	$64 \times 64 \times 64$	$64 \to 64 \to 64$
AvgPool2D	$64 \times 32 \times 32$	–
ResNet-Block	$128 \times 32 \times 32$	$64 \to 64 \to 128$
AvgPool2D	$128 \times 16 \times 16$	–
ResNet-Block	$256 \times 16 \times 16$	$128 \to 128 \to 256$
AvgPool2D	$256 \times 8 \times 8$	–
ResNet-Block	$512 \times 8 \times 8$	$256 \to 256 \to 512$
AvgPool2D	$512 \times 4 \times 4$	–
ResNet-Block	$1024 \times 4 \times 4$	$512 \to 512 \to 1024$
Conv2D	$1 \times 1 \times 1$	$1024 \to 1$
ResNet-Block	$256 \times 16 \times 16$	$128 \to 256 \to 256$
Conv2D	$3 \times 16 \times 16$	$256 \to 3$

Table 6. Discriminator architectures for tinyvideos experiment. Last layers of ω^0 are highlighted in red, and the last layers of the temporal discriminator $\omega^{1,1}$ are highlighted in blue.