Flat Metric Minimization with Applications in Generative Modeling

**Thomas Möllenhoff** [1]   **Daniel Cremers** [1]

Abstract

We take the novel perspective to view data not as a probability distribution but rather as a current. Primarily studied in the field of geometric measure theory, k-currents are continuous linear functionals acting on compactly supported smooth differential forms and can be understood as a generalized notion of oriented k-dimensional manifold. By moving from distributions (which are 0-currents) to k-currents, we can explicitly orient the data by attaching a k-dimensional tangent plane to each sample point. Based on the flat metric which is a fundamental distance between currents, we derive FlatGAN, a formulation in the spirit of generative adversarial networks but generalized to k-currents. In our theoretical contribution we prove that the flat metric between a parametrized current and a reference current is Lipschitz continuous in the parameters. In experiments, we show that the proposed shift to $k > 0$ leads to interpretable and disentangled latent representations which behave equivariantly to the specified oriented tangent planes.

1. Introduction

This work is concerned with the problem of representation learning, which has important consequences for many tasks in artificial intelligence, cf. the work of Bengio et al. (2013). More specifically, our aim is to learn representations which behave equivariantly with respect to selected transformations of the data. Such variations are often known beforehand and could for example describe changes in stroke width or rotation of a digit, changes in viewpoint or lighting in a three-dimensional scene but also the *arrow of time* (Pickup et al., 2014; Wei et al., 2018) in time-series, describing how a video changes from one frame to the next, see Fig. 1.

[1]Department of Informatics, Technical University of Munich, Garching, Germany. Correspondence to: Thomas Möllenhoff <thomas.moellenhoff@tum.de>.
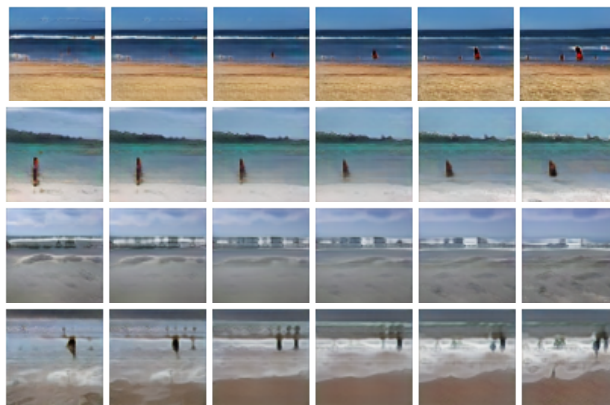
from left to right we vary the latent code z_1 (time)

Figure 1. Discovering the arrow of time by training a generative model with the proposed formalism on the tinyvideos dataset (Vondrick et al., 2016). The approach we introduce allows one to learn latent representations which behave equivariantly to specified tangent vectors (here: difference of two successive video frames).

We tackle this problem by introducing a novel formalism based on *geometric measure theory* (Federer, 1969), which we find to be interesting in itself. To motivate our application in generative modeling, recall the manifold hypothesis which states that the distribution of real-world data tends to concentrate nearby a low-dimensional manifold, see Fefferman et al. (2016) and the references therein. Under that hypothesis, a possible unifying view on prominent methods in unsupervised and representation learning such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational auto-encoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) is the following: both approaches aim to approximate the true distribution concentrating near the manifold with a distribution on some low-dimensional latent space $\mathcal{Z} \subset \mathbf{R}^l$ that is pushed through a decoder or generator $g : \mathcal{Z} \to \mathcal{X}$ mapping to the (high-dimensional) data space $\mathcal{X} \subset \mathbf{R}^d$ (Genevay et al., 2017; Bottou et al., 2017).

We argue that treating data as a distribution potentially ignores useful available geometric information such as orientation and tangent vectors to the data manifold. Such tangent vectors describe the aforementioned local variations or pertubations. Therefore we postulate that *data should not be viewed as a distribution but rather as a k-current.*
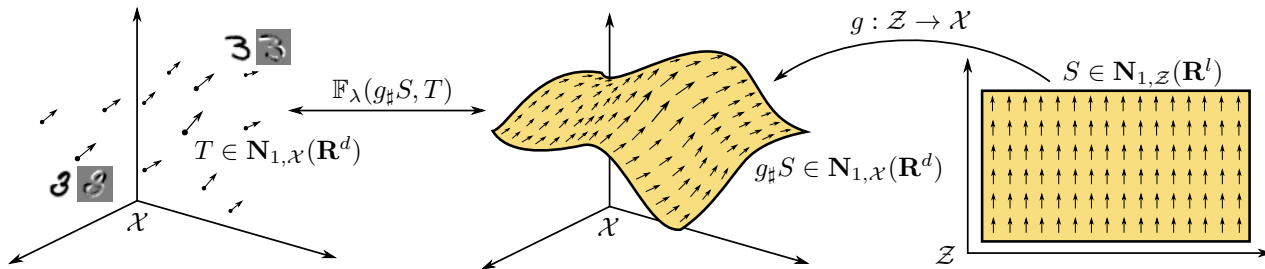
Figure 2. **Illustration of the proposed idea.** We suggest the novel perspective to view observed data (here the MNIST dataset) as a k-current T, shown as the dots with attached arrows on the left. The arrows indicate the oriented tangent space, and we selected $k = 1$ to be rotational deformation. We propose to minimize the flat distance of T to the *pushforward* $g_\sharp S$ (shown in the middle) of a current S on a low-dimensional latent space $\mathcal{Z}$ (right) with respect to a "generator" map $g : \mathcal{Z} \to \mathcal{X}$. For 0-currents (no selected tangent vectors) and sufficiently large λ, the proposed "FlatGAN" formulation specializes to the Wasserstein GAN (Arjovsky et al., 2017).

We postpone the definition of k-currents (de Rham, 1955) to Sec. 3, and informally think of them as distributions over k-dimensional oriented planes. For the limiting case $k = 0$, currents simply reduce to distributions in the sense of Schwartz (1951, 1957) and positive 0-currents with unit mass are probability measures. A seminal work in the theory of currents was written by Federer & Fleming (1960), which established compactness theorems for subsets of currents (*normal* and *integral currents*). In this paper, we will work in the space of normal k-currents with compact support in $\mathcal{X} \subset \mathbf{R}^d$, denoted by $\mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.

Similarly as probabilistic models build upon f-divergences (Csiszár et al., 2004), integral probability metrics (Sriperumbudur et al., 2012) or more general optimal transportation related divergences (Peyré & Cuturi, 2018; Feydy et al., 2018), we require a sensible notion to measure "distance" between k-currents.

In this work, we will focus on the flat norm[1] due to Whitney (1957). To be precise, we consider a scaled variant introduced and studied by Morgan & Vixie (2007); Vixie et al. (2010). This choice is motivated in Sec. 4, where we show that the flat norm enjoys certain attractive properties similar to the celebrated Wasserstein distances. For example, it metrizes the weak*-convergence for normal currents.

A potential alternative to the flat norm are kernel metrics on spaces of currents (Vaillant & Glaunès, 2005; Glaunès et al., 2008). These have been proposed for diffeomorphic registration, but kernel distances on distributions have also been sucessfully employed for generative modeling, see Li et al. (2017). Constructions similar to the Kantorovich relaxation in optimal transport but generalized to k-currents recently appeared in the context of convexifications for certain variational problems (Möllenhoff & Cremers, 2019).

---

[1]The terminology "flat" carries no geometrical significance and refers to Whitney's use of musical notation flat $|\cdot|^\flat$ and sharp $|\cdot|^\sharp$.

2. Related Work

Our main idea is illustrated in Fig. 2, which was inspired from the optimal transportation point of view on GANs given by Genevay et al. (2017).

Tangent vectors of the data manifold, either prespecified (Simard et al., 1992; 1998; Fraser et al., 2003) or learned with a contractive autoencoder (Rifai et al., 2011), have been used to train classifiers that aim to be *invariant* to changes relative to the data manifold. In contrast to these works, we use tangent vectors to learn interpretable representations and a generative model that aims to be *equivariant*. The principled introduction of tangent k-vectors into probabilistic generative models is one of our main contributions.

Various approaches to learning informative or disentangled latent representations in a completely unsupervised fashion exist (Schmidhuber, 1992; Higgins et al., 2016; Chen et al., 2016; Kim & Mnih, 2018). Our approach is orthogonal to these works, as specifying tangent vectors further encourages informative representations to be learned. For example, our GAN formulation could be combined with a mutual information term as in InfoGAN (Chen et al., 2016).

Our work is more closely related to semi-supervised approaches on learning disentangled latent representations, which similarly also require some form of knowledge of the underlying factors (Hinton et al., 2011; Denton et al., 2017; Mathieu et al., 2016; Narayanaswamy et al., 2017) and also to conditional GANs (Mirza & Osindero, 2014; Odena et al., 2017). However, the difference is the connection to geometric measure theory which we believe to be completely novel, and our specific FlatGAN formulation that seamlessly extends the Wasserstein GAN (Arjovsky et al., 2017), cf. Fig. 2.

Since the concepts we need from geometric measure theory are not commonly used in machine learning, we briefly review them in the following section.

3. Geometric Measure Theory

The book by Federer (1969) is still the formidable, definitive reference on the subject. As a more accessible introduction we recommend (Krantz & Parks, 2008) or (Morgan, 2016). While our aim is to keep the manuscript self-contained, we invite the interested reader to consult Chapter 4 in (Morgan, 2016), which in turn refers to the corresponding chapters in the book of Federer (1969) for more details.

3.1. Grassmann Algebra

Notation. Denote $\{e_1, \ldots, e_d\}$ a basis of $\mathbf{R}^d$ with dual basis $\{dx_1, \ldots, dx_d\}$ such that $dx_i : \mathbf{R}^d \to \mathbf{R}$ is the linear functional that maps every $x = (x_1, \ldots, x_d)$ to the i-th component x_i. For $k \leq d$, denote $I(d, k)$ as the ordered multi-indices $\mathbf{i} = (i_1, \ldots, i_k)$ with $1 \leq i_1 < \ldots < i_k \leq d$.

One can multiply vectors in $\mathbf{R}^d$ to obtain a new object:

$$\xi = v_1 \wedge \ldots \wedge v_k, \tag{1}$$

called a k-vector ξ in $\mathbf{R}^d$. The wedge (or exterior) product $\wedge$ is characterized by multilinearity

$$cv_1 \wedge v_2 = v_1 \wedge cv_2 = c(v_1 \wedge v_2), \quad \text{for } c \in \mathbf{R},$$
$$(u_1 + v_1) \wedge (u_2 + v_2) = \tag{2}$$
$$u_1 \wedge u_2 + u_1 \wedge v_2 + v_1 \wedge u_2 + v_1 \wedge v_2,$$

and it is alternating

$$u \wedge v = -v \wedge u, \quad u \wedge u = 0. \tag{3}$$

In general, any k-vector can be written as

$$\xi = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{i_1} \wedge \ldots \wedge e_{i_k} = \sum_{\mathbf{i} \in I(d,k)} a_{\mathbf{i}} \cdot e_{\mathbf{i}}, \tag{4}$$

for coefficients $a_{\mathbf{i}} \in \mathbf{R}$. The vector space of k-vectors is denoted by $\mathbf{\Lambda}_k \mathbf{R}^d$ and has dimension $\binom{d}{k}$. We define for two k-vectors $v = \sum_{\mathbf{i}} a_{\mathbf{i}} e_{\mathbf{i}}$, $w = \sum_{\mathbf{i}} b_{\mathbf{i}} e_{\mathbf{i}}$ an inner product $\langle v, w \rangle = \sum_{\mathbf{i}} a_{\mathbf{i}} b_{\mathbf{i}}$ and the Euclidean norm $|v| = \sqrt{\langle v, v \rangle}$.

A simple (or decomposable) k-vector is any $\xi \in \mathbf{\Lambda}_k \mathbf{R}^d$ that can be written using products of 1-vectors. Simple k-vectors such as (1) are uniquely determined by the k-dimensional space spanned by the $\{v_i\}$, their orientation and the norm $|v|$ corresponding to the area of the parallelotope spanned by the $\{v_i\}$. Simple k-vectors with unit norm can therefore be thought of as oriented k-dimensional subspaces and the rules (2)-(3) can be thought of as equivalence relations.

It turns out that the inner product of two simple k-vectors can be computed by the $k \times k$-determinant

$$\langle w_1 \wedge \ldots \wedge w_k, v_1 \wedge \ldots \wedge v_k \rangle = \det \left(W^\top V \right), \tag{5}$$

where the columns of $W \in \mathbf{R}^{d \times k}$, $V \in \mathbf{R}^{d \times k}$ contain the individual 1-vectors. This will be useful later for our practical implementation.

Not all k-vectors are simple. An illustrative example is $e_1 \wedge e_2 + e_3 \wedge e_4 \in \mathbf{\Lambda}_2 \mathbf{R}^4$, which describes two 2-dimensional subspaces in $\mathbf{R}^4$ intersecting only at zero.

The dual space of $\mathbf{\Lambda}_k \mathbf{R}^d$ is denoted as $\mathbf{\Lambda}^k \mathbf{R}^d$, and its elements are called k-covectors. They are similarly represented as (4) but with dual basis $dx_{\mathbf{i}}$. Analogously to the previous page, we can define an inner product between k-vectors and k-covectors. Next to the Euclidean norm $|\cdot|$, we define two additional norms due to Whitney (1957).

Definition 1 (Mass and comass). *The comass norm defined for k-covectors $w \in \mathbf{\Lambda}^k \mathbf{R}^n$ is given by*

$$\|w\|^* = \sup \left\{ \langle w, v \rangle : v \text{ is simple}, |v| = 1 \right\}, \tag{6}$$

and the mass norm for $v \in \mathbf{\Lambda}_k \mathbf{R}^n$ is given by

$$\|v\| = \sup \left\{ \langle v, w \rangle : \|w\|^* \leq 1 \right\}$$
$$= \inf \left\{ \sum_i |\xi_i| : \xi_i \text{ are simple}, v = \sum_i \xi_i \right\}. \tag{7}$$

The mass norm is by construction the largest norm that agrees with the Euclidean norm on simple k-vectors. For the non-simple 2-vector from before, we compute

$$\|e_1 \wedge e_2 + e_3 \wedge e_4\| = 2, \quad |e_1 \wedge e_2 + e_3 \wedge e_4| = \sqrt{2}. \tag{8}$$

Interpreting the non-simple vector as two tangent planes, we see that the mass norm gives the correct area, while the Euclidean norm underestimates it. The comass $\|\cdot\|^*$ will be used later to define the mass of currents and the flat norm.

3.2. Differential Forms

In order to define currents, we first need to introduce differential forms. A differential k-form is a k-covectorfield $\omega : \mathbf{R}^d \to \mathbf{\Lambda}^k \mathbf{R}^d$. The support $\text{spt}\, \omega$ is defined as the closure of the set $\{x \in \mathbf{R}^d : \omega(x) \neq 0\}$.

Differential forms allow one to perform coordinate-free integration over oriented manifolds. Given some manifold $\mathcal{M} \subset \mathbf{R}^d$, possibly with boundary, an *orientation* is a continuous map $\tau_{\mathcal{M}} : \mathcal{M} \to \mathbf{\Lambda}_k \mathbf{R}^d$ which assigns to each point a simple k-vector with unit norm that spans the tangent space at that point. Integration of a differential form over an oriented manifold $\mathcal{M}$ is then defined by:

$$\int_{\mathcal{M}} \omega = \int_{\mathcal{M}} \langle \omega(x), \tau_{\mathcal{M}}(x) \rangle \, d\mathcal{H}^k(x), \tag{9}$$

where the second integral is the standard Lebesgue integral with respect to the k-dimensional Hausdorff measure $\mathcal{H}^k$ restricted to $\mathcal{M}$, i.e., $(\mathcal{H}^k \llcorner \mathcal{M})(A) = \mathcal{H}^k(A \cap \mathcal{M})$. The k-dimensional Hausdorff measure assigns to sets in $\mathbf{R}^d$ their k-dimensional volume, see Chapter 2 in Morgan (2016) for a nice illustration. For $k = d$ the Hausdorff measure coincides with the Lebesgue measure.

The exterior derivative of a differential k-form is the $(k+1)$-form $d\omega : \mathbf{R}^d \to \Lambda^{k+1}\mathbf{R}^d$ defined by

$$\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1} \rangle = \lim_{h \to 0} \frac{1}{h^{k+1}} \int_{\partial P} \omega, \qquad (10)$$

where ∂P is the oriented boundary of the parallelotope spanned by the $\{hv_i\}$ at point x. The above definition is for example used in the textbook of Hubbard & Hubbard (2015). To get an intuition, note that for $k = 0$ this reduces to the familiar directional derivative $\langle d\omega(x), v_1 \rangle = \lim_{h\to 0} \frac{1}{h}(\omega(x + hv_1) - \omega(x))$. In case $\omega : \mathbf{R}^d \to \Lambda^k\mathbf{R}^d$ is sufficiently smooth, the limit in (10) is given by

$$\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1} \rangle = \qquad (11)$$
$$\sum_{i=1}^{k+1} (-1)^{i-1} \nabla_x \langle \omega(x), v_1 \wedge \ldots \wedge \hat{v}_i \wedge \ldots \wedge v_k \rangle \cdot v_i,$$

where $\hat{v}_i$ means that the vector v_i is omitted. The formulation (11) will be used in the practical implementation. Interestingly, with (9) and (10) in mind, Stokes' theorem

$$\int_{\mathcal{M}} d\omega = \int_{\partial \mathcal{M}} \omega, \qquad (12)$$

becomes almost obvious, as (informally speaking) integrating (10) one obtains (12) since the oppositely oriented boundaries of neighbouring parallelotopes cancel each other out in the interior of $\mathcal{M}$.

To define the pushforward of currents which is central to our formulation, we require the pullback of differential forms. The pullback $g^\sharp \omega : \mathbf{R}^l \to \Lambda^k\mathbf{R}^l$ by a map $g : \mathbf{R}^l \to \mathbf{R}^d$ of the k-form $\omega : \mathbf{R}^d \to \Lambda^k\mathbf{R}^d$ is given by

$$\langle g^\sharp \omega, v_1 \wedge .. \wedge v_k \rangle = \langle \omega \circ g, D_{v_1}g \wedge .. \wedge D_{v_k}g \rangle, \quad (13)$$

where $D_{v_i}g := \nabla g \cdot v_i$ and $\nabla g \in \mathbf{R}^{d \times l}$ is the Jacobian. We will also require (13) for the practical implementation.

3.3. Currents

We have now the necessary tools to define currents and the required operations on them, which will be defined through duality with differential forms. Consider the space of compactly supported and smooth k-forms in $\mathbf{R}^d$ which we denote by $\mathcal{D}^k(\mathbf{R}^d)$. When furnished with an appropriate topology (cf. §4.1 in Federer (1969) for the details) this is a locally convex topological vector space. k-currents are continuous linear functionals on smooth, compactly supported differential forms, i.e., elements from the topological dual space $\mathcal{D}_k(\mathbf{R}^d) = \mathcal{D}^k(\mathbf{R}^d)'$. Some examples for currents are given in Fig. 3. The 0-current in **(a)** could be an empirical data distribution, and the 2-current in **(b)** represents the data distribution with a two dimensional oriented tangent space at each data point. The 2-current in **(c)** simply represents the set $[0,1]^2$ as an oriented manifold, its action on a differential form is given as in (9).
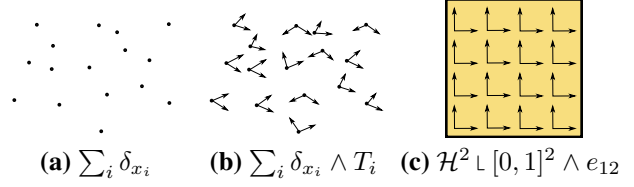

(a) $\sum_i \delta_{x_i}$ **(b)** $\sum_i \delta_{x_i} \wedge T_i$ **(c)** $\mathcal{H}^2 \mathbin{\llcorner} [0,1]^2 \wedge e_{12}$

Figure 3. Example of a 0-current **(a)**, and 2-currents **(b)**, **(c)**.

A natural notion of convergence for currents is given by the weak* topology:

$$T_i \overset{*}{\rightharpoonup} T \text{ iff } T_i(\omega) \to T(\omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbf{R}^d). \quad (14)$$

The support of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$, $\mathrm{spt}\, T$, is the complement of the largest open set, so that when testing T with compactly supported forms on that open set the answer is zero. Currents with compact support are denoted by $\mathcal{E}_k(\mathbf{R}^d)$. The boundary operator $\partial : \mathcal{D}_k(\mathbf{R}^d) \to \mathcal{D}_{k-1}(\mathbf{R}^d)$ is defined using exterior derivative

$$\partial T(\omega) = T(d\omega), \qquad (15)$$

and Stokes' theorem (12) ensures that this coincides with the intuitive notion of boundary for currents which are represented by integration over manifolds in the sense of (9).

The pushforward of a current is defined using the pullback

$$g_\sharp T(\omega) = T(g^\sharp \omega), \qquad (16)$$

where the intuition is that the pushforward transforms the current with the map g, see the illustration in Fig. 2.

The mass of a current $T \in \mathcal{D}_k(\mathbf{R}^d)$ is given by

$$\mathbb{M}(T) = \sup \{ T(\omega) : \|\omega(x)\|^* \le 1 \}. \qquad (17)$$

If the current T is an oriented manifold then the mass $\mathbb{M}(T)$ is the *volume* of that manifold. One convenient way to construct k-currents, is by combining a smooth k-vectorfield $\xi : \mathbf{R}^d \to \Lambda_k\mathbf{R}^d$ with a Radon measure μ:

$$(\mu \wedge \xi)(\psi) = \int \langle \xi, \psi \rangle \, d\mu, \text{ for all } \psi \in \mathcal{D}^k(\mathbf{R}^d). \quad (18)$$

A concrete example is illustrated in Fig. 3 **(b)**, where given samples $\{x_1, \ldots, x_N\} \subset \mathbf{R}^d$ and tangent 2-vectors $\{T_1, \ldots, T_N\} \subset \Lambda_2\mathbf{R}^d$ a 2-current is constructed.

For currents with finite mass there is a measure $\|T\|$ and a map $\vec{T} : \mathbf{R}^d \to \Lambda_k\mathbf{R}^d$ with $\|\vec{T}(\cdot)\| = 1$ almost everywhere so that we can represent it by integration as follows:

$$T(\omega) = \int \langle \omega(x), \vec{T}(x) \rangle \, d\|T\|(x) = \|T\| \wedge \vec{T}(\omega). \quad (19)$$

Another perspective is that finite mass currents are simply k-vector valued Radon measures. Currents with finite mass and finite boundary mass are called *normal currents* (Federer & Fleming, 1960). The space of normal currents with support in a compact set $\mathcal{X}$ is denoted by $\mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$.
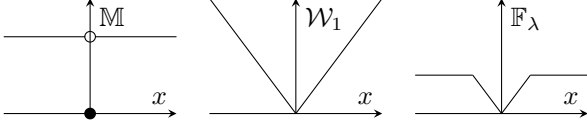
Figure 4. Illustration of distances between 0-currents on the example of two Dirac measures δ_x, δ_0. The flat metric $\mathbb{F}_\lambda$ has the following advantages: unlike the mass $\mathbb{M}$ it is continuous, and unlike Wasserstein-1 it easily generalizes to k-currents (see Fig. 5).
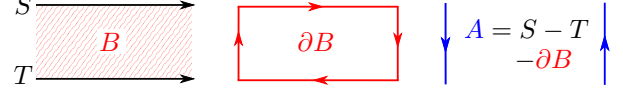

Figure 5. The flat metric $\mathbb{F}_\lambda(S, T)$ is given an optimal decomposition $S - T = A + \partial B$ into a k-current A and the boundary of a $(k + 1)$-current B with minimal weighted mass $\lambda\mathbb{M}(A) + \mathbb{M}(B)$. An intuition is that $\lambda\mathbb{M}(A)$ is a penalty that controls how closely ∂B should approximate $S - T$, while $\mathbb{M}(B)$ is the $(k + 1)$-dimensional volume of B.

4. The Flat Metric

As indicated in Fig. 2, we wish to fit a current $g_\sharp S$ that is the pushforward of a low-dimensional latent current S to the current T given by the data. A more meaningful norm on currents than the mass $\mathbb{M}$ turns out to be the flat norm.

Definition 2 (Flat norm and flat metric). *The flat norm with scale[2] $\lambda > 0$ is defined for any k-current $T \in \mathcal{D}_k(\mathbf{R}^d)$ as*

$$\mathbb{F}_\lambda(T) = \sup\{T(\omega) \mid \omega \in \mathcal{D}^k(\mathbf{R}^d), \text{ with} \tag{20}$$
$$\|\omega(x)\|^* \leq \lambda, \|d\omega(x)\|^* \leq 1, \text{ for all } x\}.$$

For $\lambda = 1$ we simply write $\mathbb{F}(\cdot) \equiv \mathbb{F}_1(\cdot)$ and $\mathbb{F}_\lambda(S, T) = \mathbb{F}_\lambda(S - T)$ will be denoted as the flat metric.

The flat norm also has a primal formulation

$$\mathbb{F}_\lambda(T) = \min_{B \in \mathcal{E}_{k+1}(\mathbf{R}^d)} \lambda\mathbb{M}(T - \partial B) + \mathbb{M}(B) \tag{21}$$

$$= \min_{T = A + \partial B} \lambda\mathbb{M}(A) + \mathbb{M}(B), \tag{22}$$

where the minimum in (21)–(22) can be shown to exist, see §4.1.12 in Federer (1969). The flat norm is finite if T is a normal current and it can be verified that it is indeed a norm.

To get an intuition, we compare the flat norm to the mass (17) and the Wasserstein-1 distance in Fig. 4 on the example of Dirac measures δ_x, δ_0. The mass $x \mapsto \mathbb{M}(\delta_x - \delta_0)$ is discontinuous and has zero gradient and is therefore unsuitable as a distance between currents. While the Wasserstein-1 metric $x \mapsto \mathcal{W}(\delta_x, \delta_0)$ is continuous in x, it does not easily generalize from probability measures to k-currents. In contrast, the flat metric $x \mapsto \mathbb{F}_\lambda(\delta_x, \delta_0)$ has a meaningful geometric interpretation also for arbitrary k-currents. In Fig. 5 we illustrate the flat norm for two 1-currents. In that figure, if S and T are of length one and are ε apart, then $\mathbb{F}_\lambda(S, T) \leq (1 + 2\lambda)\varepsilon$ which converges to zero for $\varepsilon \to 0$.

Note that for 0-currents, the flat norm (20) is strongly related to the Wasserstein-1 distance except for the additional constraint on the dual variable $\|\omega(x)\|^* \leq \lambda$, which in the example of Fig. 4 controls the truncation cutoff. Notice also

---

[2]We picked a different convention for λ as in (Morgan & Vixie, 2007), where it bounds the other constraint, to emphasize the connection to the Wasserstein-1 distance.

the similarity of (21) to the Beckmann formulation of the Wasserstein-1 distance (Beckmann, 1952; Santambrogio, 2015), with the difference being the implementation of the "divergence constraint" with a soft penalty $\lambda\mathbb{M}(T - \partial B)$. Considering the case $\lambda = \infty$ as in the Wasserstein distance is problematic in case we have $k > 0$, since not every current $T \in \mathcal{D}_k(\mathbf{R}^n)$ is the boundary of a $(k + 1)$-current, see the example above in Fig. 5.

The following proposition studies the effect of the scale parameter $\lambda > 0$ on the flat norm.

Proposition 1. *For any $\lambda > 0$, the following relation holds*

$$\min\{1, \lambda\} \cdot \mathbb{F}(T) \leq \mathbb{F}_\lambda(T) \leq \max\{1, \lambda\} \cdot \mathbb{F}(T), \tag{23}$$

meaning that $\mathbb{F}$ and $\mathbb{F}_\lambda$ are equivalent norms.

Proof. By a result of Morgan & Vixie (2007) we have the interesting relation

$$\mathbb{F}_\lambda(T) = \lambda^k \mathbb{F}(d_{\lambda^{-1}\sharp}T), \tag{24}$$

where d_λ is the λ-dilation. Using the bound $\mathbb{F}(f_\sharp T) \leq \sup\{\text{Lip}(f)^k, \text{Lip}(f)^{k+1}\}\mathbb{F}(T)$, §4.1.14 in Federer (1969), and the fact that $\text{Lip}(d_{\lambda^{-1}}) = \lambda^{-1}$, one inequality directly follows. For the other side, notice that

$$\mathbb{F}(T) = \mathbb{F}(d_{\lambda\sharp}d_{\lambda^{-1}\sharp}T) = \mathbb{F}_{\lambda^{-1}}(d_{\lambda^{-1}\sharp}T)\lambda^k$$
$$\leq \max\{1, \lambda^{-1}\}\mathbb{F}(d_{\lambda^{-1}\sharp}T)\lambda^k \tag{25}$$
$$= \max\{1, \lambda^{-1}\}\mathbb{F}_\lambda(T).$$

and dividing by $\max\{1, \lambda^{-1}\}$ yields the result. $\square$

The importance of the flat norm is due to the fact that it metrizes the weak*-convergence (14) on compactly supported normal currents with uniformly bounded mass and boundary mass.

Proposition 2. *Let $\mathcal{X} \subset \mathbf{R}^d$ be a compact set and $c > 0$ some fixed constant. For a sequence $\{T_j\} \subset \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ with $\mathbb{M}(T_j) + \mathbb{M}(\partial T_j) < c$ we have that:*

$$\mathbb{F}_\lambda(T, T_j) \to 0 \quad \text{if and only if} \quad T_j \overset{*}{\rightharpoonup} T. \tag{26}$$

Proof. Due to Prop. 1 it is enough to consider the case $\lambda = 1$, which is given by Corollary 7.3 in the paper of Federer & Fleming (1960). $\square$

5. Flat Metric Minimization

Motivated by the theoretical properties of the flat metric shown in the previous section, we consider the following optimization problem:

$$\min_{\theta \in \Theta} \mathbb{F}_\lambda(g_{\theta\sharp} S, T), \qquad (27)$$

where $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ and $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$. We will assume that $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is parametrized with parameters in a compact set $\Theta \subset \mathbf{R}^n$ and write $g_\theta : \mathcal{Z} \to \mathcal{X}$ to abbreviate $g(\cdot, \theta)$ for some $\theta \in \Theta$. We need the following assumption to be able to prove the existence of minimizers for the problem (27).

Assumption 1. *The map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ is smooth in z with uniformly bounded derivative. Furthermore, we assume that $g(z, \cdot)$ is locally Lipschitz continuous and that the parameter set $\Theta \subset \mathbf{R}^n$ is compact.*

Under this assumption, we will show that the objective in (27) is Lipschitz continuous. This will in turn guarantee existence of minimizers, as the domain is assumed to be compact.

Proposition 3. *Let $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$, $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ be normal currents with compact support. If the pushforward map $g : \mathcal{Z} \times \Theta \to \mathcal{X}$ fulfills Assumption 1, then the function $\theta \mapsto \mathbb{F}_\lambda(g_{\theta\sharp} S, T)$ is Lipschitz continuous and hence differentiable almost everywhere.*

Proof. In Appendix A. □

5.1. Application to Generative Modeling

We now turn towards our considered application illustrated in Fig. 2. There, we denote by $k \geq 0$ the number of tangent vectors we specify at each sample point. The latent current $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^l)$ is constructed by combining a probability distribution $\mu \in \mathbf{N}_{0,\mathcal{Z}}(\mathbf{R}^l)$, which could for example be the uniform distribution, with the unit k-vectorfield as follows:

$$S = \mu \wedge (e_1 \wedge \ldots \wedge e_k). \qquad (28)$$

For an illustration, see the right side of Fig. 2 and Fig. 3. The data current $T \in \mathbf{N}_{k,\mathcal{X}}(\mathbf{R}^d)$ is constructed from the samples $\{x_i\}_{i=1}^N$ and tangent vectorfields $T_i : \mathcal{X} \to \Lambda_k \mathbf{R}^d$.

$$T = \frac{1}{N} \sum_{i=1}^N \delta_{x_i} \wedge T_i, \qquad (29)$$

The tangent k-vectorfields $T_i(x) = T_{i,1} \wedge \ldots \wedge T_{i,k}$ are given by individual tangent vectors to the data manifold $T_{i,j} \in \mathbf{R}^d$. For an illustration, see the left side of Fig. 2 or Fig. 3. After solving (27), the map $g_\theta : \mathcal{Z} \to \mathcal{X}$ will be our generative model, where changes in the latent space $\mathcal{Z}$ along the unit directions $e_1, \ldots, e_k$ are expected to behave equivariantly to the specified tangent directions $T_{i,1}, \ldots, T_{i,k}$ near $g(z)$.

5.2. FlatGAN Formulation

To get a primal-dual formulation (or two player zero-sum game) in the spirit of GANs, we insert the definition of the flat norm (20) into the primal problem (27):

$$\min_{\theta \in \Theta} \sup_{\substack{\omega \in \mathcal{D}^k(\mathbf{R}^d) \\ \|\omega\|^* \leq \lambda, \|d\omega\|^* \leq 1}} S(g_\theta^\sharp \omega) - T(\omega), \qquad (30)$$

where $\theta \in \Theta$ are for example the parameters of a neural network. In the above equation, we also used the definition of pushforward (16). Notice that for $k = 0$ the exterior derivative in (30) specializes to the gradient. This yields a Lipschitz constraint, and as for sufficiently large λ the other constraint becomes irrelevant, the problem (30) is closely related to the Wasserstein GAN (Bottou et al., 2017). The novelty in this work is the generalization to $k > 0$.

Combining (28) and (29) into (30) we arrive at the objective

$$\begin{aligned} E(\theta, \omega) = &-\frac{1}{N} \sum_{i=1}^N \langle \omega(x_i), T_i \rangle \\ &+ \int \langle \omega \circ g_\theta, (\nabla_z g_\theta \cdot e_1) \wedge \ldots \wedge (\nabla_z g_\theta \cdot e_k) \rangle \, \mathrm{d}\mu. \end{aligned} \qquad (31)$$

Interestingly, due to the pullback, the discriminator ω inspects not only the output of the generator, but also parts of its Jacobian matrix. As a remark, relations between the generator Jacobian and GAN performance have recently been studied by Odena et al. (2018).

The constraints in (30) are implemented using penalty terms. First notice that due to the definition of the comass norm (6), the first constraint is equivalent to imposing $|\langle \omega(x), v \rangle| \leq \lambda$ for all simple k-covectors with $|v| = 1$. We implement this with the a penalty term with parameter $\rho > 0$ as follows:

$$\rho \cdot \int_{\mathcal{X}} \int \max\{0, |\langle \omega(x), v \rangle| - \lambda\}^2 \, \mathrm{d}\gamma_{k,d}(v) \, \mathrm{d}x, \quad (32)$$

where $\gamma_{k,d}$ denotes the Haar measure on the Grassmannian manifold $\mathbf{Gr}(d, k) \subset \Lambda_k \mathbf{R}^d$ of k-dimensional subspaces in $\mathbf{R}^d$, see Chapter 3.2 in Krantz & Parks (2008). Similarly, the constraint on the exterior derivative is implemented by another penalty term as follows:

$$\rho \cdot \int_{\mathcal{X}} \int \max\{0, |\langle d\omega(x), v \rangle| - 1\}^2 \, \mathrm{d}\gamma_{k+1,d}(v) \, \mathrm{d}x. \quad (33)$$

5.3. Implementation with Deep Neural Networks

For high dimensional practical problems it is completely infeasible to directly work with $\Lambda_k \mathbf{R}^d$ due to the curse of dimensionality. For example, already for the MNIST dataset augmented with two tangent vectors ($k = 2$, $d = 28^2$), we have that $\dim(\Lambda_k \mathbf{R}^d) \approx 3 \cdot 10^5$.

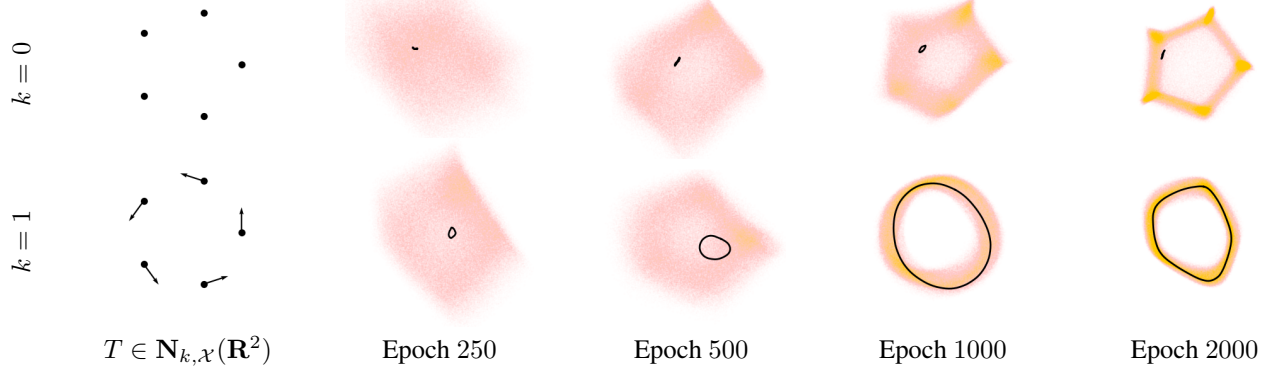*Figure 6.* We illustrate the effect of moving from $k = 0$ to $k = 1$ and plot the measure $\|g_\sharp S\|$ of the pushforward of a k-current $S \in \mathbf{N}_{k,\mathcal{Z}}(\mathbf{R}^5)$ (shown in orange) for different epochs. The black curve illustrates a walk along the first latent dimension z_1. For $k = 0$, which is similar to WGAN-GP (Gulrajani et al., 2017), the latent walk is not meaningful. The proposed approach ($k = 1$) allows to specify tangent vectors at the samples to which the first latent dimension behaves equivariantly, yielding an interpretable representation.
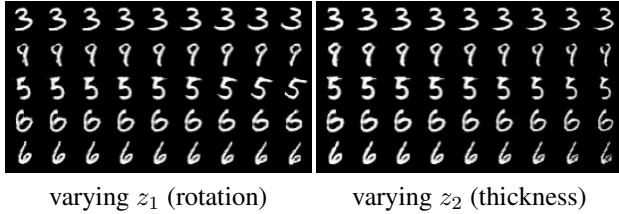

varying z_1 (rotation)   varying z_2 (thickness)

Figure 7. We show the effect of varying the first two components in 128-dimensional latent space, corresponding to the two selected tangent vectors which are rotation and thickness. As seen in the figure, varying the corresponding latent representation yields an interpretable effect on the output, corresponding to the specified tangent direction.

To overcome this issue, we unfortunately have to resort to a few heuristic approximations. To that end, we first notice that in the formulations the dual variable $\omega : \mathbf{R}^d \to \Lambda^k \mathbf{R}^d$ only appears as an inner product with simple k-vectors, so we can implement it by implicitly describing its action, i.e., interpret it as a map $\omega : \mathbf{R}^d \times \Lambda_k \mathbf{R}^d \to \mathbf{R}$:

$$\omega(x, v_1 \wedge \ldots \wedge v_k) \tag{34}$$
$$= \omega^0(x) + \alpha \langle \omega^{1,1}(x) \wedge \ldots \wedge \omega^{1,k}(x), v_1 \wedge \ldots \wedge v_k \rangle,$$

Theoretically, the "affine term" $\omega^0(x)$ is not fully justified as the map does not describe an inner product on $\Lambda_k \mathbf{R}^d$ anymore, but we found it to improve the quality of the generative model. An attempt to justify this in the context of GANs is that the function $\omega^0 : \mathbf{R}^d \to \mathbf{R}$ is the usual "discriminator" while the $\omega^{1,i} : \mathbf{R}^d \to \mathbf{R}^d$ are combined to discriminate oriented tangent planes.

In practice, we parametrize $\omega^0, \omega^{1,i}$ using deep neural networks. For efficiency reasons, the networks share their parameters up until the last few layers.

The inner product in (34) between the simple vectors is implemented by a $k \times k$-determinant, see (5). The reason we do this is to satisfy the properties of the Grassmann algebra (2) – (3). This is important, since otherwise the "discriminator" ω could distinguish between different representations of the same oriented tangent plane.

For the implementation of the penalty term (33), we use the definition of the exterior derivative (11) together with the "approximate form" (34). To be compatible with the affine term we use a seperate penalty on ω^0, which we also found to give better results:

$$|d\omega(x, v_1 \wedge \ldots \wedge v_{k+1})| \approx (k+1)\|\nabla_x \omega^0(x)\|$$
$$+ \alpha \left| \sum_{i=1}^{k+1} (-1)^{i-1} \nabla_x \det(W(x)^\top V_i) \cdot v_i \right|. \tag{35}$$

In the above equation, $V_i \in \mathbf{R}^{d \times k}$ is the matrix with columns given by the vectors $v_1, \ldots, v_{k+1}$ but with v_i omitted and $W(x) \in \mathbf{R}^{d \times k}$ is the matrix with columns given by the $\omega^{1,i}(x)$. Another motivation for this implementation is, that in the case $k = 0$ the second term in (35) disappears and one recovers the well-known "gradient penalty" regularizer proposed by Gulrajani et al. (2017).

For the stochastic approximation of the penalty terms (32) – (33) we sample from the Haar measure on the Grassmannian (i.e., taking random k-dimensional and $(k+1)$-dimensional subspaces in $\mathbf{R}^d$) by computing singular value decomposition of random $k \times d$ Gaussian matrices. Furthermore, we found it beneficial in practice to enforce the penalty terms only at the data points as for example advocated in the recent work (Mescheder et al., 2018). The right multiplied Jacobian vector products (also referred to as "`rop`" in some frameworks) in (35) as well as in the loss function (31) are implemented using two additional backpropagations.
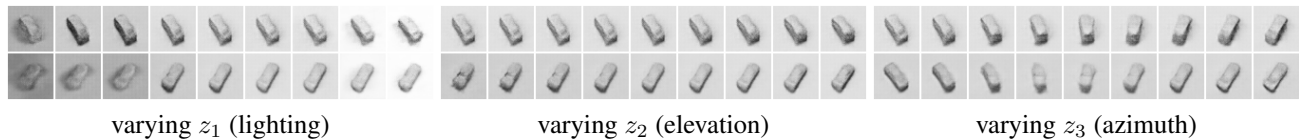
varying z_1 (lighting)   varying z_2 (elevation)   varying z_3 (azimuth)

Figure 8. From left to right we vary the latent codes in $[-1, 1]$ after training on the smallNORB dataset (LeCun et al., 2004).

varying z_1 (time)

Figure 9. Varying the learned latent representation of time. The model captures behaviours such as people walking on the beach, see also the results shown in Fig. 1.

6. Experiments

The specific hyperparameters, architectures and tangent vector setups used in practice[3] are detailed in Appendix B.

6.1. Illustrative 2D Example

As a first proof of concept, we illustrate the effect of moving from $k = 0$ to $k = 1$ on a very simple dataset consisting of five points on a circle. As shown in Fig. 6, for $k = 0$ (corresponding to a WGAN-GP formulation) varying the first latent variable has no clear meaning. In contrast, with the proposed FlatGAN formulation, we can specify vectors tangent to the circle from which the data is sampled. This yields an interpretable latent representation that corresponds to an angular movement along the circle. As the number of epochs is increasing, both formulations tend to concentrate most of the probability mass on the five data points. However, since $g_\theta : \mathcal{Z} \to \mathcal{X}$ is continuous by construction an interpretable path remains.

6.2. Equivariant Representation Learning

In Fig. 7 and Fig. 8 we show examples for $k = 2$ and $k = 3$ on MNIST respectively the smallNORB dataset of LeCun et al. (2004). For MNIST, we compute the tangent vectors manually by rotation and dilation of the digits, similar as done by Simard et al. (1992; 1998). For the smallNORB example, the tangent vectors are given as differences between the corresponding images. As observed in the figures, the proposed formulation leads to interpretable latent codes

---

[3]See https://github.com/moellenh/flatgan for a `PyTorch` implementation to reproduce Fig. 6 and Fig. 7.

which behave equivariantly with the generated images. We remark that the goal was not to achieve state-of-the-art image quality but rather to demonstrate that specifying tangent vectors yields disentangled representations. As remarked by Jaderberg et al. (2015), representing a 3D scene with a sequence of 2D convolutions is challenging and a specialized architecture based on a voxel representation would be more appropriate for the smallNORB example.

6.3. Discovering the Arrow of Time

In our last experiment, we set $k = 1$ and specify the tangent vector as the difference of two neighbouring frames in video data. We train on the tinyvideo beach dataset (Vondrick et al., 2016), which consists of more than 36 million frames. After training for about half an epoch, we can already observe a learned latent representation of time, see Fig. 1 and Fig. 9. We generate individual frames by varying the latent coordinate z_1 from -12.5 to 12.5.

Even though the model is trained on individual frames in random order, a somewhat coherent representation of time is discovered which captures phenomena such as ocean waves or people walking on the beach.

7. Discussion and Conclusion

In this work, we demonstrated that k-currents can be used introduce a notion of orientation into probabilistic models. Furthermore, in experiments we have shown that specifying partial tangent information to the data manifold leads to interpretable and equivariant latent representations such as the camera position and lighting in a 3D scene or the arrow of time in time series data.

The difference to purely unsupervised approaches such as InfoGAN or β-VAE is, that we can encourage potentially very complex latent representations to be learned. Nevertheless, an additional mutual information term as in (Chen et al., 2016) can be directly added to the formulation so that some representations could be encouraged through tangent vectors and the remaining ones are hoped to be discovered in an unsupervised fashion.

Generally speaking, we believe that geometric measure theory is a rather underexploited field with many possible application areas in probabilistic machine learning. We see this work as a step towards leveraging this potential.

Acknowledgements

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Beckmann, M. A continuous model of transportation. *Econometrica: Journal of the Econometric Society*, pp. 643–660, 1952.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Bottou, L., Arjovsky, M., Lopez-Paz, D., and Oquab, M. Geometrical insights for implicit generative modeling. *arXiv:1712.07822*, 2017.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

Csiszár, I., Shields, P. C., et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

de Rham, G. *Variétés différentiables, formes, courants, formes harmoniques*, volume 1222. Hermann, 1955.

Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, 2017.

Federer, H. *Geometric Measure Theory*. Springer, 1969.

Federer, H. and Fleming, W. H. Normal and integral currents. *Annals of Mathematics*, pp. 458–520, 1960.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouvé, A., and Peyré, G. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *arXiv:1810.08278*, 2018.

Fraser, A. M., Hengartner, N. W., Vixie, K. R., and Wohlberg, B. E. Incorporating invariants in Mahalanobis distance based classifiers: Application to Face Recognition. In *International Joint Conference on Neural Networks*, 2003.

Genevay, A., Peyré, G., and Cuturi, M. GAN and VAE from an optimal transport point of view. *arXiv:1706.01807*, 2017.

Glaunès, J., Qiu, A., Miller, M. I., and Younes, L. Large deformation diffeomorphic metric curve mapping. *International Journal of Computer Vision (IJCV)*, 80(3):317, 2008.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *arXiv:1704.00028*, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β–VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, 2011.

Hubbard, J. H. and Hubbard, B. B. *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Matrix Editions, 2015.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

Kim, H. and Mnih, A. Disentangling by factorising. *arXiv:1802.05983*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2014.

Krantz, S. G. and Parks, H. R. *Geometric Integration Theory*. Birkhäuser Boston, 2008.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, 2016.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually Converge? In *International Conference on Machine Learning*, 2018.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Möllenhoff, T. and Cremers, D. Lifting vectorial variational problems: A natural formulation based on geometric measure theory and discrete exterior calculus. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Morgan, F. *Geometric Measure Theory: A Beginner's Guide*. Academic Press, 5th edition, 2016.

Morgan, S. P. and Vixie, K. R. L^1TV computes the flat norm for boundaries. In *Abstract and Applied Analysis*, 2007.

Narayanaswamy, S., Paige, T. B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, 2017.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, 2017.

Odena, A., Buckman, J., Olsson, C., Brown, T. B., Olah, C., Raffel, C., and Goodfellow, I. Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning*, 2018.

Peyré, G. and Cuturi, M. Computational optimal transport. *arXiv:1803.00567*, 2018.

Pickup, L. C., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Schölkopf, B., and Freeman, W. T. Seeing the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082*, 2014.

Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, 2011.

Santambrogio, F. *Optimal Transport for Applied Mathematicians*. Birkhäuser, New York, 2015.

Schmidhuber, J. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Schwartz, L. *Théorie des distributions I, II*, volume 1245, 1122. Hermann, 1951, 1957.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop – a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, 1992.

Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. Transformation invariance in pattern recognition – tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pp. 239–274, 1998.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R., et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Vaillant, M. and Glaunès, J. Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, 2005.

Vixie, K. R., Clawson, K., Asaki, T. J., Sandine, G., Morgan, S. P., and Price, B. Multiscale flat norm signatures for shapes and images. *Applied Mathematical Sciences*, 4 (14):667–680, 2010.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, 2016.

Wei, D., Lim, J. J., Zisserman, A., and Freeman, W. T. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Whitney, H. *Geometric Integration Theory*. Princeton University Press, 1957.