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.

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 \(\mathcal{Z} \subset \mathbb{R}^l\) that is pushed through a decoder or generator \(g : \mathcal{Z} \to \mathcal{X}\) mapping to the (high-dimensional) data space \(\mathcal{X} \subset \mathbb{R}^d\) (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.
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 \( X \subset \mathbb{R}^d \), denoted by \( \mathbb{N}_{k,X}(\mathbb{R}^d) \).

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 successfully 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 (Möllerhoff & Cremers, 2019).

\(^1\)The terminology “flat” carries no geometrical significance and refers to Whitney’s use of musical notation flat \( \| \cdot \| \) and sharp \( \| \cdot \|^\flat\).

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 principle 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 \( \mathbb{R}^d \) with dual basis \( \{dx_1, \ldots, dx_d\} \) such that \( dx_i : \mathbb{R}^d \to \mathbb{R} \) 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 \( I = (i_1, \ldots, i_k) \) with \( 1 \leq i_1 < \ldots < i_k \leq d \).

One can multiply vectors in \( \mathbb{R}^d \) to obtain a new object:

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

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

\[
(c v_1 \wedge v_2) = c (v_1 \wedge v_2), \quad (u_1 + v_1) \wedge (u_2 + v_2) = u_1 \wedge u_2 + u_1 \wedge v_2 + v_1 \wedge u_2 + v_1 \wedge v_2, \tag{2}
\]

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

\[
\xi = \sum_{i \in I(d,k)} a_i e_{i_1} \wedge \ldots \wedge e_{i_k} = \sum_{i \in I(d,k)} a_i e_i, \tag{4}
\]

for coefficients \( a_i \in \mathbb{R} \). The vector space of \( k \)-vectors is denoted by \( \Lambda_k \mathbb{R}^d \) and has dimension \( \binom{d}{k} \). We define for two \( k \)-vectors \( v = \sum_i a_i e_i, w = \sum_i b_i e_i \) an inner product \( \langle v, w \rangle = \sum_i a_i b_i \) and the Euclidean norm \( |v| = \sqrt{\langle v, v \rangle} \).

A simple (or decomposable) \( k \)-vector is any \( \xi \in \Lambda_k \mathbb{R}^d \) 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 parallelepiped 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

\[
\langle w_1 \wedge \ldots \wedge w_k, v_1 \wedge \ldots \wedge v_k \rangle = \det(W^T V), \tag{5}
\]

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

Not all \( k \)-vectors are simple. An illustrative example is \( e_1 \wedge e_2 + e_3 \wedge e_4 \in \Lambda_2 \mathbb{R}^d \), which describes two 2-dimensional subspaces in \( \mathbb{R}^d \) intersecting only at zero.

The dual space of \( \Lambda_k \mathbb{R}^d \) is denoted as \( \Lambda^k \mathbb{R}^d \), and its elements are called \( k \)-covectors. They are similarly represented as (4) but with dual basis \( dx_i \). 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 \( w \in \Lambda^k \mathbb{R}^n \) is given by

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

and the mass norm for \( v \in \Lambda_k \mathbb{R}^n \) is given by

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

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 \( \omega : \mathbb{R}^d \to \Lambda^k \mathbb{R}^d \). The support \( \text{spt} \omega \) is defined as the closure of the set \( \{x \in \mathbb{R}^d : \omega(x) \neq 0\} \).

Differential forms allow one to perform coordinate-free integration over oriented manifolds. Given some manifold \( M \subset \mathbb{R}^d \), possibly with boundary, an orientation is a continuous map \( \tau_M : M \to \Lambda_k \mathbb{R}^d \) 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 \( M \) is then defined by:

\[
\int_M \omega = \int_M \langle \omega(x), \tau_M(x) \rangle d\mathcal{H}^k(x), \tag{9}
\]

where the second integral is the standard Lebesgue integral with respect to the \( k \)-dimensional Hausdorff measure \( \mathcal{H}^k \) restricted to \( M \), i.e., \( (\mathcal{H}^k \circ M)(A) = \mathcal{H}^k(A \cap M) \). The \( k \)-dimensional Hausdorff measure assigns to sets in \( \mathbb{R}^d \) 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 \(d\omega : \mathbb{R}^d \to \Lambda^{k+1}\mathbb{R}^d\) defined by
\[
\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1} \rangle = \lim_{h \to 0} \frac{1}{h} \int_{\partial P} \omega,
\]
where \(\partial P\) is the oriented boundary of the parallelotope spanned by \(\{v_i\}\) 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 \(\omega : \mathbb{R}^d \to \Lambda^k\mathbb{R}^d\) is sufficiently smooth, the limit in (10) is given by
\[
\langle d\omega(x), v_1 \wedge \ldots \wedge v_{k+1} \rangle = \sum_{i=1}^{k+1} (-1)^{i-1} \nabla_x \langle \omega(x), v_1 \wedge \ldots \wedge \hat{v}_i \wedge \ldots \wedge v_{k+1} \rangle \cdot v_i,
\]
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
\[
\int_M d\omega = \int_{\partial M} \omega,
\]
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 \(M\).

To define the pushforward of currents which is central to our formulation, we require the pullback of differential forms. The pullback \(g^*\omega : \mathbb{R}^d \to \Lambda^k\mathbb{R}^d\) by a map \(g : \mathbb{R}^d \to \mathbb{R}^d\) of the \(k\)-form \(\omega : \mathbb{R}^d \to \Lambda^k\mathbb{R}^d\) is given by
\[
\langle g^*\omega(x), v_1 \wedge \ldots \wedge v_k \rangle = \langle \omega \circ g, D_{v_1} g \wedge \ldots \wedge D_{v_k} g \rangle,
\]
where \(D_{v_i} g := \nabla g \cdot v_i\) and \(\nabla g \in \mathbb{R}^{d \times d}\) 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 \(\mathbb{R}^d\) which we denote by \(\mathcal{D}^k(\mathbb{R}^d)\). 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 \(\mathcal{D}'(\mathbb{R}^d) = \mathcal{D}^{k\prime}(\mathbb{R}^d)\). 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 natural notion of convergence for currents is given by the weak* topology:
\[
T_i \rightharpoonup T \text{ iff } T_i(\omega) \to T(\omega), \text{ for all } \omega \in \mathcal{D}^k(\mathbb{R}^d).
\]

The support of a current \(T \in \mathcal{D}_k(\mathbb{R}^d)\), \(\text{supp} \, T\), 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 \(\mathcal{E}_k(\mathbb{R}^d)\). The boundary operator \(\partial : \mathcal{D}_k(\mathbb{R}^d) \to \mathcal{D}_{k-1}(\mathbb{R}^d)\) is defined using exterior derivative
\[
\partial T(\omega) = T(d\omega),
\]
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
\[
g_*T(\omega) = T(g^*\omega),
\]
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 \(T \in \mathcal{D}_k(\mathbb{R}^d)\) is given by
\[
\mathcal{M}(T) = \sup \{T(\omega) : \|\omega(x)\|^* \leq 1\}.
\]

If the current \(T\) is an oriented manifold then the mass \(\mathcal{M}(T)\) is the volume of that manifold. One convenient way to construct \(k\)-currents, is by combining a smooth \(k\)-vectorfield \(\xi : \mathbb{R}^d \to \Lambda^k\mathbb{R}^d\) with a Radon measure \(\mu\):
\[
(\mu \wedge \xi)(\psi) = \int \langle \xi, \psi \rangle \, d\mu, \text{ for all } \psi \in \mathcal{D}^k(\mathbb{R}^d).
\]

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

For currents with finite mass there is a measure \(\|T\|\) and a map \(\bar{T} : \mathbb{R}^d \to \Lambda_k\mathbb{R}^d\) with \(\|\bar{T}(\cdot)\| = 1\) almost everywhere so that we can represent it by integration as follows:
\[
T(\omega) = \int \langle \omega(x), \bar{T}(x) \rangle \, d\|T\| = \|T\| \wedge \bar{T}(\omega).
\]

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 \(X\) is denoted by \(\mathcal{N}_{k,X}(\mathbb{R}^d)\).
For Figure 4. As indicated in Fig. 2, we wish to fit a current \( g_2 S \) 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 \( M \) turns out to be the flat norm.

**Definition 2** (Flat norm and flat metric). The flat norm with scale \( \lambda > 0 \) is defined for any \( k \)-current \( T \in D_k(\mathbb{R}^d) \) as

\[
F_\lambda(T) = \sup\{ T(\omega) \mid \omega \in D^k(\mathbb{R}^d), \|\omega(x)\|^* \leq \lambda, \|d\omega(x)\|^* \leq 1, \text{ for all } x \}.
\]

For \( \lambda = 1 \) we simply write \( F(\cdot) \equiv F_1(\cdot) \) and \( F_\lambda(S,T) = F_\lambda(S-T) \) will be denoted as the flat metric.

The flat norm also has a primal formulation

\[
F_\lambda(T) = \min_{B \in E_{k+1}(\mathbb{R}^d)} \lambda M(T - \partial B) + M(B), \quad \text{(20)}
\]

\[
= \min_{T = A + \partial B} \lambda M(A) + M(B), \quad \text{(21)}
\]

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 \( \delta_x, \delta_0 \). The mass \( x \mapsto M(\delta_x - \delta_0) \) is discontinuous and has zero gradient and is therefore unsuitable as a distance between currents. While the Wasserstein-1 metric \( x \mapsto W_1(\delta_x, \delta_0) \) is continuous in \( x \), it does not easily generalize from probability measures to \( k \)-currents. In contrast, the flat metric \( x \mapsto F_\lambda(\delta_x, \delta_0) \) 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 \( \varepsilon \) apart, then \( F_\lambda(S,T) \leq (1+2\lambda)\varepsilon \) 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

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

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 \( \lambda M(T - \partial B) \). Considering the case \( \lambda = \infty \) as in the Wasserstein distance is problematic in case we have \( k > 0 \), since not every current \( T \in D_k(\mathbb{R}^d) \) 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

\[
\min\{1, \lambda\} \cdot F(T) \leq F_\lambda(T) \leq \max\{1, \lambda\} \cdot F(T), \quad \text{(23)}
\]

meaning that \( F \) and \( F_\lambda \) are equivalent norms.

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

\[
F_\lambda(T) = \lambda^k F(d_{\lambda^{-1}} T), \quad \text{(24)}
\]

where \( d_{\lambda} \) is the \( \lambda \)-dilation. Using the bound \( F(f_2 T) \leq \sup\{\text{Lip}(f)^k, \text{Lip}(f)^{k+1}\} F(T) \), §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

\[
F(T) = F(d_{\lambda} d_{\lambda^{-1}} T) = F_{\lambda^{-1}}(d_{\lambda^{-1}} T) \lambda^k \leq \max\{1, \lambda^{-1}\} F(d_{\lambda^{-1}} T) \lambda^k \quad \text{(25)}
\]

and dividing by \( \lambda^k \) yields the result.

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 \( X \subset \mathbb{R}^d \) be a compact set and \( c > 0 \) some fixed constant. For a sequence \( \{T_j\} \subset N_{k,c}(\mathbb{R}^d) \) with \( M(T_j) + M(\partial T_j) < c \) we have that

\[
F_X(T_j, T) \to 0 \quad \text{if and only if} \quad T_j \overset{\lambda}{\rightharpoonup} T. \quad \text{(26)}
\]

**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).
5. Flat Metric Minimization

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

$$\min_{\delta \in \Theta} F_\lambda (g_{\theta_0} S, T),$$

(27)

where $S \in N_{k, Z}(R^l)$ and $T \in N_{k, X}(R^d)$. We will assume that $g : Z \times \Theta \to X$ is parametrized with parameters in a compact set $\Theta \subset R^n$ and write $g_\theta : Z \to X$ 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 $g : Z \times \Theta \to X$ is smooth in $z$ with uniformly bounded derivative. Furthermore, we assume that $g(z, \cdot)$ is locally Lipschitz continuous and that the parameter set $\Theta \subset R^n$ 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 $S \in N_{k, Z}(R^l)$, $T \in N_{k, X}(R^d)$ be normal currents with compact support. If the pushforward map $g : Z \times \Theta \to X$ fulfills Assumption 1, then the function $\theta \mapsto F_\lambda (g_{\theta} S, T)$ 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 $S \in N_{k, Z}(R^l)$ is constructed by combining a probability distribution $\mu \in N_{0, Z}(R^l)$, which could for example be the uniform distribution, with the unit $k$-vectorfield as follows:

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

(28)

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

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

(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 $T_{i,j} \in R^d$. For an illustration, see the left side of Fig. 2 or Fig. 3. After solving (27), the map $g_\theta : Z \to X$ will be our generative model, where changes in the latent space $Z$ 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):

$$\min_{\theta \in \Theta} \sup_{\omega \in \mathcal{H}^n(X^d)} S(g_\theta \omega) - T(\omega),$$

(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 $\lambda$ 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

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

(31)

Interestingly, due to the pullback, the discriminator $\omega$ 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:

$$\rho \cdot \int_X \int \max\{0, |\langle \omega(x), v \rangle| - \lambda\}^2 d\gamma_{k,d}(v) dx,$$

(32)

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

$$\rho \cdot \int_X \int \max\{0, |\langle d\omega(x), v \rangle| - 1\}^2 d\gamma_{k+1,d}(v) dx.$$

(33)

5.3. Implementation with Deep Neural Networks

For high dimensional practical problems it is completely infeasible to directly work with $A_k R^d$ 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 $\dim(A_k R^d) \approx 3 \cdot 10^5$. 
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” $\omega$ 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 separate penalty on $\omega^0$, which we also found to give better results:

$$
|d\omega(x, v_1 \ldots v_{k+1})| \approx (k+1)\|\nabla_x \omega^0(x)\| + \alpha \sum_{i=1}^{k+1}(-1)^{i-1}\det(W(x)^T v_i) \cdot v_i.
$$

(35)

In the above equation, $V_i \in \mathbb{R}^{d \times k}$ is the matrix with columns given by the vectors $v_1, \ldots, v_{k+1}$ but with $v_i$ omitted and $W(x) \in \mathbb{R}^{d \times k}$ 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 $\mathbb{R}^d$) 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 “crop” in some frameworks) in (35) as well as in the loss function (31) are implemented using two additional backpropagations.
Flat Metric Minimization with Applications in Generative Modeling

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

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 $g_\theta : Z \to X$ 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 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 to 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 $\beta$-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.

\(^3\)See [https://github.com/moellenh/flatgan](https://github.com/moellenh/flatgan) for a PyTorch implementation to reproduce Fig. 6 and Fig. 7.
Acknowledgements

We thank Kevin R. Vixie for his detailed feedback and comments on the manuscript. The work was partially supported by the German Research Foundation (DFG); project 394737018 “Functional Lifting 2.0 – Efficient Convexifications for Imaging and Vision”.

References


