Conservativeness of untied auto-encoders

Daniel Jiwoong Im[*]
Montreal Institute for Learning Algorithms
University of Montreal
Montreal, QC, H3C 3J7
imdaniel@iro.umontreal.ca

Mohamed Ishmael Belghazi[*]
HEC Montreal
3000 Ch de la Cte-Ste-Catherine
Montreal, QC, H3T 2A7
mohamed.2.belghazi@hec.ca

Roland Memisevic
Montreal Institute for Learning Algorithms
University of Montreal
Montreal, QC, H3C 3J7
roland.memisevic@umontreal.ca

In this supplementary material, we present some background information, derivations, and supporting explanations.

Conservative untied auto-encoders

Towards Poincare's criterion for untied AE using differential forms

This section provides detailed derivations of *Proposition 1* in Section 3.

Proposition 1. *Consider an m-hidden-layer auto-encoder defined as*

$$r(\mathbf{x};\theta) = W^{(m)}h^{(m)}\Big(W^{(m-1)}h^{(m-1)} \Big(\cdots W^{(1)}h^{(1)}\left(\mathbf{x}\right)\cdots\Big) + \mathbf{c}^{(m-1)}\Big) + \mathbf{c}^{(m)},$$

where $\theta = \cup_{k=0}^{m}\theta^{(k)}$ such that $\theta^{(k)} = \{W^{(k)}, \mathbf{c}^{(k)}\}$ are the parameters of the model, and $h^{(k)}(\cdot)$ is a smooth elementwise activation function at layer k. Then the auto-encoder is said to be conservative over a smooth simply connect domain $K \subseteq \mathbb{R}^D$ if and only if its reconstruction's Jacobian $\frac{\partial r(\mathbf{x})}{\partial \mathbf{x}}$ is symmetric for all $\mathbf{x} \in K$.

The high level idea is that simply finding the antiderivative of an auto-encoder vector field as proposed in (Kamyshanska 2013) does not work for untied auto-encoders. This is due to the difference in solving first order ordinary differential equations for tied auto-encoders and first order partial differential equations for untied auto-encoders. Therefore, here we present a different approach that uses differential forms to facilitate the derivation of the existence condition of a potential energy function in the case of untied auto-encoders.

The advantage of differential forms is that they allow us to work with a generalized, coordinate free system. A differential form α of degree l (l-form) on a smooth domain $K \subseteq \mathbb{R}^d$ is an expression:

$$\alpha = \sum_{i=1}^{D} f_i dx_i. \tag{1}$$

Using differential form algebra and exterior derivatives, we can show that the 1-form implied by an untied auto-encoder is exact, which means that α can be expressed as $\alpha = d\beta$ for some $\beta \in \Lambda^{l-1}(K)$. Let α be the 1-form implied by the vector field of an untied auto-encoder. Then, we have

$$\alpha = \sum_{i=1}^{D} r_i dx_i, \text{ and } d\alpha = \sum_{i=1}^{D} d(r_i \wedge dx_i) \tag{2}$$

where $\wedge$ is the exterior multiplication, d is the differential operator on differential forms, and $r(\cdot)$ is the reconstruction function of the auto-encoder. Based on the exterior derivative properties, i) if $f \in \Lambda^0(K)$ then $df = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} dx_i$ and ii) if $\alpha \in \Lambda^l(K)$ and $\beta \in \Lambda^m(K)$ then $\alpha\beta = (-1)^{lm}\beta\alpha$,

$$d\alpha = \sum_{i=1}^{D} d(r_i \wedge dx_i) \tag{3}$$

$$= \sum_{i,j=1}^{D} \frac{\partial r_i}{\partial x_j}(dx_j \wedge dx_i) \tag{4}$$

$$= -\sum_{1\leq i<j<D} \frac{\partial r_i}{\partial x_j} dx_i \wedge dx_j + \sum_{1\leq i<j<D} \frac{\partial r_j}{\partial x_i} dx_i \wedge dx_j \tag{5}$$

$$= \sum_{1\leq i<j<D} \left(\frac{\partial r_i}{\partial x_j} - \frac{\partial r_j}{\partial x_i}\right) dx_i \wedge dx_j \tag{6}$$

According to the Poincare's theorem, which states that every exact form is closed and conversely, if α is closed then it is exact in a simply connected region and $\alpha \in \Lambda^l(K)$, where α is closed if $d\alpha = 0$. Then, by Poincare's theorem, we see that

$$d\alpha = \sum_{1\leq i<j<D} \left(\frac{\partial r_i}{\partial x_j} - \frac{\partial r_j}{\partial x_i}\right) dx_i \wedge dx_j = 0 \tag{7}$$

This is equivalent to requiring the Jacobian to be symmetric for all $\mathbf{x} \in K$

Relations between the sufficient conditions

Let's re-state the two sufficient conditions:

1. If $R = C\hat{W}$ such that C is symmetric and commutes with WW^T, then the auto-encoder vector field is conservative.

[*] Authors constributed equally.

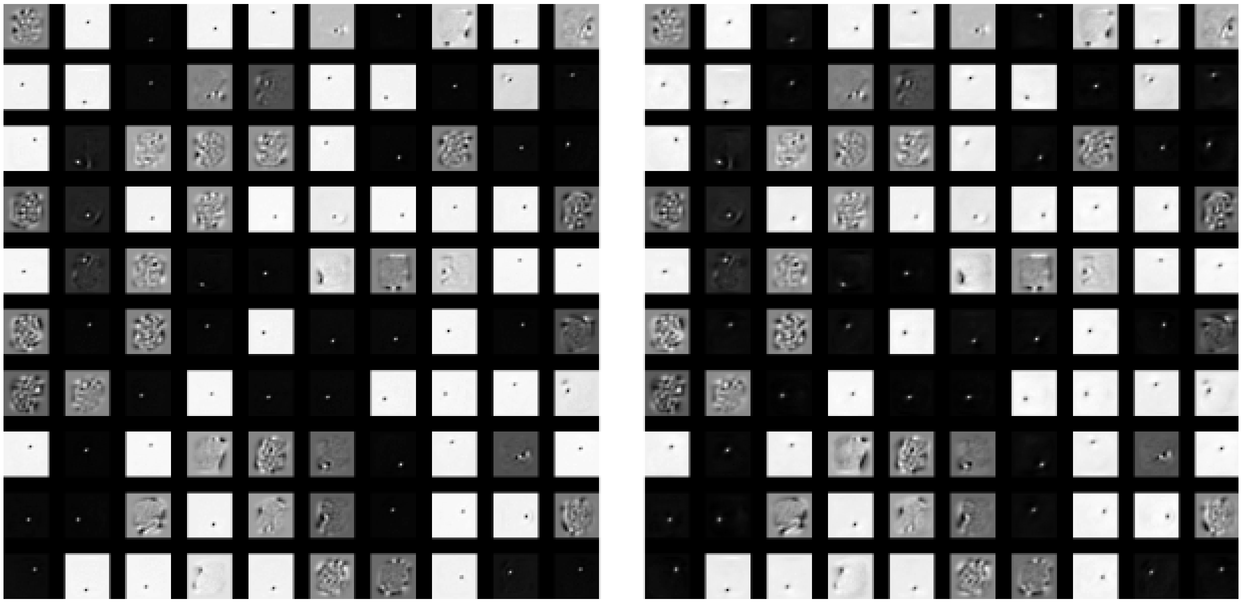


Figure 1: The weights of encoder W (Left) and weights of decoder R^T (Right) for a contractive auto-encoder trained with weight length constraints are shown.

2. If $R = \hat{W}E$ such that E is diagonal matrix, then the auto-encoder vector field is conservative.

We can try to understand the role of symmetric matrix C through the lens of spectral decomposition. Note that two symmetric matrices commute if they share the same eigenspace. Then,

$$CWD_{h'}D_{h'}W^T = Q\Lambda Q^T Q\Sigma Q^T = Q\Lambda\Sigma Q^T \qquad (8)$$

where $Q\Lambda Q^T$ is the eigen decomposition of C and $Q\Sigma Q^T$ is the eigen decomposition of $WD_{h'}D_{h'}W^T$. This illustrates that one can find a C based on choosing an appropriate matrix Λ, where Λ merely stretches or shrinks along the direction of the eigenvectors. Additionally, the role of the diagonal E in $R = WD_{h'}E$ can be explained as scaling the pre-activation of the hidden units. This can be directly observed re-expressing the condition in terms of elementwise operations as

$$R_{jl} = W_{jl}E_{ll}\,\forall l = 1\cdots H, \forall j = 1\cdots D. \qquad (9)$$

This amounts to "brightening" or "dimming" the filters $R_{\cdot l}$ depending on the diagonal matrix of E_{ll}.

Now, we find C and E given the parameters W and R. If $R = C\hat{W}$ such that C is symmetric and commutes with $\hat{W}\hat{W}^T$, then we know that $R\hat{W}^T = \hat{W}R^T$. Then, we can find C as follows:

$$\begin{aligned} R\hat{W}^T &= \hat{W}R^T \\ C\hat{W}\hat{W}^T &= \hat{W}R^T \\ C &= \hat{W}R^T A^T(AA^T)^{-1} \end{aligned}$$

where $A = \hat{W}\hat{W}^T$ and $A^T(AA^T)^{-1}$ is the right pseudo-inverse. Similarly, we can also compute E as follows:

$$\begin{aligned} R\hat{W}^T &= \hat{W}R^T \\ \hat{W}E\hat{W}^T &= \hat{W}R^T \\ E &= (\hat{W}^T\hat{W})^{-1}\hat{W}^T\hat{W}R^T\hat{W}(\hat{W}^T\hat{W})^{-1} \\ E &= R^T\hat{W}(\hat{W}^T\hat{W})^{-1} \end{aligned}$$

since $(W^TW)^{-1}W^TW = I$.

Further experiments on symmetry

One natural way to regularize the auto-encoder is to use a weight length constraint $||\mathbf{w}_i||^2 = \alpha$ for all $i = 1\cdots H$. A weight length constraints also implies a form of contraction, because the contractive term (norm of Jacobian matrix w.r.t hidden units) contains the term $\|W\|^2$.

Figure 1 demonstrates that weight length indeed helps $\frac{\partial r(\mathbf{x})}{\partial \mathbf{x}}$ to be more symmetric. An auto-encoder trained with weight length constraints, denoted as "AE sig wl", achieves a symmetry score of 0.9914 and, and a CAE with weight length constraints denoted as "CAE sig wl", a symmetric score of 0.9916.

Furthermore, as shown in Figurew1, the encoder and decoder weights of the contrastive auto-encoder trained with weight length constraints. are indistinguishable from one another. The increase in the symmetricity is clear when they are compared to Figure 1 in the original paper. In that case, the decoder weights are "smoothed" versions of the encoder weights, which is not the case here. This implies that R is becoming like W when we straightly enforce $\|W\| = \|R\|$, which brings back to having a symmetric auto-encoder.

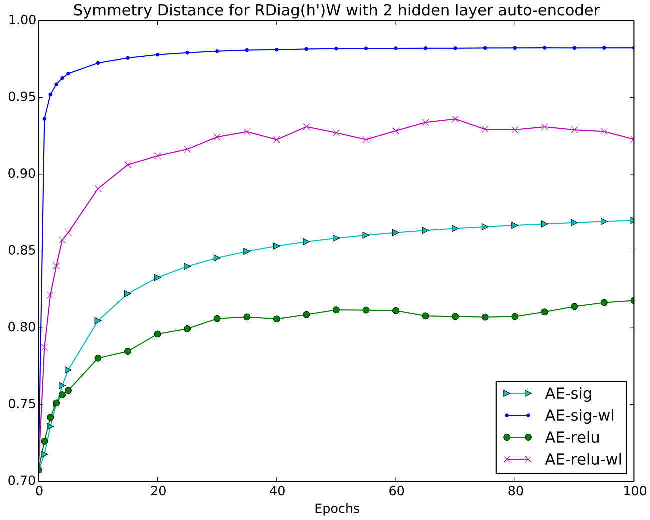


Figure 2: Figure presents symmetric distance of $\frac{\partial r(\mathbf{x})}{\partial \mathbf{x}}$ for two hidden layer untied auto-encoder

Based on above results, we see that our sufficient condition could explain what is going on with filters for the auto-encoder with sigmoid activation. Moreover, having a weight length constraints, $\|W\| = \|R\|$ leads to $R \rightarrow W$.

$$\sum_k h'(\mathbf{w}_k^T \mathbf{x}) \|\mathbf{w}\|^2 < D \tag{10}$$

will make the point $\mathbf{x}$ to be sink of the auto-encoders dynamics is desirable condition.

Next, an obvious question to ask is will deeper auto-encoders be symmetric as well? We plotted $\frac{\partial r(\mathbf{x})}{\partial \mathbf{x}}$ for two hidden layer untied auto-encoders. Figure 2 illustrates that they have harder time becoming fully symmetric, but still desire for symmetricity.

Another way of explicitly measuring the conservativeness as the auto-encoder gets trained is to look at the curl. In our experiments, we created three 2D syntehtic datasets to understand the auto-encoder's dynamics while learning. The three datasets consists of manifolds that looks like line, circle, and spiral. We looked at the changes in the vector field of before training, intermediate stage, and final stage. We also examined the magnitude of the curl and show that magnitude of curl decreases during the training. We also notice that sigmoid activation deforms the vector fields very slowly whereas, having ReLU activion function, changes the vector fields very rapidly. Figure 4, 5, and 6 shows the initial, final vector fields, and the magnitude of curl near the manifold. The colour of vector field indicates the mean reconstruction error rates.

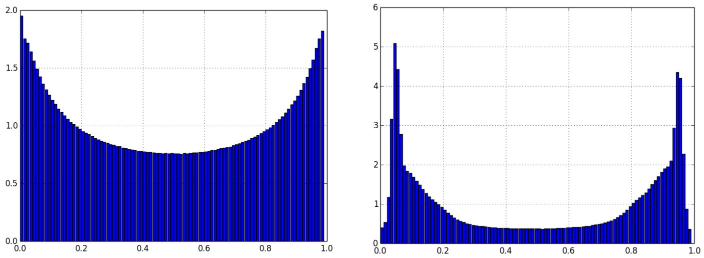


Figure 3: Histograms of hidden activations for sigmoid units without weight constraints (left) and with weight constraints (right)

In fact, after conducting this experiments, we argue that studying the dynamics of vector field is a excellent way of understanding the changes in the energy surface during the training, because we get the idea of how energy surface deforms by observing the changes in the vector fields.

Autoencoder	ReLU	ReLU+wl	sig.+wl	sig.+wl
AE	95.9%	98.7%	95.1%	99.1%
CAE	95.2%	98.6%	97.4%	99.1%

Table 1: Symmeticity of ADW after training AEs with 500 units on MNIST for 100 epochs. We denote the auto-encoders with weight length constraints as '+wl'.

Decomposing the vector field

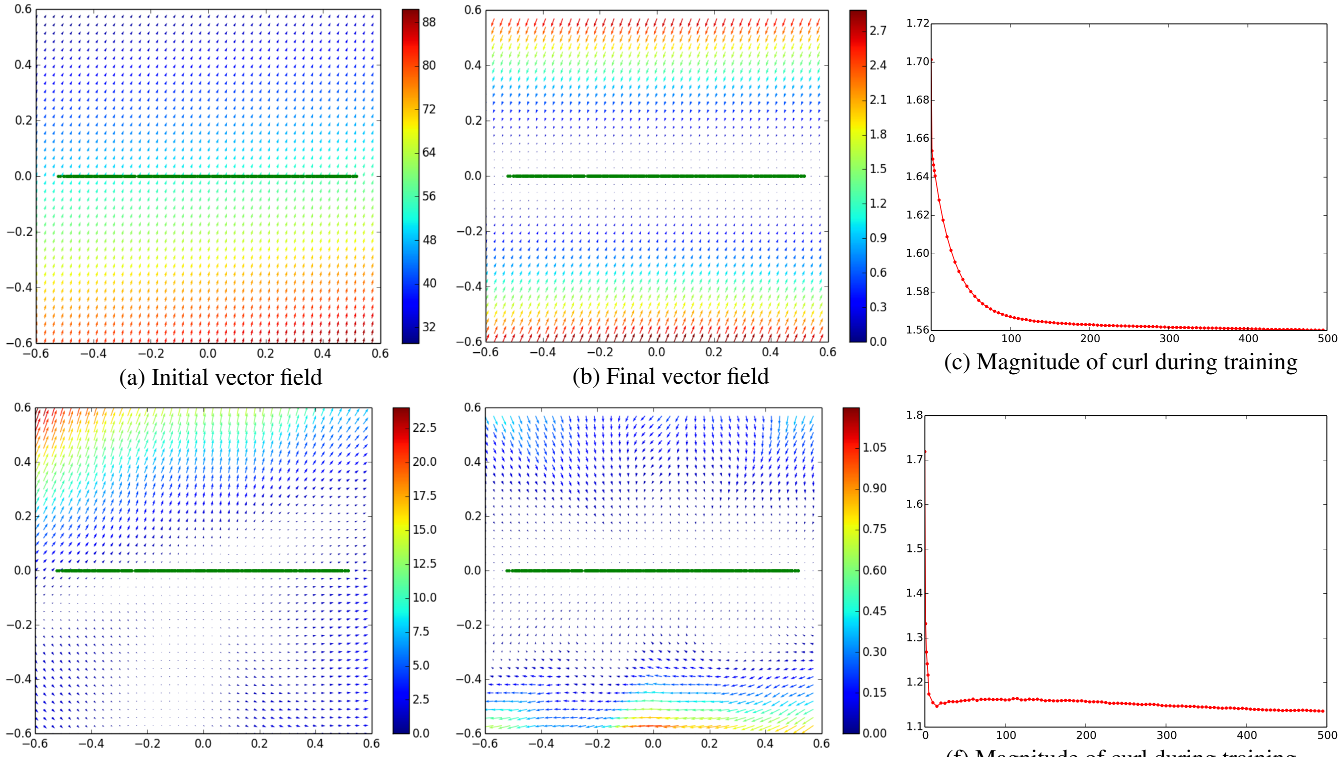
Before describing the Hodge-Helmholtz decomposition, it is worth mentioning that many communities such as fluid mechanics, physics, and mathemetics have developed various projection methods to decompose the dynamics of incompressible fluids for the simulations(Chorin 1997; 1968). As well, the Helmholtz-Hodge decomposition has be employed to various research fields such as computer graphics(Foster and Metaxas 1996; 1997), computer vision(Gao et al. 2010; Guo, Mandal, and Li 2004), and robotics (Mochizuki and Imiyya 2009). The literature review paper by (Bhatia et al. 2013) expounds much more applications of the Helmholtz-Hodge decomposition in different disciplines including fluid mechanics, physics, computer graphics, and such.

The Helmholtz-Hodge decomposition

The fundamental theorem of vector calculus, also known as Helmhotz decomposition (James 1966), states that any vector field can be expressed as the sum of an irrotational and a solenoidal field. Furthermore, extending from $\mathbb{R}^3$ to differential forms on a Riemannian manifold, the Hodge Helmholtz decomposition of any arbitrary k-form in terms of a $k-1$-form, $k+1$ form and a harmonic k-form:

$$\omega = d\alpha + \delta\beta + \gamma \tag{11}$$

where d is the exterior derivative, δ the co-differential, and $\Delta\gamma = 0$. This means that any vector field can be decomposed into scaler(symmetric), solenoidal (anti-symmetric tensor), and harmonic (rotational) vector fields and they are orthogonal to each other. Here we work with 1-forms since they correspond to vector fields. Based on Hodge decomposition theorem, any 1-form (vector field) can be orthogonally decomposed into a direct sum of a scalar, solenoidal, and harmonic components. This shows that it is always possible, in theory, to get the closest, in a least square sense, conservative vector field to a non-conservative one.



(a) Initial vector field

(b) Final vector field

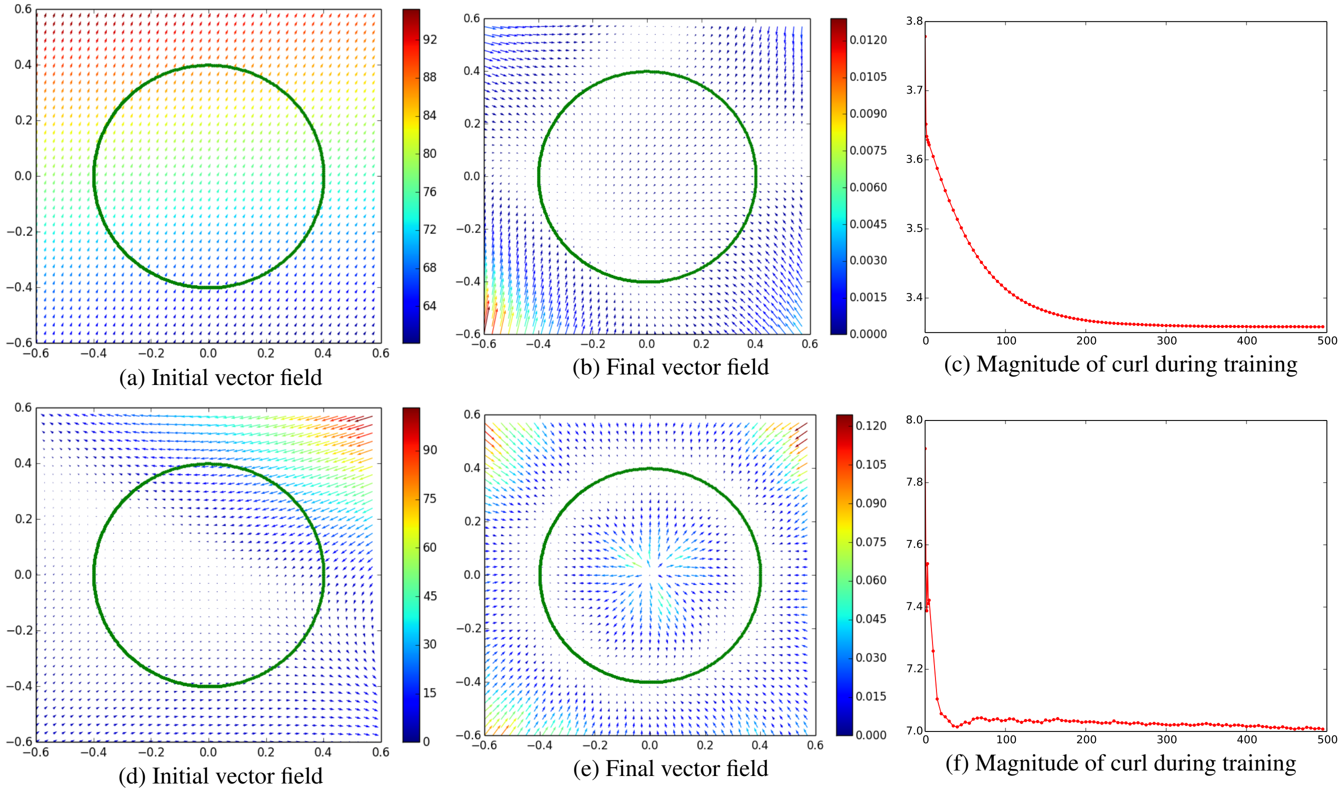
(c) Magnitude of curl during training

(d) Initial vector field

(e) Final vector field

(f) Magnitude of curl during training

Figure 4: Initial and final vector field after training untied auto-encoder on line dataset.



(a) Initial vector field

(b) Final vector field

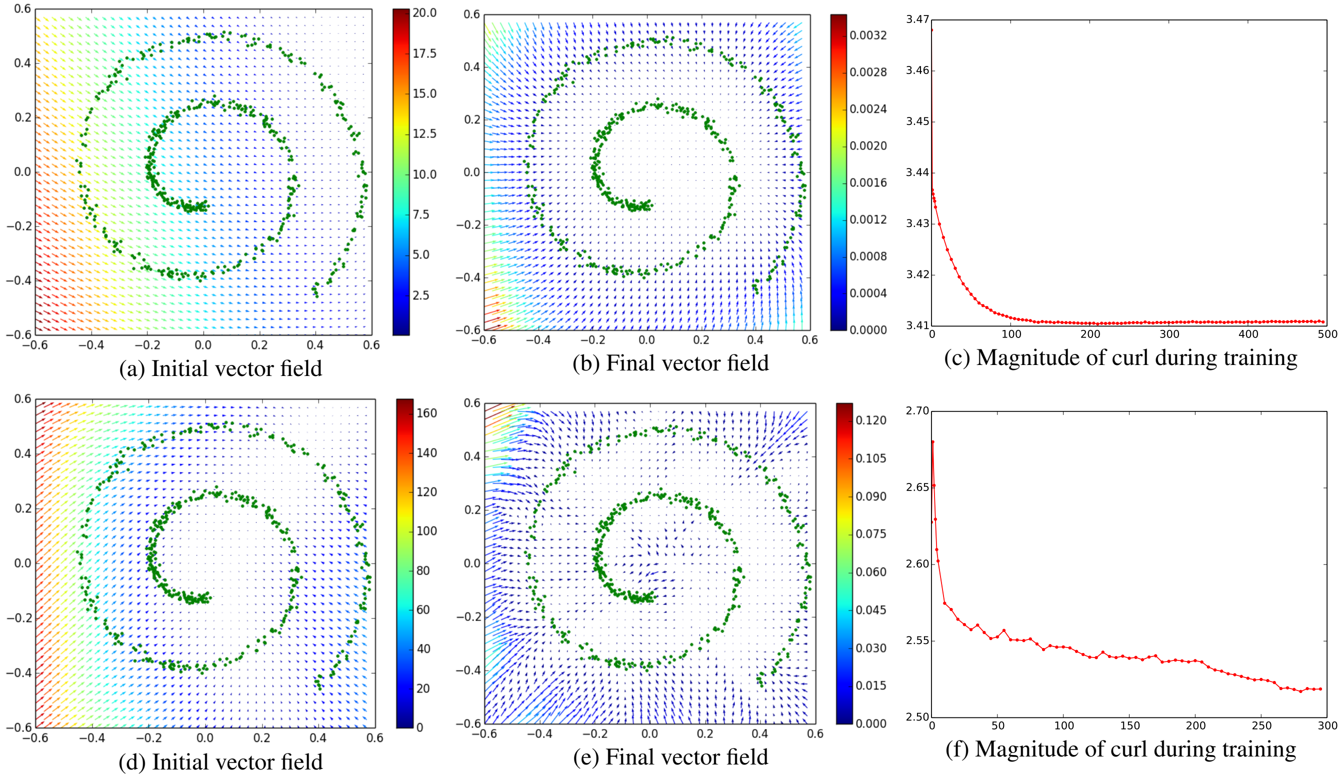
(c) Magnitude of curl during training

(d) Initial vector field

(e) Final vector field

(f) Magnitude of curl during training

Figure 5: Initial and final vector field after training untied auto-encoder on circle dataset.



(a) Initial vector field

(b) Final vector field

(c) Magnitude of curl during training

(d) Initial vector field

(e) Final vector field

(f) Magnitude of curl during training

Figure 6: Initial and final vector field after training untied auto-encoder on spiral dataset.

References

Bhatia, H.; Norgard, G.; Pascucci, V.; and Bremer, P.-T. 2013. The helmtholz-hodge decomposition-a survey. *IEEE Transactions on visualization and computer graphics* 19:1386–1404.

Chorin, A. J. 1968. Numerical solution of the naiver-stokes equations. *Mathematics of computation* 104:745–762.

Chorin, A. J. 1997. A numerical method for solving incompressible viscous flow problem. *Jounral of computations physics* 135:118–125.

Foster, N., and Metaxas, D. 1996. Realistic animation of liquids. *Graphical Models and Image Processing* 58:471–483.

Foster, N., and Metaxas, D. 1997. Modeling the motion of a hot, turbulent gas. In *Annual Conference Computer graphics and interactive techniques*, 181–188.

Gao, H.; Mandal, M. K.; Guo, G.; and Wan, J. 2010. Singular point detection using discrete hodge helmholtz decomposition in fingerprint images. In *IEEE International conference Acoustic Speech and Signal Processing*, 1094–1097.

Guo, Q.; Mandal, M. K.; and Li, M. Y. 2004. Efficient hodge-helmholtz decomposition of motion fields. *Pattern Recognition Letters* 26:493–501.

James, R. 1966. Advanced calculus. In *Advanced Calculus*. Belmont, CA: Wadsworth.

Kamyshanska, H. 2013. On autoencoder scoring. In *Proceedings of the International Conference on Machine Learning (ICML)*, 720–728.

Mochizuki, Y., and Imiyya, A. 2009. Spatial reasoning for robot navigation using the helmholtz-hodge decompoistion of ominidirectional optical flow. In *International conference Image and vision computing New Zealand*.