

# Tangent-Normal Adversarial Regularization for Semi-supervised Learning

Bing Yu<sup>1\*</sup>, Jingfeng Wu<sup>1\*</sup>, Jinwen Ma<sup>1</sup>, Zhanxing Zhu<sup>1,2,3†</sup>

<sup>1</sup>School of Mathematical Sciences, Peking University

<sup>2</sup>Center for Data Science, Peking University

<sup>3</sup> Beijing Institute of Big Data Research

{byu, pkujwf}@pku.edu.cn

jwma@math.pku.edu.cn

zhanxing.zhu@pku.edu.cn

## Abstract

Compared with standard supervised learning, the key difficulty in semi-supervised learning is how to make full use of the unlabeled data. A recently proposed method, virtual adversarial training (VAT), smartly performs adversarial training without label information to impose a local smoothness on the classifier, which is especially beneficial to semi-supervised learning. In this work, we propose **tangent-normal adversarial regularization (TNAR)** as an extension of VAT by taking the data manifold into consideration. The proposed TNAR is composed by two complementary parts, the tangent adversarial regularization (TAR) and the normal adversarial regularization (NAR). In TAR, VAT is applied along the tangent space of the data manifold, aiming to enforce local invariance of the classifier on the manifold, while in NAR, VAT is performed on the normal space orthogonal to the tangent space, intending to impose robustness on the classifier against the noise causing the observed data deviating from the underlying data manifold. Demonstrated by experiments on both artificial and practical datasets, our proposed TAR and NAR complement with each other, and jointly outperforms other state-of-the-art methods for semi-supervised learning.

## 1. Introduction

The main challenge in semi-supervised learning (SSL) is how to utilize the large amount of the unlabeled data to obtain useful information, benefiting the supervised learning on the relatively insufficient amount of labeled data. For this purpose, one of the important line of research focuses on the manifold assumption on the data distribution, i.e., the observed data is distributed on a low dimensional manifold that could be characterized using the large amount of the unlabeled data, and aims to learn a proper classifier based on

the data manifold [1, 24, 19, 11, 13]. Following this stream, we sort out three reasonable assumptions to motivate our idea for semi-supervised learning:

**The manifold assumption** The observed data  $x$  presented in high dimensional space  $\mathbb{R}^D$  is with high probability concentrated in the vicinity of some underlying manifold with much lower dimensionality [3, 18, 4, 24], denoted as  $\mathcal{M} \cong \mathbb{R}^d$ .

**The noisy observation assumption** The observed data  $x$  can be decomposed into two parts as  $x = x_0 + n$ , where  $x_0$  is exactly supported on the underlying manifold  $\mathcal{M}$  and  $n$  is some noise independent of  $x_0$  [2, 23].

**The semi-supervised learning assumption** If two points  $x_1, x_2 \in \mathcal{M}$  are close in manifold distance, then the conditional probability  $p(y|x_1)$  and  $p(y|x_2)$  are similar [1, 24, 19]. In other words, the true classifier, or the true condition distribution  $p(y|X)$  varies smoothly along the underlying manifold  $\mathcal{M}$ .

According to the three assumptions, the best classifier we aim to obtain should be 1) smooth along the data manifold; 2) robust to the off-manifold noise. Hence it is natural to formulate a loss function [1, 11] for SSL as,

$$L_{\text{ssl}} := L_{\text{supervised}} + \mathcal{R}_{\text{manifold}} + \mathcal{R}_{\text{noise}}, \quad (1)$$

where the first term in Eq. (1) is the supervised learning loss, the second term penalize the manifold smoothness of the classifier, and the third term smooths the classifier so that it is robust to noise, respectively. While the supervised learning loss  $L_{\text{supervised}}$  concerning the labeled data is standard, the key ingredient lies on how to design  $\mathcal{R}_{\text{manifold}}$  and  $\mathcal{R}_{\text{noise}}$  smartly to 1) be effective for inducing the desired smoothness on the classifier, 2) be efficient for optimization, and 3) make full use of the unlabeled data.

Existing works construct  $\mathcal{R}_{\text{manifold}}$  based on the Jacobian, for instance, *tangent propagation* [26, 11]

$$\mathcal{R}_{\text{manifold}} = \mathbb{E}_{x \sim p(x)} \sum_{v \in T_x \mathcal{M}} \|(J_x f) \cdot v\|, \quad (2)$$

\*Equal contributions.

†Corresponding author.

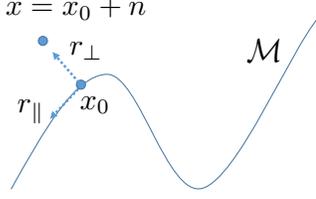


Figure 1. Illustration for tangent-normal adversarial regularization.  $x = x_0 + n$  is the observed data, where  $x_0$  is exactly supported on the underlying manifold  $\mathcal{M}$  and  $n$  is the noise independent of  $x_0$ .  $r_{\parallel}$  is the adversarial perturbation along the tangent space to induce invariance of the classifier on manifold;  $r_{\perp}$  is the adversarial perturbation along the normal space to impose robustness on the classifier against noise  $n$ .

and *manifold Laplacian norm* [1, 13, 22]

$$\begin{aligned} \mathcal{R}_{\text{manifold}} &= \int_{x \in \mathcal{M}} \|\nabla_{\mathcal{M}} f(x)\| dp(x) \\ &\approx \mathbb{E}_{z \sim p(z)} \|\nabla_{\mathcal{M}} f(g(z))\| \\ &\approx \mathbb{E}_{z \sim p(z)} \|J_z f(g(z))\|, \end{aligned} \quad (3)$$

where  $J$  is the Jacobian,  $f$  is the classifier,  $T_x \mathcal{M}$  is the tangent space of the data manifold and  $x = g(z)$  is the manifold representation of data. They regularize the manifold smoothness of the classifier under the sense of the norm of its Jacobian along the data manifold. The typical choice of  $\mathcal{R}_{\text{noise}}$  is in a corresponding form as  $\mathcal{R}_{\text{manifold}}$  except the Jacobian to penalize is with respect to the observation space other than the tangent space [1, 11].

On the other hand, inspired by adversarial training [8], virtual adversarial training (VAT) [17, 16] was proposed for SSL, not relying on the label information. Unlike the smoothness induced by  $L_p$ -norm of the Jacobian, VAT leads to the robustness of classifier by involving virtual adversarial examples, thus inducing a new local smoothness of the classifier. Empirical results [17, 21] show that VAT achieves state-of-the-art performance for SSL tasks, demonstrating the superiority of the smoothness imposed by virtual adversarial training.

Encouraged by the effectiveness of VAT, we propose to construct manifold regularizer based on VAT, instead of the  $L_p$ -norm of the Jacobian. Concretely, we propose *tangent adversarial regularization* (TAR) by performing VAT along the tangent space of the data manifold, and *normal adversarial regularization* (NAR) by applying VAT orthogonal to the tangent space of the data manifold, which are intuitively demonstrated in Figure 1. TAR enforces the local smoothness of the classifier along the underlying manifold, while NAR imposes robustness on the classifier against the noise carried in the observed data. The two terms, complementing with each other, establish our proposed approach *tangent-normal adversarial regularization* (TNAR).

To realize TNAR, we have two challenges to conquer:

1) how to estimate the underlying manifold and 2) how to efficiently perform TNAR. For the first issue, we take advantage of the generative models equipped with an extra encoder, to characterize the coordinate chart of manifold [11, 13, 22]. More specifically, in this work we choose variational autoencoder (VAE) [10] and localized GAN [22] to estimate the underlying manifold from data. For the second problem, we further extend the techniques introduced in [17] with some sophisticatedly designed auxiliary functions, implementing VAT restricted in tangent space (TAR) and normal space (NAR) efficiently. The details are elaborated in Section 3.

The remaining of the paper is organized as follows. In Section 2 we introduce VAT and two generative models as the background of TNAR. Based on that, we elaborate about the technical details of TNAR in Section 3. In Section 4 we compare TNAR with other related approaches and analyze the advantages of TNAR over VAT and other manifold-based regularization. Various experiments are conducted for demonstrating the effectiveness of TNAR in Section 5. And in Section 6 and Section 7, we discuss an existing problem about TNAR for future exploration and conclude the paper.

## 2. Background

### 2.1. Notations

The labeled and unlabeled dataset are denoted as  $\mathcal{D}_l = \{(x_l, y_l)\}$  and  $\mathcal{D}_{ul} = \{x_{ul}\}$  respectively, thus  $\mathcal{D} := \mathcal{D}_l \cup \mathcal{D}_{ul}$  is the full dataset. The output of classification model is written as  $p(y|x, \theta)$ , with  $\theta$  being the model parameters to be trained.  $\ell(\cdot, \cdot)$  represents the supervised loss function. For data example, the observed space  $\mathbb{R}^D$  and the underlying manifold is  $\mathcal{M}$ . The decoder (generator) and the encoder are denoted as  $g$  and  $h$  respectively, which form the coordinate chart of manifold together. If not stated otherwise, we always assume  $x$  and  $z$  correspond to the coordinate of the same data point in observed space  $\mathbb{R}^D$  and on manifold  $\mathcal{M}$ , i.e.,  $g(z) = x$  and  $h(x) = z$ . The tangent space of  $\mathcal{M}$  at point  $x$  is  $T_x \mathcal{M} = J_z g(\mathbb{R}^d) \cong \mathbb{R}^d$ , where  $J_z g$  is the Jacobian of  $g$  at point  $z$ .  $T_x \mathcal{M}$  is also the span of the columns of  $J_z g$ . We use  $J$  to represent the Jacobian when there is no ambiguity.

The perturbation in the observed space  $\mathbb{R}^D$  is denoted as  $r \in \mathbb{R}^D$ , while the perturbation on the manifold representation is denoted as  $\eta \in \mathbb{R}^d$ . Hence the perturbation on manifold is  $g(z + \eta) - g(z)$ . When the perturbation  $\eta$  is small enough for the holding of the first order Taylor's expansion, the perturbation on manifold is approximately equal to the perturbation on its tangent space,  $g(z + \eta) - g(z) \approx J \cdot \eta \in T_x \mathcal{M}$ . Therefore we say a perturbation  $r \in \mathbb{R}^D$  is actually on manifold, if there is a perturbation  $\eta \in \mathbb{R}^d$ , such that  $r = J \cdot \eta$ .

## 2.2. Virtual adversarial training

VAT [17] is an effective regularization method for SSL. The virtual adversarial loss introduced in VAT is defined by the robustness of the classifier against local perturbation in the input space  $\mathbb{R}^D$ . Hence VAT imposes a kind of smoothness condition on the classifier. Mathematically, the virtual adversarial loss in VAT for SSL is

$$L(\mathcal{D}_l, \mathcal{D}_{ul}, \theta) := \mathbb{E}_{(x_l, y_l) \in \mathcal{D}_l} \ell(y_l, p(y|x_l, \theta)) + \alpha \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{vat}}(x, \theta). \quad (4)$$

The VAT regularization  $\mathcal{R}_{\text{vat}}$  is defined as

$$\mathcal{R}_{\text{vat}}(x; \theta) := \max_{\|r\|_2 \leq \epsilon} \text{dist}(p(y|x, \theta), p(y|x+r, \theta)), \quad (5)$$

where  $\text{dist}(\cdot, \cdot)$  is some distribution distance measure and  $\epsilon$  controls the magnitude of the adversarial example. For simplicity, define

$$F(x, r, \theta) := \text{dist}(p(y|x, \theta), p(y|x+r, \theta)). \quad (6)$$

Then  $\mathcal{R}_{\text{vat}} = \max_{\|r\|_2 \leq \epsilon} F(x, r, \theta)$ . And the so called virtual adversarial example is  $r^* := \text{argmax}_{\|r\|_2 \leq \epsilon} F(x, r, \theta)$ . Once we have  $r^*$ , the VAT loss can be optimized with the objective as  $L(\mathcal{D}_l, \mathcal{D}_{ul}, \theta) = \mathbb{E}_{(x_l, y_l) \in \mathcal{D}_l} \ell(y_l, p(y|x_l, \theta)) + \alpha \mathbb{E}_{x \in \mathcal{D}} F(x, r^*, \theta)$ .

To obtain the virtual adversarial example  $r^*$ , [17] suggested to apply second order Taylor's expansion to  $F(x, r, \theta)$  around  $r = 0$  as

$$F(x, r, \theta) \approx \frac{1}{2} r^T H r, \quad (7)$$

where  $H := \nabla_r^2 F(x, r, \theta)|_{r=0}$  denotes the Hessian of  $F$  with respect to  $r$ . The vanishing of the first two terms in Taylor's expansion occurs because that  $\text{dist}(\cdot, \cdot)$  is a distance measure with minimum zero and  $r = 0$  is the corresponding optimal value, indicating that at  $r = 0$ , both the value and the gradient of  $F(x, r, \theta)$  are zero. Therefore for small enough  $\epsilon$ ,  $r^* \approx \text{argmax}_{\|r\|_2 \leq \epsilon} \frac{1}{2} r^T H r$ , which is an eigenvalue problem and the direction of  $r^*$  can be solved by power iteration.

## 2.3. Generative models for data manifold

We take advantage of generative model with both encoder  $h$  and decoder  $g$  to estimate the underlying data manifold  $\mathcal{M}$  and its tangent space  $T_x \mathcal{M}$ . As assumed by previous works [11, 13], perfect generative models with both decoder and encoder can describe the data manifold, where the decoder  $g(z)$  and the encoder  $h(x)$  together serve as the coordinate chart of manifold  $\mathcal{M}$ . Note that the encoder is indispensable for it helps to identify the manifold coordinate  $z = h(x)$  for point  $x \in \mathcal{M}$ . With the trained generative model, the tangent space is given by  $T_x \mathcal{M} = J_z g(\mathbb{R}^d)$ , or the span of the columns of  $J = J_z g$ .

In this work, we adopt VAE [10] and localized GAN [22] to learn the targeted underlying data manifold  $\mathcal{M}$  as summarized below.

VAE [10] is a well known generative model consisting of both encoder and decoder. The training of VAE is by optimizing the variational lower bound of log likelihood,

$$\log p(x, \theta) \geq \mathbb{E}_{z \sim q(z|x, \theta)} [\log p(x|z, \theta)] - KL(q(z|x, \theta) \| p(z)). \quad (8)$$

Here  $p(z)$  is the prior of hidden variable  $z$ , and  $q(z|x, \theta)$ ,  $p(x|z, \theta)$  models the encoder and decoder in VAE, respectively. The derivation of the lower bound with respect to  $\theta$  is well defined thanks to the reparameterization trick, thus it could be optimized by gradient based method. The lower bound could also be interpreted as a reconstruction term plus a regularization term [10]. With a trained VAE, the encoder and decoder are given as  $h(x) = \text{argmax}_z q(z|x)$  and  $g(z) = \text{argmax}_x q(x|z)$  accordingly.

**Localized GAN** Localized GAN [22] suggests to use a localized generator  $G(x, z)$  to replace the global generator  $g(z)$  in vanilla GAN [7]. The key difference between localized GAN and previous generative model for manifold is that, localized GAN learns a distinguishing local coordinate chart for each point  $x \in \mathcal{M}$ , which is given by  $G(x, z)$ , rather than one global coordinate chart. To model the local coordinate chart in data manifold, localized GAN requires the localized generator to satisfy two more regularity conditions:

**locality**  $G(x, 0) = x$ , so that  $G(x, z)$  is localized around  $x$ ;

**orthogonality**  $\left(\frac{\partial G(x, z)}{\partial z}\right)^T \frac{\partial G(x, z)}{\partial z} = I$ , to ensure  $G(x, z)$  is non-degenerated.

The two conditions are achieved by the following penalty during training of localized GAN:

$$\mathcal{R}_{\text{localized GAN}} := \mu_1 \|G(x, 0) - x\|^2 + \mu_2 \left\| \left(\frac{\partial G(x, z)}{\partial z}\right)^T \frac{\partial G(x, z)}{\partial z} - I \right\|^2. \quad (9)$$

Since  $G(x, z)$  defines a local coordinate chart for each  $x$  separately, in which the latent encode of  $x$  is  $z = 0$ , there is no need for the extra encoder to provide the manifold representation of  $x$ .

## 3. Method

In this section we elaborate our proposed *tangent-normal adversarial regularization* (TNAR) strategy. The TNAR

loss to be minimized for SSL is

$$\begin{aligned} L(D_l, D_{ul}, \theta) := & \mathbb{E}_{(x_l, y_l) \in \mathcal{D}_l} \ell(y_l, p(y|x_l, \theta)) \\ & + \alpha_1 \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{tangent}}(x, \theta) \\ & + \alpha_2 \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{normal}}(x, \theta). \end{aligned} \quad (10)$$

The first term in Eq. (10) is a common used supervised loss, e.g., negative cross entropy.  $\mathcal{R}_{\text{tangent}}$  and  $\mathcal{R}_{\text{normal}}$  is the so called *tangent adversarial regularization* (TAR) and *normal adversarial regularization* (NAR) accordingly, jointly forming the proposed TNAR. In the following section, we assume that we already have a well trained generative model for the underlying data manifold  $\mathcal{M}$ , with encoder  $h$  and decoder  $g$ , which can be obtained as described in Section 2.3.

### 3.1. Tangent adversarial regularization

Vanilla VAT penalizes the variety of the classifier against local perturbation in the input space  $\mathbb{R}^D$  [17], which might overly regularize the classifier, since the semi-supervised learning assumption only indicates that the true conditional distribution varies smoothly along the underlying manifold  $\mathcal{M}$ , but not the whole input space  $\mathbb{R}^D$  [1, 24, 19]. To avoid this shortcoming of vanilla VAT, we propose the tangent adversarial regularization (TAR), which restricts virtual adversarial training to the tangent space of the underlying manifold  $T_x \mathcal{M}$ , to enforce manifold invariance property of the classifier.

$$\mathcal{R}_{\text{tangent}}(x; \theta) := \max_{\|r\|_2 \leq \epsilon, r \in T_x \mathcal{M} = J_z g(\mathbb{R}^d)} F(x, r, \theta), \quad (11)$$

where  $F(x, r, \theta)$  is defined as in Eq. (6). To optimize Eq. (11), we first apply Taylor's expansion to  $F(x, r, \theta)$  so that

$$\mathcal{R}_{\text{tangent}}(x; \theta) \approx \max_{\|r\|_2 \leq \epsilon, r \in T_x \mathcal{M} = J_z g(\mathbb{R}^d)} \frac{1}{2} r^T H r, \quad (12)$$

where the notations and the derivation are as in Eq. (7). We further reformulate  $\mathcal{R}_{\text{tangent}}$  as

$$\begin{aligned} & \underset{r \in \mathbb{R}^D}{\text{maximize}} \quad \frac{1}{2} r^T H r, \\ & \text{s.t.} \quad \|r\|_2 \leq \epsilon, \quad r = J\eta, \quad \eta \in \mathbb{R}^d. \\ & (J := J_z g \in \mathbb{R}^{D \times d}, \quad H \in \mathbb{R}^{D \times D}) \end{aligned} \quad (13)$$

Or equivalently,

$$\underset{\eta \in \mathbb{R}^d}{\text{maximize}} \quad \frac{1}{2} \eta^T J^T H J \eta, \quad \text{s.t.} \quad \eta^T J^T J \eta \leq \epsilon^2. \quad (14)$$

This is a classic generalized eigenvalue problem, the optimal solution  $\eta^*$  of which could be obtained by power iteration and conjugate gradient (and scaling). The iteration

framework is as

$$\begin{aligned} v & \leftarrow J^T H J \eta; \\ \mu & \leftarrow (J^T J)^{-1} v; \\ \eta & \leftarrow \frac{\mu}{\|\mu\|_2}. \end{aligned} \quad (15)$$

Now we elaborate the detailed implementation of each step in Eq. (15).

**Computing  $J^T H J \eta$ .** Note that  $z = h(x)$ ,  $x = g(z)$ . Define

$$r(\eta) := g(z + \eta) - g(z). \quad (16)$$

For

$$F(x, r(\eta), \theta) = \text{dist}(p(y|x, \theta) \| p(y|x + r(\eta), \theta)), \quad (17)$$

we have

$$\begin{aligned} \nabla_\eta^2 F(x, r(\eta), \theta) & = (J_{z+\eta} g)^T \nabla_r^2 F(x, r(\eta), \theta) (J_{z+\eta} g) \\ & \quad + \nabla_\eta^2 g(z + \eta) \cdot \nabla_r F(x, r(\eta), \theta). \end{aligned} \quad (18)$$

While on the other hand, since  $\text{dist}(\cdot, \cdot)$  is some distance measure with minimum zero and  $r(0) = 0$  is the corresponding optimal value, we have

$$F(x, r(0), \theta) = 0, \quad \nabla_r F(x, r(0), \theta) = 0. \quad (19)$$

Therefore,

$$\nabla_\eta^2 F(x, r(0), \theta) = (J_z g)^T \nabla_r^2 F(x, r(0), \theta) J_z g = J^T H J. \quad (20)$$

Thus the targeted matrix vector product could be efficiently computed as

$$J^T H J \eta = \nabla_\eta^2 F(x, r(0), \theta) \cdot \eta = \nabla_\eta (\nabla_\eta F(x, r(0), \theta) \cdot \eta). \quad (21)$$

Note that  $\nabla_\eta F(x, r(0), \theta) \cdot \eta$  is a scalar, hence the gradient of which could be obtained by back propagating the network for once. And it only costs twice back propagating for the computation of  $J^T H J \eta$ .

**Solving  $J^T J \mu = v$ .** Similarly, define

$$K(\eta) := (g(z + \eta) - g(z))^T (g(z + \eta) - g(z)). \quad (22)$$

We have

$$\nabla_\eta^2 K(\eta) = (J_{z+\eta} g)^T J_{z+\eta} g + \nabla_\eta^2 g(z + \eta) \cdot K(\eta). \quad (23)$$

Since  $K(0) = 0$ , we have

$$\nabla_\eta^2 K(0) = (J_z g)^T J_z g = J^T J. \quad (24)$$

Thus the matrix vector product  $J^T J \mu$  could be evaluated similarly as

$$J^T J \mu = \nabla_\eta (\nabla_\eta K(0) \cdot \mu). \quad (25)$$

The extra cost for evaluating  $J^T J \mu$  is still back propagating the network for twice. Due to  $J^T J$  being positive definite ( $g$  is non-degenerated), we can apply several steps of conjugate gradient to solve  $J^T J \mu = v$  efficiently.

By iterating Eq. (15), we obtain the optimal solution  $\eta_{\parallel}$  of Eq. (14). The desired optimal solution is then  $r_{\parallel} = \epsilon J \eta_{\parallel} / \|J \eta_{\parallel}\|$ , hence  $\mathcal{R}_{\text{tangent}}(x; \theta) = F(x, r_{\parallel}, \theta)$ , which could be optimized by popular gradient optimizers.

### 3.2. Normal adversarial regularization

Motivated by the noisy observation assumption indicating that the observed data contains noise driving them off the underlying manifold, we further come up with the normal adversarial regularization (NAR) to enforce the robustness of the classifier against such noise, by performing virtual adversarial training in the normal space. The mathematical description is

$$\begin{aligned} \mathcal{R}_{\text{normal}}(x; \theta) &:= \max_{\|r\|_2 \leq \epsilon, r \perp T_x \mathcal{M}} F(x, r, \theta) \\ &\approx \max_{\|r\|_2 \leq \epsilon, r \perp T_x \mathcal{M}} \frac{1}{2} r^T H r. \end{aligned} \quad (26)$$

Note that  $T_x \mathcal{M}$  is spanned by the columns of  $J = J_z g$ , thus  $r \perp T_x \mathcal{M} \Leftrightarrow J^T \cdot r = 0$ . Therefore we could reformulate Eq. (26) as

$$\begin{aligned} \text{maximize}_{r \in \mathbb{R}^D} \quad & \frac{1}{2} r^T H r, \\ \text{s.t.} \quad & \|r\|_2 \leq \epsilon, \quad J^T \cdot r = 0. \end{aligned} \quad (27)$$

However, Eq. (27) is not easy to optimize since  $J^T \cdot r$  cannot be efficiently computed. To overcome this, instead of requiring  $r$  being orthogonal to the whole tangent space  $T_x \mathcal{M}$ , we take a step back to demand  $r$  being orthogonal to only one specific tangent direction, i.e., the tangent space adversarial perturbation  $r_{\parallel}$ . Thus the constraint  $J^T \cdot r = 0$  is relaxed to  $(r_{\parallel})^T \cdot r = 0$ . And we further replace the constraint by a regularization term,

$$\begin{aligned} \text{maximize}_{r \in \mathbb{R}^D} \quad & \frac{1}{2} r^T H r - \lambda r^T (r_{\parallel} r_{\parallel}^T) r, \\ \text{s.t.} \quad & \|r\|_2 \leq \epsilon, \end{aligned} \quad (28)$$

where  $\lambda$  is a hyperparameter introduced to control the orthogonality of  $r$ .

Since Eq. (28) is again an eigenvalue problem, and we can apply power iteration to solve it. Note that a small identity matrix  $\lambda \|r_{\parallel}\| I$  is needed to be added to keep  $\frac{1}{2} H - \lambda r_{\parallel} r_{\parallel}^T + \lambda \|r_{\parallel}\| I$  semi-positive definite, which does not change the optimal solution of the eigenvalue problem. The power iteration is as

$$r \leftarrow \frac{1}{2} H r - \lambda (r_{\parallel})^T r_{\parallel} r + \lambda \|r_{\parallel}\| r. \quad (29)$$

And the evaluation of  $H r$  is by

$$H r = \nabla_r (\nabla_r F(x, 0, \theta) \cdot r), \quad (30)$$

which could be computed efficiently. After finding the optimal solution of Eq. (28) as  $r_{\perp}$ , the NAR becomes  $\mathcal{R}_{\text{normal}}(x, \theta) = F(x, r_{\perp}, \theta)$ .

Finally, as suggested in [17], we add *entropy regularization* to our loss function. It ensures neural networks to output more determinate predictions and has implicit benefits for performing virtual adversarial training.

$$\mathcal{R}_{\text{entropy}}(x, \theta) := - \sum_y p(y|x, \theta) \log p(y|x, \theta). \quad (31)$$

Our final loss for SSL is

$$\begin{aligned} L(D_l, D_{ul}, \theta) &:= \mathbb{E}_{(x_l, y_l) \in \mathcal{D}_l} \ell(y_l, p(y|x_l, \theta)) \\ &+ \alpha_1 \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{tangent}}(x, \theta) \\ &+ \alpha_2 \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{normal}}(x, \theta) \\ &+ \alpha_3 \mathbb{E}_{x \in \mathcal{D}} \mathcal{R}_{\text{entropy}}(x, \theta). \end{aligned} \quad (32)$$

## 4. Comparison to other methods

**Virtual adversarial training** Our proposed TNAR serves as an extension of VAT, by taking the information of data manifold into consideration. VAT equally penalizes the smoothness along each dimension of the whole observation space, not discriminating different directions. In contrast, TNAR enforces the smoothness of the classifier along the manifold and orthogonal to the manifold separately. This separate treatment along the two directions allows TNAR to impose different scales of smoothness along the tangent space and the normal space of the data manifold, which is particularly crucial for inducing desired regularization effect. To illustrate this, considering an image sample, its Euclidean neighborhood in the input space could contain many inter-class samples, besides intra-class ones, as demonstrated in Figure 2. Thus the output of the ideal classifier must vary significantly inside such Euclidean neighborhood to correctly classify the contained samples, which makes it essentially improper for VAT to enforce that the classifier does not change much inside this Euclidean ball. A more reasonable treatment is to adopt the manifold assumption and impose different scales of smoothness of the classifier along the manifold and its orthogonal direction, as TNAR has done.

**Jacobian based manifold regularization** As explained in Eq. (2) and Eq. (3), tangent propagation [26, 11] and manifold Laplacian norm [1, 13, 22] are also popular methods for realizing manifold regularization for SSL. However, our TNAR is the first to use VAT constructing manifold regularization. The difference between TNAR and Jacobian norm based manifold regularization is two folds.

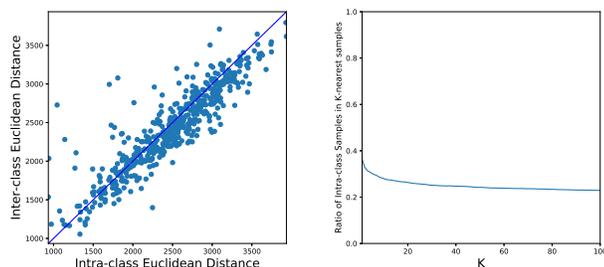


Figure 2. Left: the smallest intra-class distance vs. the smallest inter-class distance for CIFAR-10 dataset. X-axis: the smallest Euclidean distance to the other examples of the same class. Y-axis: the smallest Euclidean distance to the other examples of the different class. We only plot such coordinate for 500 examples. Right: the ratio of intra-class example among its  $K$ -nearest neighborhood for CIFAR-10 dataset. The ratio is averaged over 500 examples. From the figures we clearly see that for most examples, 1) the smallest inter-class distance is shorter or at least about the same scale as the least intra-class distance, and 2) its  $K$ -nearest neighborhood contains more inter-class examples than intra-class examples.

Firstly, they lead to different manifold smoothness conditions on the classifier. Tangent propagation and manifold Laplacian norm smooth the classifier by regularizing its norm of the manifold Jacobian. TNAR, on the other hand, smooths the classifier through penalizing the virtual adversarial loss defined by the distance of an example with its tangent directional virtual adversarial example. This involves the second order information of the virtual adversarial loss along the manifold. Theoretically, it is not easy to say that one smoothness is superior to the other. Nonetheless, empirical experiments on multiple datasets (Section 5) suggest that our proposed TNAR achieves better performance on SSL. We leave the theoretical analysis as future work.

Secondly, as shown in Eq. (2) and Eq. (3), all the existing Jacobian based manifold regularization requires evaluating the Jacobian of either the classifier, or the generator as manifold coordinate chart, which is prohibitively feasible for modern high-dimensional datasets given large neural networks. Alternatively, some works suggested stochastically evaluating these Jacobian based regularization terms. Kumar et al. [11] proposed to randomly preserve several columns of  $J_z g$  as the approximation of the tangent space  $T_x \mathcal{M}$ , and Lecouat et al. [13] applied the norm of several directional gradients to approximate the norm of the Jacobian. However, such stochastic strategies, unfortunately with high variance, could cause implicit side effects on the manifold smoothness of the classifier. Compared with them, the computational cost of our proposed TNAR does not rely on the dimensionality of the datasets, since performing VAT only requires several times of power iteration (typically once), and TNAR adds constant extra times of back or

forward propagation to VAT. This advantage makes TNAR a potentially better manifold regularization method for modern semi-supervised learning tasks.

**Other approaches for SSL** There is also a wide class of SSL framework based on GAN [25, 20, 6, 5]. Most of them modify the discriminator to include a classifier, by splitting the real class of original discriminator into  $K$  subclasses, where  $K$  is the number of classes of labeled data. The features extracted for distinguishing the example being real or fake, which can be viewed as a kind of coarse label, have implicit benefits for supervised classification task. Though in TNAR, GAN with encoder could be adopted as a method to identify the underlying manifold, these two kinds of approaches are motivated from different perspectives. TNAR focuses on the manifold regularization other than the feature sharing as in the GAN frameworks for SSL.

Besides above, there are also other strategies for SSL, e.g., Tripple GAN [14], Mean Teacher [27],  $\Pi$  model [12], CCLP [9] etc. We leave the comparison of the performance with TNAR in Section 5.

## 5. Experiments

To demonstrate the advantages of our proposed TNAR for SSL, we conduct a series of experiments on both artificial and real datasets. The tested TNAR based methods for SSL include: (1) **TNAR-VAE**: TNAR with the underlying manifold estimated by VAE; (2) **TNAR-LGAN**: TNAR with the underlying manifold estimated by localized GAN; (3) **TNAR-Manifold**: TNAR with oracle underlying manifold for the observed data, only used for artificial dataset; (4) **TNAR-AE**: TNAR with the underlying manifold estimated roughly by autoencoder, only used for artificial dataset; (5) **TAR**: tangent adversarial regularization for ablation study; (6) **NAR**: normal adversarial regularization for ablation study. If not stated otherwise, all the above methods contain entropy regularization term.

### 5.1. Two-rings artificial dataset

We first introduce experiments on a two-rings artificial dataset to show the effectiveness of TNAR intuitively. In this experiments, there is 3,000 unlabeled data (gray dots) and 6 labeled data (blue dots), 3 for each class. The detailed construction could be found in Supplementary Materials.

The performance of each compared methods is shown in Table 1, and the corresponding classification boundary is demonstrated in Figure 3. The TNAR under true underlying manifold (TNAR-Manifold) perfectly classifies the two-rings dataset with merely 6 labeled data, while the other methods fail to predict the correct decision boundary. The failure of VAT supports our claims of its shortcut in Section 4. Even with the underlying manifold roughly approxi-

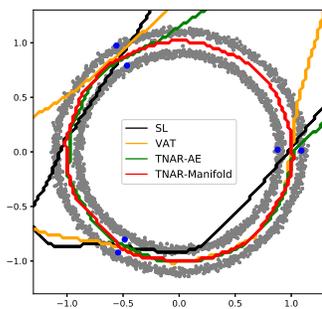


Figure 3. The decision boundaries of compared methods on two-rings artificial dataset. Gray dots distributed on two rings: the unlabeled data. Blue dots (3 in each ring): the labeled data. Colored curves: the decision boundaries found by compared methods.

Table 1. Classification errors (%) of compared methods on two-ring artificial dataset. We test with and without entropy regularization in each method and report the best one. In VAT and TNAR-AE, without entropy regularization is better; For TNAR-Manifold, adding entropy regularization is better.

Model	Error (%)
Labeled data only	32.95
VAT	23.80
TNAR-AE	12.45
TNAR-Manifold	9.90
TNAR-Manifold (ent)	0

mated by an autoencoder, our approach (TNAR-AE) outperforms VAT in this artificial dataset. However, the performance of TNAR-AE is worse than TNAR-Manifold, indicating that the effectiveness of TNAR relies on the quality of estimating the underlying manifold.

## 5.2. FashionMNIST

We also conduct experiments on FashionMNIST dataset<sup>1</sup>. There are three sets of experiments with the number of labeled data being 100, 200 and 1,000, respectively. The details about the networks are in Supplementary Materials.

The corresponding results are shown in Table 2, from which we observe at least two phenomena. The first is that our proposed TNAR methods (TNAR-VAE, TNAR-LGAN) achieve lower classification errors than VAT in all circumstances with different number of labeled data. The second is that the performance of our method depends on the estimation of the underlying manifold of the observed data. In this case, TNAR-VAE brings larger improvement than TNAR-LGAN, since VAE produces better diverse examples

<sup>1</sup><https://github.com/zalandoresearch/fashion-mnist>

according to our observation.

## 5.3. CIFAR-10 and SVHN

There are two classes of experiments for demonstrating the effectiveness of TNAR in SSL, SVHN with 1,000 labeled data, and CIFAR-10 with 4,000 labeled data. The experiment setups are identical with [17]. We test two kinds of convolutional neural networks as classifier (denoted as "small" and "large") as in [17]. We test both VAE and Localized GAN as the underlying data manifold. More detailed experimental settings are included in Supplementary Materials. We test the performance of TNAR with or without data augmentation, with the identical augmentation strategy used in [17]. Note that when perform TNAR with data augmentation, the corresponding data manifold should also be trained with data augmentation. It is worth to remark that VAT [17] and VAT + SNTG [15] adopts ZCA as pre-processing on CIFAR-10 experiments, while we do not use this trick implementing TNAR experiments.

In Table 3 we report the experiments results on SVHN and CIFAR-10, without data augmentation. And in Table 4 the results on SVHN and CIFAR-10 with data augmentation are presented. The comparison demonstrates that our proposed TNAR outperforms all the other state-of-the-art SSL methods as far as we known on both SVHN and CIFAR-10, with or without data augmentation. Especially, compared with VAT or manifold regularization like Improved GAN + JacobRegu + tangent [11] or Improved GAN + ManiReg [13], TNAR brings an evident improvements to them, as our analysis in Section 4 has suggested. Similar to experiments on FashionMNIST datasets, we observe that for TNAR, the underlying manifold identified by VAE benefits more than the manifold identified by Localized GAN. We attribute this phenomenon to the relatively lacking of diversity of the images generated by Localized GAN.

## 5.4. Ablation study

We conduct ablation study on FashionMNIST, SVHN and CIFAR-10 datasets to demonstrate that both of the two regularization terms in TNAR are crucial for SSL. The results are reported in Table 2 and the last two lines in Table 3. Removing either tangent adversarial regularization or normal adversarial regularization will harm the final performance, since they fail to enforce the manifold invariance or the robustness against the off-manifold noise. In together, the proposed TNAR achieves the best performance.

Furthermore, the adversarial perturbations and adversarial examples from FashionMNIST and CIFAR-10 are shown in Figure 4. We can easily observe that the tangent adversarial perturbation focuses on the edges of foreground objects, while the normal space perturbation mostly appears as certain noise over the whole image. This is consistent with our understanding on the role of perturbation along the

Table 2. Classification errors (%) of compared methods on FashionMNIST dataset.

Method	100 labels	200 labels	1000 labels
VAT	27.69	20.85	14.51
TNAR/TAR/NAR-LGAN	23.65/24.87/28.73	18.32/19.16/24.49	13.52/14.09/15.94
TNAR/TAR/NAR-VAE	<b>23.35/26.45/27.83</b>	<b>17.23/20.53/24.81</b>	<b>12.86/14.02/15.44</b>

Table 3. Classification errors (%) of compared methods on SVHN and CIFAR-10 datasets without data augmentation.

Method	SVHN	CIFAR-10
	1,000 labels	4,000 labels
VAT (small) [17]	6.83 ± 0.24	14.87 ± 0.13
VAT (large) [17]	4.28 ± 0.10	13.15 ± 0.21
VAT + SNTG [15]	4.02 ± 0.20	12.49 ± 0.36
II model [12]	5.43 ± 0.25	16.55 ± 0.29
Mean Teacher [27]	5.21 ± 0.21	17.74 ± 0.30
CCLP [9]	5.69 ± 0.28	18.57 ± 0.41
ALI [6]	7.41 ± 0.65	17.99 ± 1.62
Improved GAN [25]	8.11 ± 1.3	18.63 ± 2.32
Tripple GAN [14]	5.77 ± 0.17	16.99 ± 0.36
Bad GAN [5]	4.25 ± 0.03	14.41 ± 0.30
LGAN [22]	4.73 ± 0.16	14.23 ± 0.27
Improved GAN + JacobRegu + tangent [11]	4.39 ± 1.20	16.20 ± 1.60
Improved GAN + ManiReg [13]	4.51 ± 0.22	14.45 ± 0.21
TNAR-LGAN (small)	4.25 ± 0.09	12.97 ± 0.31
TNAR-LGAN (large)	4.03 ± 0.13	12.76 ± 0.04
TNAR-VAE (small)	3.99 ± 0.08	12.39 ± 0.11
TNAR-VAE (large)	<b>3.80 ± 0.12</b>	<b>12.06 ± 0.35</b>
TAR-VAE (large)	5.62 ± 0.19	13.87 ± 0.32
NAR-VAE (large)	4.05 ± 0.04	15.91 ± 0.09

Table 4. Classification errors (%) of compared methods on SVHN and CIFAR-10 datasets with data augmentation.

Method	SVHN	CIFAR-10
	1,000 labels	4,000 labels
VAT (large) [17]	3.86 ± 0.11	10.55 ± 0.05
VAT + SNTG [15]	3.83 ± 0.22	9.89 ± 0.34
II model [12]	4.82 ± 0.17	12.36 ± 0.31
Temporal ensembling [12]	4.42 ± 0.16	12.16 ± 0.24
Mean Teacher [27]	3.95 ± 0.19	12.31 ± 0.28
LGAN [22]	-	9.77 ± 0.13
TNAR-VAE (large)	<b>3.74 ± 0.04</b>	<b>8.85 ± 0.03</b>

two directions that capture the different aspects of smoothness.

## 6. Discussion

As shown in our experiments, the data manifold is crucial for the improvement of our proposed TNAR. Though TNAR seems to work with a wide range of manifold coordinate chart, e.g., VAE and Localized GAN, it is still not clear which kind of manifold benefits most for TNAR. Dai

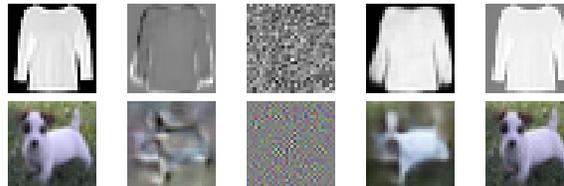


Figure 4. The perturbations and adversarial examples in the tangent space and the normal space. Note that the perturbations is actually too small to distinguish easily, thus we show the scaled perturbations. First row: FashionMNIST dataset; Second row: CIFAR-10 dataset. From left to right: original example, tangent adversarial perturbation, normal adversarial perturbation, tangent adversarial example, normal adversarial example.

at.el [5] suggested that a bad generator works better for GAN based framework for semi-supervised learning. Our experiments agree with this argument to some extent. Localized GAN could produce detailed images than VAE, but the latter cooperates better with TNAR in all our experiments. At current stage, we conjecture that a more diverse generator helps more for TNAR, since diversity on generator enables TNAR to explore more different directions along the data manifold. The throughout analysis is left for future work.

## 7. Conclusion

We present the tangent-normal adversarial regularization for semi-supervised learning, a novel regularization strategy based on virtual adversarial training and manifold regularization. TNAR is composed of regularization on the tangent and normal space separately. The tangent adversarial regularization enforces manifold invariance of the classifier, while the normal adversarial regularization imposes robustness of the classifier against the noise contained in the observed data. Experiments on synthetic and real datasets demonstrate that our approach outperforms other state-of-the-art methods for semi-supervised learning.

## Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No.61806009), Beijing Natural Science Foundation (Grant No. 4184090), and Intelligent Manufacturing Action Plan of Industrial Solid Foundation Program (Grant No. JCKY2018204C004).

## References

- [1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.
- [2] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.
- [3] L. Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.
- [4] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.
- [5] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in Neural Information Processing Systems*, pages 6510–6520, 2017.
- [6] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [8] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [9] K. Kamnitsas, D. C. Castro, L. L. Folgoc, I. Walker, R. Tanno, D. Rueckert, B. Glocker, A. Criminisi, and A. Nori. Semi-supervised learning via compact latent space clustering. *arXiv preprint arXiv:1806.02679*, 2018.
- [10] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [11] A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems*, pages 5540–5550, 2017.
- [12] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.
- [13] B. Lecouat, C.-S. Foo, H. Zenati, and V. R. Chandrasekhar. Semi-supervised learning with gans: Revisiting manifold regularization. *arXiv preprint arXiv:1805.08957*, 2018.
- [14] C. Li, K. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. *arXiv preprint arXiv:1703.02291*, 2017.
- [15] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. *arXiv preprint arXiv:1711.00258*, 2017.
- [16] T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.
- [17] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017.
- [18] H. Narayanan and S. Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, pages 1786–1794, 2010.
- [19] P. Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *The Journal of Machine Learning Research*, 14(1):1229–1250, 2013.
- [20] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.
- [21] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018.
- [22] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua. Global versus localized generative adversarial nets. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.
- [24] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302, 2011.
- [25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.
- [26] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri. Transformation invariance in pattern recognition: tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 1998.
- [27] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.