On Data-Augmentation and Consistency-based Semi-Supervised Learning

Atin Ghosh, Alexandre H. Thiery

Department of Statistics and Applied Probability National University of Singapore

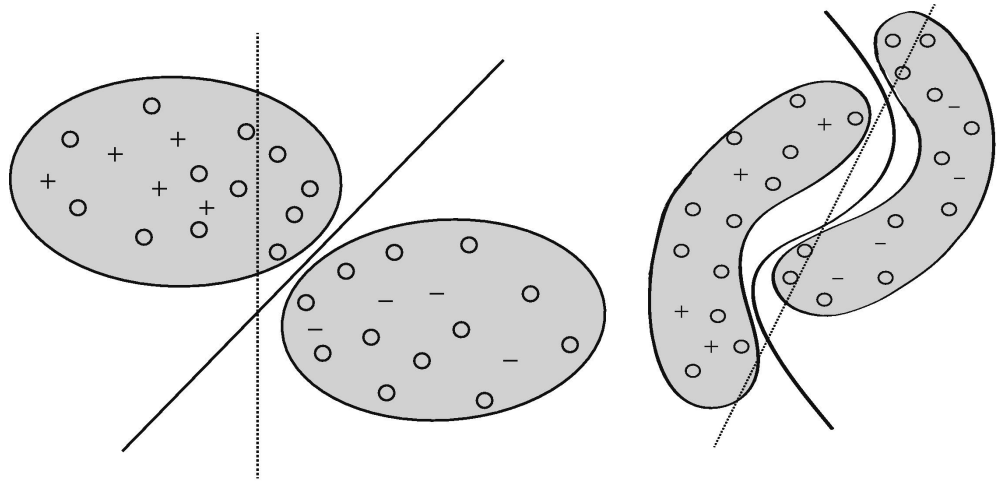
06/08/2021

Semi-supervised Learning(SSL)

What is SSL? SSL is concerned with the use of both the labelled and unlabeled data for training.

Challenge: The design of methods that can exploit the information contained in the distribution of the unlabeled data.

Assumption Cluster assumption, Low-density separation and Manifold assumption.



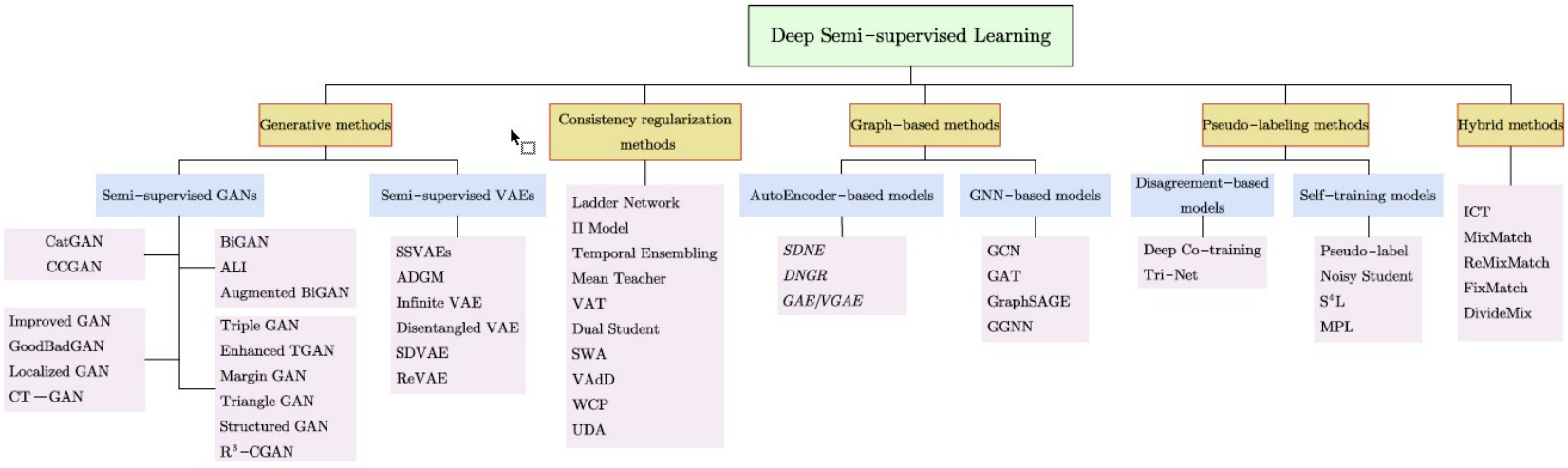
(a) Cluster Assumption (b) Manifold Assumption

Figure 1: The positive and negative signs show labeled examples from two different classes. The circles depict the unlabeled examples. These decision boundaries are moved to regions with lower density (solid line) using unlabeled data.

Manifold learning

- Manifold learning assumes that the observed data lie on a low-dimensional manifold embedded in a higher-dimensional space.
- The manifold assumption states that the space of natural images has the differential-geometric structure of a low-dimensional manifold embedded in the high-dimensional pixel space.
- It should be emphasized that manifold learning refers to a set of methods based on the manifold assumption.

The taxonomy Deep Semi-supervised learning method.



- DSSL studies how to effectively utilize both labeled and unlabeled data by deep neural networks.
- DSSL can be classified into five categories: generative methods, consistency regularization methods, graph-based methods, pseudo-labeling methods, and hybrid methods.

Consistency-based SSL

- Assume that the samples $x_i \in \mathcal{X} \subset \mathbb{R}^D$ and the labels $y_i \in \mathcal{Y} \equiv \{1, \dots, C\}$. We have empirical risk minimization

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}_L|} \sum_{i \in \mathsf{I}_L} \ell(\mathcal{F}_\theta(x_i), y_i) + \mathcal{R}(\theta) \tag{1}$$

- Consistency-based SSL algorithms regularize the learning by enforcing that the learned function $x \mapsto \mathcal{F}_\theta(x)$ respects local derivative and invariance constraints. In the Π-model, the consistency regularization term can be written as follows:

$$\mathcal{R}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathsf{I}_L \cup \mathsf{I}_U} \mathbb{E}_\omega \Big\{ \big\| \mathcal{F}_\theta[\mathcal{S}_\omega(x_i)] - \mathcal{F}_{\theta_\star}(x_i) \big\|^2 \Big\}. \tag{2}$$

Where $\theta_\star$ denotes a copy of the parameter θ, i.e. $\theta_\star = \theta$.

Consistency-based SSL

- To help *propagating* the information of labeled samples to unlabeled samples, in the experiment, we have adopted instead the following regularization term

$$\begin{aligned}\mathcal{R}(\theta) = &\frac{1}{|\mathcal{D}_L|}\sum_{i\in \mathsf{I}_L}\mathbb{E}_\omega\left\{\left\|\mathcal{F}_\theta[\mathcal{S}_\omega(x_i)] - \mathcal{F}_{\theta_\star}(x_i)\right\|^2\right\} \\ &+ \frac{1}{|\mathcal{D}_U|}\sum_{j\in \mathsf{I}_U}\mathbb{E}_\omega\left\{\left\|\mathcal{F}_\theta[\mathcal{S}_\omega(x_j)] - \mathcal{F}_{\theta_\star}(x_j)\right\|^2\right\}\end{aligned} \tag{3}$$

- The quality and variety of the data-augmentation scheme $\mathcal{S} : \mathcal{X} \times \Omega \to \mathcal{X}$ is pivotal to the success of consistency-based SSL methods
- Designing good data-augmentation schemes is an efficient manner of injecting expert/prior knowledge into the learning process.

The contribution of paper

One sentence summary: This paper propose a simple and natural framework leveraging the Hidden Manifold Model to study modern SSL methods.

- We analyse consistency-based methods in setting where analytically tractable low-dimensional can be obtained.
- We establish links with Manifold Tangent Classifiers and demonstrate the consistency-based are more powerful.
- We propose an extension of the Hidden Manifold Model to investigate the properties of consistency-based SSL methods.

Approximate Manifold Tangent Classifier

- Consider the data manifold $\mathcal{M} \subset \mathcal{X} \subset \mathbb{R}^D$ and the dimension of $\mathcal{M}$ is $1 \leq d \leq D$. For $x \in \mathcal{M}$, the tangent plane T_x is also of dimension d with an orthonormal basis $\mathrm{e}_1^x, \ldots, \mathrm{e}_d^x \in \mathbb{R}^D$. Given small coefficients $\omega_1, \ldots, \omega_d \in \mathbb{R}$, the transformed sample $\overline{x} \in \mathcal{X}$ also lies, or is very close to, the data manifold $\mathcal{M}$. $\overline{x}$ is defined as

$$\overline{x} \;=\; x + \sum_{j=1}^{d} \omega_j \, \mathrm{e}_j^x$$

- A possible stochastic data-augmentation scheme can therefore be defined as $\mathcal{S}_\omega(x) = x + V_\omega$ where $V_\omega = \sum_{j=1}^{d} \omega_j \, \mathrm{e}_j^x$.
- To enforce that the function $x \to \mathcal{F}_\theta(x)$ is locally approximately constant along the manifold $\mathcal{M}$, one can thus penalize the derivatives of $\mathcal{F}_\theta$ at x in the directions V_ω.

Approximate Manifold Tangent Classifier

- Denoting by $\mathsf{J}_x \in \mathbb{R}^{C,D}$ the Jacobian with respect to $x \in \mathbb{R}^D$ of $\mathcal{F}_\theta$ at $x \in \mathcal{M}$, this can be implemented by adding a penalization term of the type $\mathbb{E}_\omega[\|\mathsf{J}_x\, V_\omega\|^2] = \mathsf{Tr}\big(\Gamma \otimes \mathsf{J}_x^T \mathsf{J}_x\big)$, where $\Gamma \in \mathbb{R}^{D,D}$ is the covariance matrix of the random vector $\omega \to V_\omega$.
- For any $x, \omega \in \mathcal{X} \times \Omega$, $\mathcal{S}_{\varepsilon\,\omega}(x) = x + \varepsilon\ , \mathsf{D}(x,\omega) + \mathcal{O}(\varepsilon^2)$, for some derivative mapping $\mathsf{D} : \mathcal{X} \times \Omega \to \mathcal{X}$, it follows that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2}\, \mathbb{E}_\omega\left[\|\mathcal{F}_\theta[\mathcal{S}_{\varepsilon\,\omega}(x)] - \mathcal{F}_\theta(x)\|^2\right] = \mathbb{E}_\omega\left[\|\mathsf{J}_x\, \mathsf{D}(x,\omega)\|^2\right]$$
$$= \mathsf{Tr}\Big(\Gamma_{x,\mathcal{S}} \otimes \mathsf{J}_x^T\, \mathsf{J}_x\Big)$$

where $\Gamma_{x,\mathcal{S}}$ is the covariance matrix of the $\mathcal{X}$-valued random vector $\omega \mapsto \mathsf{D}(x,\omega) \in \mathcal{X}$. This shows that consistency-based methods can be understood as approximated Jacobian regularization.

Manifold Tangent Classifier

- The MTC uses a Contractive-Auto-Encoder (CAE) to extract in an unsupervised manner a good representation of the dataset.
- The CAE can subsequently be leveraged to obtain an approximate basis of each tangent plane T_{x_i} for $x_i \in \mathcal{D}$, which can then be used for penalizing the Jacobian of the mapping $x \mapsto \mathcal{F}_\theta(x)$ in the direction of the tangent plane to $\mathcal{M}$ at x.

How to build and train a deep network with MTC method.

- Train a stack of K CAE+H layers. Each is trained in turn on the representation learned by the previous layer.
- For each $x_i \in \mathcal{D}$ compute the Jacobian of the last layer representation $\frac{\partial h^{(k)}}{\partial x}(x_i)$ and its SVD.Store the leading d_M singular vectors in set.
- On top of the K pre-trained layers, stack an output layer of size the number of classes. Fine-tune the whole network, using for each x_i and tangent directions $\mathcal{B}_{x_i}$

Limitations

Somewhat simplistic approach consisting in adding an isotropic Gaussian noise to the data samples is unlikely to deliver satisfying results, Why?

- This mechanism does not take at all into account the local-geometry of the data-manifold.
- It is equivalent to penalizing the Frobenius norm $\|\mathrm{J}_x\|_{\mathrm{F}}^2$ of the Jacobian of the mapping $x \mapsto \mathcal{F}_\theta(x)$; in a linear model, that is equivalent to the standard *ridge regularization*.
- Jacobian penalization techniques are not efficient at learning highly non-linear manifolds that are common.

Limitations

In general, domain-specific data-augmentation schemes lead to much better regularization than Jacobian penalization.

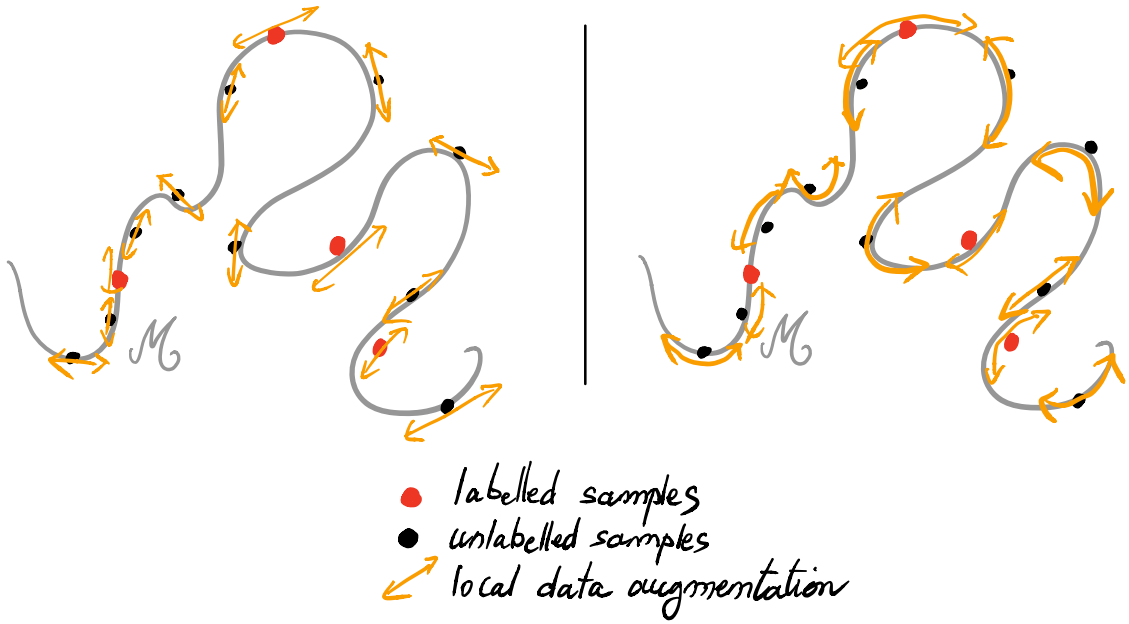


Figure 2: **Left:** Jacobian (i.e. first order) Penalization method are short-sighted and do not exploit fully the data-manifold **Right:** Data-Augmentation respecting the geometry of the data-manifold.

Limitations

- Jacobian penalization techniques are not efficient at learning highly non-linear manifolds.
- For example, in "pixel space", a simple image translation is a highly non-linear transformation only well approximated by a first order approximation for very small translations.
- In other words, if $x \in \mathcal{X}$ represents an image and $g(x, v)$ is its translated version by a vector v, the approximation $g(x, v) \approx x + \nabla_v g(x)$, with $\nabla_v g(x) \equiv \lim_{\varepsilon \to 0} (g(x, \varepsilon\, v) - g(x)/\varepsilon$, becomes poor as soon as the translation vector v is not extremely small.

Limitations

- In computer vision, translations, rotations and dilatations are often used as sole data-augmentation schemes: this leads to a poor local exploration of the data-manifold since this type transformations only generate a very low dimensional exploration manifold.
- Enriching the set of data-augmentation degrees of freedom with transformations such as elastic deformation or non-linear pixel intensity shifts is crucial to obtaining a high-dimensional local exploration manifold that can help propagating the information on the data-manifold efficiently.

4. Asymptotic Properties

4.1 Fluid limit

- Consider the standard Π-model trained with a standard Stochastic Gradient Descent (SGD). Denote by $\theta_t \in \Theta$ the current value of the parameter and $\eta > 0$ the learning rate. We have

$$\theta_{k+1} = \theta_k - \eta \nabla_\theta \Bigg\{ \frac{1}{|\mathcal{B}_L|} \sum_{i \in \mathcal{B}_L} \ell(\, \mathcal{F}_{\theta_k}(x_i),\, y_i \,) + \frac{\lambda}{|\mathcal{B}_L|} \sum_{j \in \mathcal{B}_L} \Big\| \mathcal{F}_{\theta_k}(\mathcal{S}_\omega[x_j]) - f_j \Big\|^2 + \frac{\lambda}{|\mathcal{B}_U|} \sum_{k \in \mathcal{B}_U} \Big\| \mathcal{F}_{\theta_k}(\mathcal{S}_\omega[x_k]) - f_k \Big\|^2 \Bigg\} \tag{4}$$

for a parameter $\lambda > 0$ that controls the trade-off between supervised and consistency losses, as well as subsets $\mathcal{B}_L$ and $\mathcal{B}_U$ of labelled and unlabelled data samples, and $f_j \equiv \mathcal{F}_{\theta\star}(x_j)$ for $\theta_\star \equiv \theta_k$.

4.2 Minimizers are harmonic functions

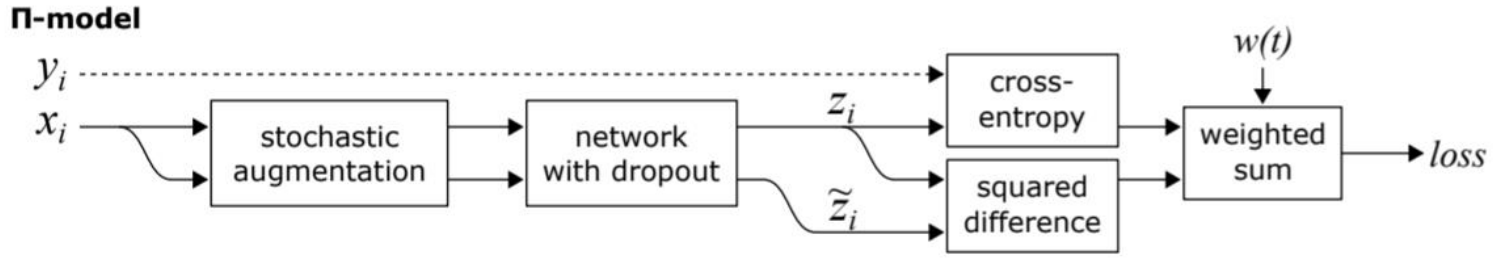
- Assume that $\mathcal{M} \subset \mathbb{R}^D$ can be globally parametrized by $\Phi : \mathbb{R}^d \to \mathcal{M} \subset \mathbb{R}^D$. Consider a data-augmentation $\mathcal{S}_{\varepsilon\omega}(x) = \Phi(z + \varepsilon\omega)$ for $z = \Phi^{-1}(x)$ and a sample ω from a $\mathbb{R}^d$-valued centred and isotropic Gaussian distribution. We investigate the regime $\varepsilon \to 0$ and, the minimization of the consistency-regularized objective

$$\mathcal{L}_L(\theta) + \frac{\lambda}{\varepsilon^2} \int_{\mathbb{R}^d} \mathbb{E}_\omega \left\{ \left\| \mathcal{F}_\theta[\mathcal{S}_{\varepsilon\omega}(\Phi(z))] - \mathcal{F}_\theta(\Phi(z)) \right\|^2 \right\} \mu(dz). \tag{5}$$

- As $\varepsilon \to 0$,the objective function approaches the quantity

$$\mathsf{G}(f_\theta) \equiv \frac{1}{|\mathcal{D}_L|} \sum_{i \in \mathsf{I}_L} \ell(f_\theta(z_i), y_i) + \lambda \int_{\mathbb{R}^d} \|\nabla_z f_\theta(z)\|^2 \, \mu(dz). \tag{6}$$

4.2 Minimizers are harmonic functions



- The generalization properties of consistency-based SSL methods will typically be insensitive to this parameter, in the regime of small data-augmentation at least.
- consistency-based SSL methods are indeed based on the same principles as more standard graph-based approaches.

4.2 Minimizers are harmonic functions

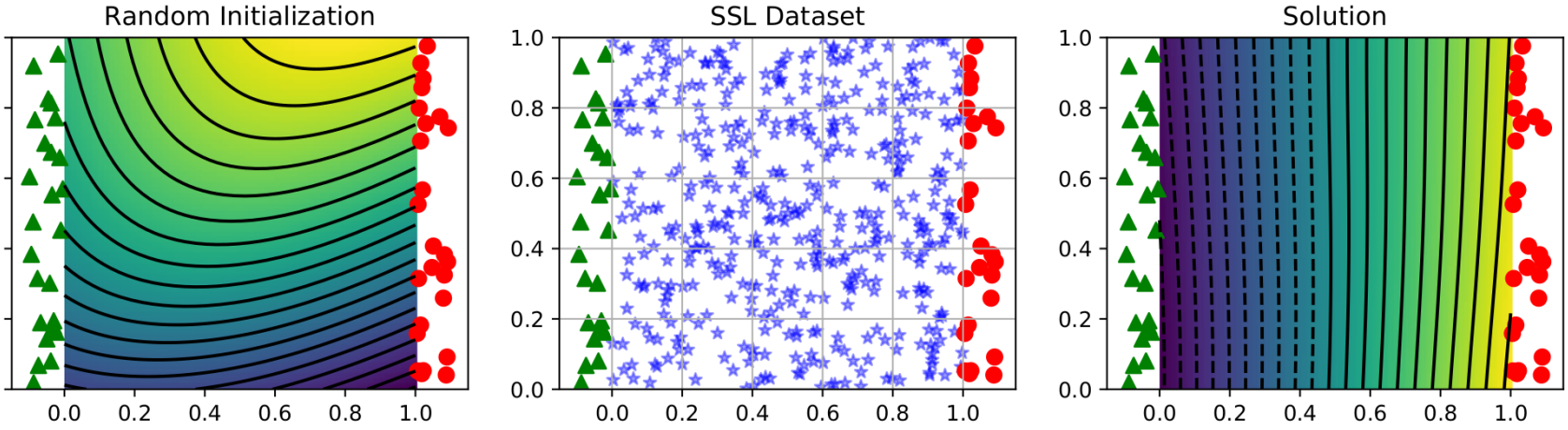


Figure 3: Labelled data samples with class $y = 0$ (green triangle) and $y = +1$ (red dot) are placed on the Left/Right boundary of the unit square. Unlabelled data samples (blue stars) are uniformly placed within the unit square. We consider a simple regression setting with loss function $\ell(f, y) = \frac{1}{2}(f - y)^2$. **Left:** Randomly initialized neural network. **Middle:** labelled/unlabelled data **Right:** Solution of f obtained by training a standard Π-model. It is the harmonic function $f(u, v) = u$.

4.3 Generative model for Semi-Supervised Learning

- The goal of this paper is to understand the mechanisms that are at play when consistency-based SSL methods are used to uncover the structures present in real datasets.
- It is important to build simplified and tractable generative models of data that **(1)** respect these low-dimensional structures and **(2)** allow the design of efficient data-augmentation schemes.
- Follow the Hidden Manifold Model framework, the author introduce a model of synthetic data concentrating near low-dimensional structures and analyze the learning curve associated to a class of two-layered neural networks.

Low-dimensional structure

- The mapping Φ is chosen to be a neural network with a single hidden layer with H neurons.
- For $z = (z^1, \ldots, z^d) \in \mathbb{R}^d$, set $\Phi(z) = A^{1\to 2}\,\varphi(A^{0\to 1}z + b^1)$ for matrices $A^{0\to 1} \in \mathbb{R}^{H,d}$ and $A^{1\to 2} \in \mathbb{R}^{D,H}$, bias vector $b^1 \in \mathbb{R}^H$ and non-linearity $\varphi : \mathbb{R} \to \mathbb{R}$ applied element-wise.
- We adopt the standard normalization $A^{0\to 1}_{i,j} = w^{(1)}_{i,j}/\sqrt{d}$ and $A^{1\to 2}_{i,j} = w^{(2)}_{i,j}/\sqrt{H}$ for weights $w^{(k)}_{i,j}$ drawn i.i.d from a centred Gaussian distribution with unit variance.

Data augmentation

- consider the natural data-augmentation $\mathcal{S}_{\varepsilon\omega}(x_i) = \Phi(z_i + \varepsilon\omega)$, where the sample $\omega \in \mathbb{R}^d$ samples from an isotropic Gaussian distribution with unit covariance and $\varepsilon > 0$.
- $\mathcal{S}_{\varepsilon\omega}(x_i)$ belongs to the data-manifold $\mathcal{M}$ for any perturbation vector $\varepsilon\,\omega$. For any value of ε, the data-augmentation preserves the low-dimensional manifold: perturbed samples $\mathcal{S}_{\varepsilon\omega}(x_i)$ *exactly* lie on the data-manifold.
- The larger ε, the more efficient the data-augmentation scheme.

Neural architecture and optimization

- Fitting a two-layered neural network $\mathcal{F}_\theta : \mathbb{R}^D \to \mathbb{R}$ by minimising $\mathcal{L}_L(\theta) \equiv (1/|\mathcal{D}_L|) \sum_i \ell[\mathcal{F}_\theta(x_i), y_i]$ where $\ell(f, y) = \log(1 + \exp[-y\, f])$.
- Assume that $|\mathcal{D}_L| = 10$ labeled data pairs $\{x_i, y_i\}_{i=1_L}$, as well as $|\mathcal{D}_U| = 1000$ unlabeled data samples. The ambient space has dimension $D = 100$ and the data manifold $\mathcal{M}$ has dimension $d = 10$
- For minimizing the consistency-based SSL objective $\mathcal{L}_L(\theta) + \lambda\, \mathcal{R}(\theta)$, we use the standard strategy consisting in first minimizing the un-regularized objective alone $\mathcal{L}_L$ for a few epochs in order for the function $\mathcal{F}_\theta$ to be learned in the neighbourhood of the few labeled data-samples before switching on the consistency-based regularization.

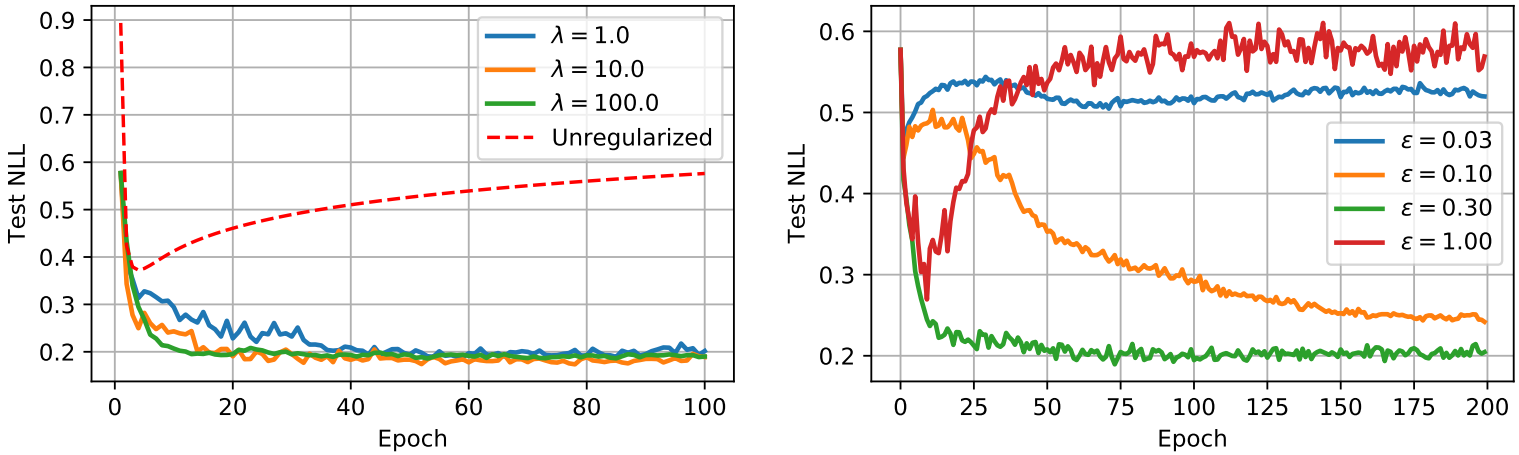


Figure 4: **Left:** For a fixed data-augmentation scheme, generalization properties for λ spanning two orders of magnitude. **Right:** Influence of the quantity of the data-augmentation of the generalization properties.

- Much larger or smaller values of λ do lead to convergence and stability issues.
- Too low an amount of data-augmentation (i.e. $\varepsilon = 0.03$) and the final performance is equivalent to the un-regularized method. Too large an amount of data-augmentation (i.e. $\varepsilon = 1.0$) also leads to poor generalization properties.

Quality of the Data-augmentation

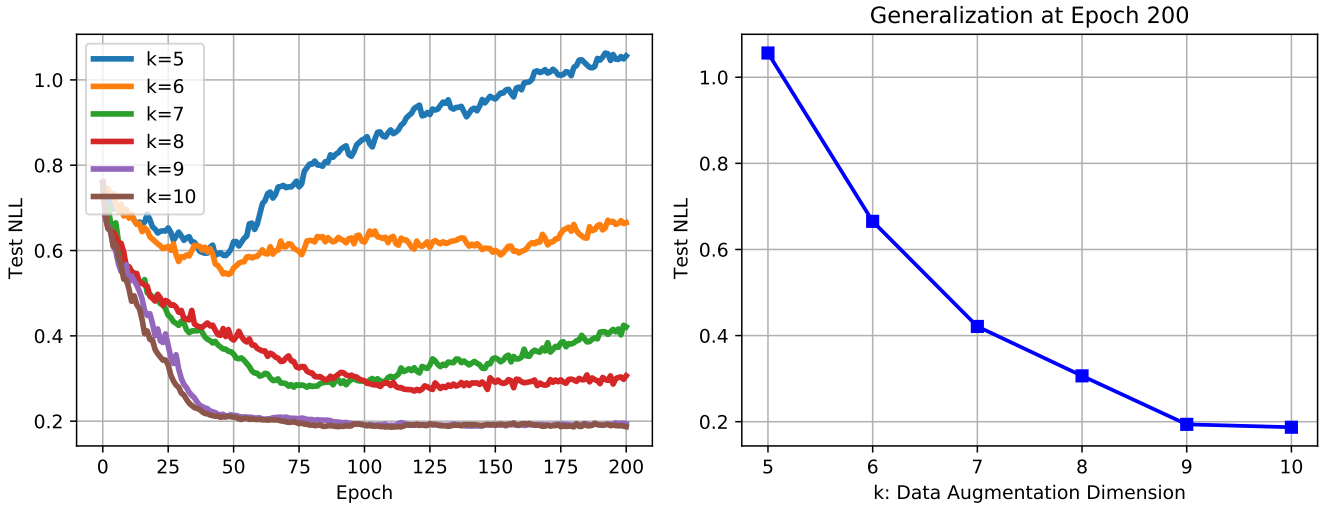


Figure 5: **Left:** Learning curves (Test NLL) for data-augmentation dimension $k \in [5, 10]$ **Right:** Test NLL at epoch $N = 200$ (see left plot) for data-augmentation dimension $k \in [5, 10]$.

Consider a perturbation $\mathcal{S}_{\varepsilon\omega[k]}(x_i) = \Phi(z_i + \omega[k])$ for $x_i = \Phi(z_i)$ where the noise term $\omega[k]$ is defined as follows. For a *data-augmentation dimension* parameter $1 \leq k \leq d$ we have $\omega[k] = (\xi_1, \ldots, \xi_k, 0, \ldots, 0)$ for i.i.d standard Gaussian samples $\xi_1, \ldots, \xi_k \in \mathbb{R}$.

Mean-Teacher versus Π-model

- MT parameter: $\theta_{\text{avg},k} = \beta_{\text{MT}}\,\theta_{\text{avg},k-1} + (1-\beta_{\text{MT}})\theta_k$ for the $\theta_{\text{avg},k}$ with different scales $\beta_{\text{MT}} \in \{0.9, 0.95, 0.99, 0.995\}$.

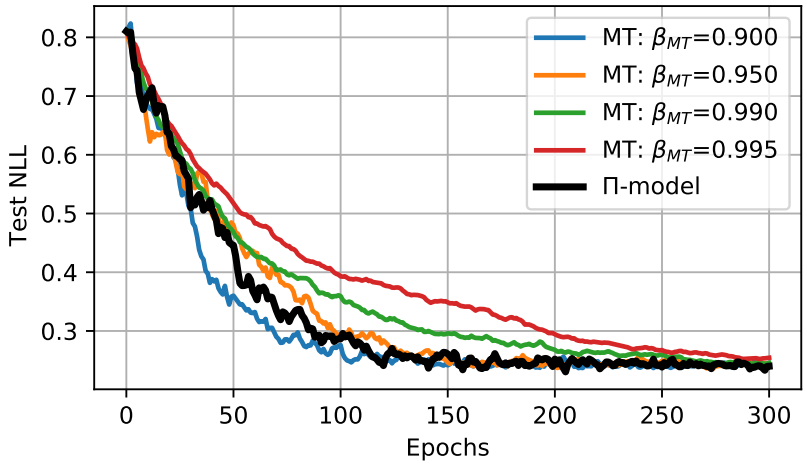


Figure 6: Mean-Teacher (MT) learning curves (Test NLL) for different values of the exponential smoothing parameter $\beta_{\text{MT}} \in (0, 1)$. For $\beta_{\text{MT}} \in \{0.9, 0.95, 0.99, 0.995\}$, the final test NLL obtained through the MT approach is identical to the test NLL obtained through the Π-model. In all the experiments, we used $\lambda = 10$ and used SGD with momentum $\beta = 0.9$.

Comment of reviewer

Final Decision

ICLR 2021 Conference Program Chairs

08 Jan 2021 (modified: 13 Jan 2021) ICLR 2021 Conference Paper1052 Decision Readers: Everyone

Decision: Accept (Poster)

Comment:

This paper provides some theoretical perspective on the use of data augmentation in consistency regularization-based semi-supervised learning. The framework used in the paper argues that high-quality data augmentation should move along the data manifold. This generic view allows the paper's ideas to be applied across datasets (as opposed to image-specific data augmentation used in state-of-the-art semi-supervised learning algorithms). I am not aware of any other work raising these points, and indeed this paper is significant in that it provides a new and potentially useful perspective on the most performative semi-supervised learning approach. Reviewers agreed that the paper was clear and useful. The main concern was that the paper only included experiments in toy settings. Indeed, it would have been much more impactful to apply these ideas to state-of-the-art semi-supervised learning methods, but I think it can be excused given the theoretical focus of the work.

- This paper provides some theoretical perspective on the use of data augmentation in consistency regularization-based semi-supervised learning.
- The framework used in the paper argues that high-quality data augmentation should move along the data manifold.
- This generic view allows the paper's ideas to be applied across datasets (as opposed to image-specific data augmentation used in state-of-the-art semi-supervised learning algorithms).