Why Does Deep Learning Work?

Bhavya Bhatt

Indian Institute of Technology Mandi

July 2020

Overview

1. Recap of IB principle
2. Information Geometry
3. Future Work

Information Bottleneck Principle

The central idea of IB-Theory is[1]

"Regularization by optimal intermediate representations"

[1] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *CoRR*, vol. abs/1503.02406, 2015. arXiv: 1503.02406. [Online]. Available: http://arxiv.org/abs/1503.02406.

Mathematical Recapitulation

Given $p_{XY}(x, y)$ for the dataset, IB objective is as follows

$$L(p(\hat{X}|X)) = I(\hat{X}, X) - \beta I(\hat{X}, Y)$$

and probability distribution of minimum sufficient statistics (optimal) is

$$p^*(\hat{X}|X) = \underset{p(\hat{X}|X)}{\arg\min}\, L(p(\hat{X}|X))$$

So intermediate representation $\hat{X}$ is a stochastic compressed representation of X

True optimal and Empirical optimal

The optimal curve for different β with true distribution $p_{XY}(x, y)$ (in black) and with empirical distribution $\hat{p}_{XY}(x, y)$ estimated with finite samples in the dataset is as follows [2]

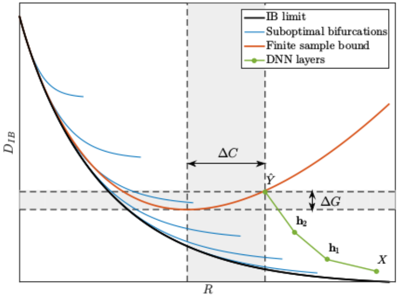


Figure: A qualitative information plane, with a hypothesized path of the layers in a typical DNN (green line) on the training data. The black line is the optimal achievable IB limit, and the blue lines are sub-optimal IB bifurcations.

[2] Figure taken from [1]

Two phase training dynamics - Generalization and Compression phase

First observed in [2]

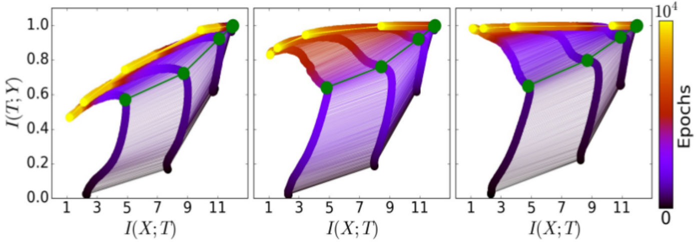


Figure: The evolution of the layers with the training epochs in the information plane. On the left - 5%, middle - 45%, and right - 85% of the data.

3

[3] Figure taken from R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017. arXiv: 1703.00810. [Online]. Available: `http://arxiv.org/abs/1703.00810`

The Controversy

Paper [3] attacked the original paper claiming[4]

- IB-Theory is not fundamental theory
- Depends on specific activation used
- Showed two phase dynamics do not hold for RELU activation

But later more accurate MI estimators published and observed the two phases !

Even then , IB-Theory have issues in case of deterministic networks

[4] A. M. Saxe, Y. Bansal, J. Dapello, *et al.*, "On the information bottleneck theory of deep learning," in *International Conference on Learning Representations*, 2018. [Online]. Available: `https://openreview.net/forum?id=ry_WPG-A-`.

Resolution - Accurate Mutual Information estimation methods

Improvement in estimation methods

- Use more accurate MI estimation methods like EDGE, MINE [5]etc.
- Use tight bounded MI representations
- Research for which estimation method to use when

[5]MINE - I. Belghazi, S. Rajeswar, A. Baratin, *et al.*, "MINE: mutual information neural estimation," *CoRR*, vol. abs/1801.04062, 2018. arXiv: 1801.04062. [Online]. Available: http://arxiv.org/abs/1801.04062
EDGE - M. Noshad and A. O. H. III, "Scalable mutual information estimation using dependence graphs," *CoRR*, vol. abs/1801.09125, 2018. arXiv: 1801.09125. [Online]. Available: http://arxiv.org/abs/1801.09125

Problems with Mutual Information

Problem with mutual information is as follows:

- Difficult to estimate in practice fewer samples available.
- Suffers discontinuity in case of non-stochastic deterministic networks[6].
- No measure for robustness to noise.

[6] R. A. Amjad and B. C. Geiger, "How (not) to train your neural network using the information bottleneck principle," *CoRR*, vol. abs/1802.09766, 2018. arXiv: 1802.09766. [Online]. Available: http://arxiv.org/abs/1802.09766.

Dependence Criterion

Instead of MI as dependence criterion use more robust criterion which captures

- **inform about** Y - $\hat{X}$ should be sufficient statistics
- **be maximally compressed** - representation $\hat{X}$ should not tell about X
- **admit a simple decision function** - Y can be estimated from $\hat{X}$ using simple functions
- **be robust** - small noise should not change $\hat{X}$ with big differences

First two are captured by mutual information criterion but there is a need for criterion which captures last two too !

Hilbert Schmidt Independence Criteria

An alternate criterion HSIC was first introduced in paper[7] which gave one of the most interesting application of IB principle

"Deep Learning without Back-Propagation"

[7] W.-D. K. Ma, J. P. Lewis, and W. B. Kleijn, *The hsic bottleneck: Deep learning without back-propagation*, 2019. arXiv: 1908.01580 [cs.LG].

PyGlow: Python Package for Information Theory of Deep Learning

PyGlow

Information Theory of Deep Learning

[8]

[8] GitHub repo on: https://github.com/spino17/PyGlow
PyGlow Docs: https://pyglow.github.io/

Need for Geometric Reformulation

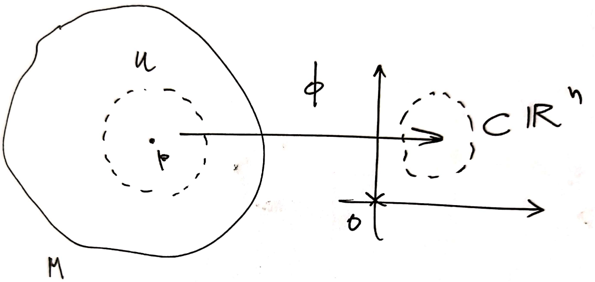
Need for a geometric framework of information theory:

- Models should not be specified with particular value of parameter[9].
- Learning dynamics should be independent of parameters.
- Distance measure in information theory is non-euclidean.
- Coordinate-free formulation to make deep connections of information theory with DL visible.

---

[9] G. Desjardins, K. Simonyan, R. Pascanu, *et al.*, *Natural neural networks*, 2015. arXiv: 1507.00210 [stat.ML].

Manifolds - Topological spaces which locally looks like $\mathbb{R}^n$.

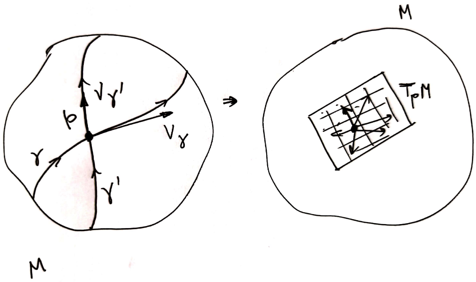


For any point $p \in M$, there exists an open set U such that we have a homeomorphism from U to $\mathbb{R}^n$.

Ingredients - Vectors

Vectors - Objects that captures the rate with which a function $f\colon M \to \mathbb{R}$ on manifold changes along a curve $\gamma\colon [0,1] \to M$. They live in tangent space T_pM.

$$V_\gamma[f] = \frac{\mathrm{d} f \circ \gamma(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=0}$$



Basis of T_PM - $\{\partial_\alpha = \frac{\partial}{\partial x^\alpha}\Big|_p\}_{\alpha=1\ldots n}$

Ingredients - One-forms

One-forms - Objects that take vectors from abstract objects to a point in $\mathbb{R}^n$ (components of vectors in common treatments). They live in cotangent dual space T_p^*M.

$$df(V) := V[f] \in \mathbb{R}$$

This object is called as gradient of a function f. We now define dual basis ω_α of T_p^*M, which are dual to the coordinate basis of T_pM as

$$\omega_\alpha\left(\frac{\partial}{\partial x^\beta}\bigg|_p\right) := \delta^\alpha_\beta$$

Basis of T_p^*M - $\{dx^\alpha\}_{\alpha=1\ldots n}$

Ingredients - Metric Tensor

Metric Tensor - Gives a notion of length (via infinitesimal lengths) on the manifold. $g_p \colon T_pM \times T_pM \to \mathbb{R}$ as

$$g_p(U, V) = \langle U, V \rangle$$

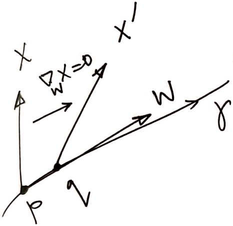
for any $U, V \in T_pM$. In coordinate basis,

$$g = g_{\alpha\beta}(p) dx^\alpha \otimes dx^\beta$$

for example: $ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$ is infinitesimal length on sphere in spherical coordinates basis.

Ingredients - Affine Connections

Affine Connection - Provides us with a definition of parallel transport and directional derivative.



For a vector $X \in T_pM$, the parallel transport vector $X' \in T_qM$ along the curve γ with tangent vector W is

$$X'^{\alpha}(q) = X^{\alpha}(p) - \Gamma^{\alpha}_{\beta\gamma}(p) W^{\gamma} X^{\beta} \Delta\lambda$$

where $\Gamma^{\alpha}_{\beta\gamma}$ are christoffel symbols which captures curvi-linearity of coordinate systems.

Affine Connections

The covariant derivative $\nabla_W X$ w.r.t to connection ∇ is given by

$$\nabla_W X = \frac{\mathrm{d}X^\alpha}{\mathrm{d}\lambda} + \Gamma^\alpha_{\beta\gamma} W^\gamma X^\beta$$

Metric Induced Connection:

- Metric compatibility - X[g(Y, Z)] = 0 for Y, Z being parallel transported along X.
- Torsion Free - $\Gamma^\alpha_{\beta\gamma} = \Gamma^\alpha_{\gamma\beta}$.

These restriction are enough to obtain closed form expression for christoffel symbols

$${}^{LC}\Gamma^\alpha_{\beta\gamma} = \frac{1}{2} g^{\alpha\lambda}(\partial_\gamma g_{\lambda\beta} + \partial_\beta g_{\gamma\lambda} - \partial_\lambda g_{\beta\gamma})$$

where ${}^{LC}\Gamma^\alpha_{\beta\gamma}$ is Levi-Civita christoffel symbols.

Fundamental Theorem of Riemannian Geometry - The above Levi-Civita connections are unique.

Information Manifolds - (M, g, ∇, ∇^*)

Conjugate-Connection Manifolds - Spaces of allowed parameters for a statistical decision making problem. The CCM's have a dual structure of connections (∇, ∇^*) with ∇^* defined by:

$$X[g(Y, Z)] = g(\nabla_X Y, Z) + g(Y, \nabla^\star_X Z)$$

for any arbitrary vector fields X, Y and Z.

Note: $(\nabla^\star)^\star = \nabla$. Equivalent definition -

$$\langle U, V \rangle = \langle \prod_{\gamma(t)}^{\nabla} U, \prod_{\gamma(t)}^{\nabla^\star} V \rangle$$

which is statement that dual connections are metric-preserving. Note from this definition Levi-Civita Connection is self-dual $({}^{LC}\nabla, {}^{LC}\nabla^* = {}^{LC}\nabla)$

Statistical Manifolds

A statistical manifold (M, g, C) is a manifold equipped with a metric tensor g and a totally symmetric 3-tensor C Amari-Chentsov tensor where

$$C(X, Y, Z) := \langle \nabla_X Y - \nabla^{\star}_X Y, Z \rangle$$

or in component form

$$C^k_{ij} = \Gamma^k_{ij} - \tilde{\Gamma}^k_{ij}$$

Also for given pair $(\Gamma^{\alpha}_{\beta\gamma}, \tilde{\Gamma}^{\alpha}_{\beta\gamma})$, we have a self-dual connection

$$^{LC}\nabla = \frac{1}{2}(\nabla + \nabla^{\star})$$

which from fundamental theorem of Riemannian Geometry is Levi-Civita connections.

One-parameter family of Conjugate Connections

For a given $(\Gamma^{\alpha}_{\beta\gamma}, \tilde{\Gamma}^{\alpha}_{\beta\gamma})$ with Amari-Chentsov tensor $C^{\alpha}_{\beta\gamma}$.

Then for a continous parameter λ, $\lambda C^{\alpha}_{\beta\gamma}$ is also totally symmetric tensor and can be Amari-Chentsov tensor of some other pair of dual connections $(\nabla^{-\lambda}, (\nabla^{-\lambda})^* = \nabla^{\lambda})$ with christoffel symbols as:

$$\Gamma^{\lambda}_{ijk} = \Gamma^{0}_{ijk} - \frac{\lambda}{2} C_{ijk}$$
$$\Gamma^{-\lambda}_{ijk} = \Gamma^{0}_{ijk} + \frac{\lambda}{2} C_{ijk}$$

where Γ^{0}_{ijk} is Levi-Civita connections.

Fundamental Theorem of Information Geometry - A manifold $(M, g, \nabla^{-\lambda}, \nabla^{\lambda})$ is ∇^{λ}-flat if and only if it is $\nabla^{-\lambda}$-flat.

Big Question - How can we obtain above structures ?

Two ways to obtain CCM structure (g, ∇, ∇^*) canonically from the problem is:

- From divergence considered in the problem to capture dissimilarity between probability distributions.
- From probability distribution itself by using max-log likelihood principle.

Conjugate Connection Structure from Divergences

Divergence:
$D\colon M \times M \to [0, \infty)$ on a manifold M with a local chart $\theta \subset \mathbb{R}^D$ with following properties

- $D(\theta : \theta') \geq 0 \;\forall\; \theta, \theta' \in \Theta$ where equality holds if and only if $\theta = \theta'$.
- $\partial_{i,.} D(\theta, \theta')\Big|_{\theta=\theta'} = \partial_{.,j} D(\theta, \theta'\Big|_{\theta=\theta'} = 0 \;\forall i, j$
- $-\partial_{.,i}\partial_{.,j} D(\theta, \theta')\Big|_{\theta=\theta'}$ is positive-definite.

We can also define a dual-divergence by swapping the arguments

$$D^\star(\theta, \theta') = D(\theta, \theta')$$

Intuition:

$$\begin{aligned}
D(\theta, \theta + \delta\theta) &= D(\theta, \theta) + \frac{\partial}{\partial\theta^i} D\Big|_\theta \delta\theta + \frac{\partial^2}{\partial\theta^i \partial\theta^j} D\Big|_\theta \delta\theta^i \delta\theta^j \\
&= \frac{\partial^2}{\partial\theta^i \partial\theta^j} D\Big|_\theta \delta\theta^i \delta\theta^j \\
&= \partial_{i,j} D(\theta, \theta')\Big|_\theta \delta\theta^i \delta\theta^j
\end{aligned}$$

Conjugate Connection Structure from Divergences

we define the conjugate connection structure naturally induced by divergence $D(\theta, \theta')$ as follows:

- $g := -\partial_{i,j} D(\theta, \theta')\Big|_{\theta=\theta'}$
- $\Gamma_{ijk} := -\partial_{ij,k} D(\theta, \theta')\Big|_{\theta=\theta'}$
- $\tilde{\Gamma}_{ijk} := -\partial_{k,ij} D(\theta, \theta')\Big|_{\theta=\theta'}$

Conjugate Connection Structure from Probability Distribution

Let $\mathfrak{P}$ be a parametric family of probability distribution

$$\mathfrak{P} = \{p_\theta(X)\}_{\theta \in \Theta}$$

where Θ is parameter space. Order of parameter space is dimension of parameter space. We use the familiar log-likelihood function

$$l(\theta; x) = \log p_\theta(x)$$

We now define metric as

$$I(\theta) = g_{ij} = \mathbb{E}_\theta[\partial_i l \partial_j l]$$

This relates with the Cramer-Rao lower bound on variance of estimator

$$Var_\theta[\hat{\theta}_n(x)] \geq \frac{1}{n} I^{-1}(\theta)$$

Conjugate Connection Structure from Probability Distribution

Now we define two types of connections which is naturally defined in terms of the above probability distribution

- Exponential connection - ${}^{e}\Gamma_{ijk} := \mathbb{E}_\theta[(\partial_i\partial_j l)\partial_k l]$
- Mixing connection - ${}^{m}\Gamma_{ijk} := \mathbb{E}_\theta[(\partial_i\partial_j l + \partial_i l \partial_j l)\partial_k l]$

For a complete treatment on the subject refer the paper[10].

[10] F. Nielsen, *An elementary introduction to information geometry*, 2018. arXiv: 1808.08271 [cs.LG].

Why this is relevant in Deep Learning?

Natural Gradient Descent - If the problem has non-euclidean information manifold (which is the case most of the time) then the update rule in gradient descent algorithm is given by

$$\theta_{t+1} = \theta_t - \eta_t \tilde{\nabla} l(\theta)$$

Here $\tilde{\nabla}$ is natural gradient defined as[11]

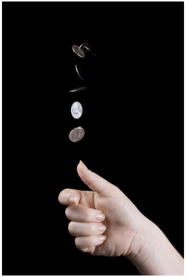
$$\tilde{\nabla} l(\theta) = G^{-1} \nabla l(\theta)$$

where $\nabla = (\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n})$ is usual gradient and $G = g_{\alpha\beta}$.

[11] S.-i. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998. DOI: 10.1162/089976698300017746. eprint: https://doi.org/10.1162/089976698300017746. [Online]. Available: https://doi.org/10.1162/089976698300017746.

Scope of Future Work - Classical VS Non-Classical Models

Classical Probability - Source of stochastic outcome comes from the missing of intractable hidden variables from the dynamics. For example: tossing of coin[12]



Non-classical Probability - Stochastic outcomes are intrinsic even if the dynamics is complete (no local hidden variables missing). For example: quantum probability

[12] Image Source - https://www.bellevuerarecoins.com/history-coin-flip/

Scope of Future Work - Bell's Inequality

How to distinguish between classical and non-classical probabilities?

"Bell's Inequality"

Classical Probability - Obeys Bell's Inequality.

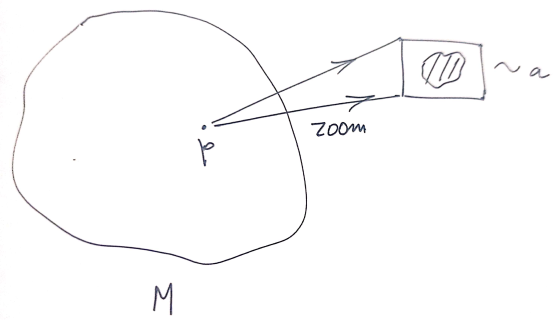
Non-Classical Probability - Does not obey Bell's inequality.[13]

[13] J. S. Bell, "On the einstein podolsky rosen paradox," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, 2004 [1964], pp. 14–21.

Scope of Future Work - Classical VS Non-Classical Models

Two main features of dynamical theories which obey Bell's Inequality:

- Counter-factual definiteness.
- Local hidden variables.



In this diagram above, a is the order of intrinsic uncertainty. For quantum mechanics a is $\hbar$.

Scope of Future Work - Classical VS Non-Classical Models

Classical information manifolds are non-classical information manifolds in the limit $a \longrightarrow 0$.
For reverse, from classical information manifolds to non-classical information manifolds, the process is called

"Quantization"

Quantization: real chart coordinates $\theta_\alpha \in \mathbb{R} \longrightarrow$ matrix-valued operators Θ_α (finite or infinite dimensional)

$$[\Theta_\alpha, \Theta_\beta] \sim a\Delta_{\alpha\beta}\mathbb{I}$$

where $\Delta_{\alpha\beta}$ quantifies uncertainty coefficient between different pairs (α, β).

Scope of Future Work - Are Deep Neural Networks non-classical?

Do hidden layers have non-classical stochastic nature ?

"Analyse Classical Information Manifolds for DNN"

Thank You

References I

[1] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *CoRR*, vol. abs/1503.02406, 2015. arXiv: 1503.02406. [Online]. Available: http://arxiv.org/abs/1503.02406.

[2] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017. arXiv: 1703.00810. [Online]. Available: http://arxiv.org/abs/1703.00810.

[3] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ry_WPG-A-.

[4] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. C. Courville, "MINE: mutual information neural estimation," *CoRR*, vol. abs/1801.04062, 2018. arXiv: 1801.04062. [Online]. Available: http://arxiv.org/abs/1801.04062.

[5] M. Noshad and A. O. H. III, "Scalable mutual information estimation using dependence graphs," *CoRR*, vol. abs/1801.09125, 2018. arXiv: 1801.09125. [Online]. Available: http://arxiv.org/abs/1801.09125.

[6] R. A. Amjad and B. C. Geiger, "How (not) to train your neural network using the information bottleneck principle," *CoRR*, vol. abs/1802.09766, 2018. arXiv: 1802.09766. [Online]. Available: http://arxiv.org/abs/1802.09766.

[7] W.-D. K. Ma, J. P. Lewis, and W. B. Kleijn, *The hsic bottleneck: Deep learning without back-propagation*, 2019. arXiv: 1908.01580 [cs.LG].

References III

[8] G. Desjardins, K. Simonyan, R. Pascanu, and K. Kavukcuoglu, *Natural neural networks*, 2015. arXiv: 1507.00210 [stat.ML].

[9] F. Nielsen, *An elementary introduction to information geometry*, 2018. arXiv: 1808.08271 [cs.LG].

[10] S.-i. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998. DOI: 10.1162/089976698300017746. eprint: https://doi.org/10.1162/089976698300017746. [Online]. Available: https://doi.org/10.1162/089976698300017746.

[11] J. S. Bell, "On the einstein podolsky rosen paradox," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, 2004 [1964], pp. 14–21.