The Variational Deficiency Bottleneck

Pradeep Kr. Banerjee
pradeep@mis.mpg.de

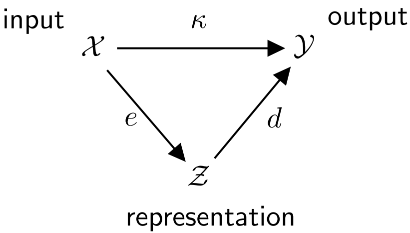
Joint work with Guido Montúfar (UCLA and MPI MiS)

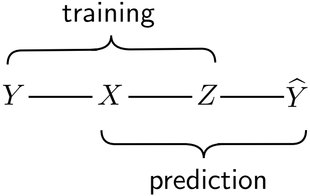
International Joint Conference on Neural Networks
WCCI 2020, Glasgow

Max-Planck-Institut für
Mathematik
in den Naturwissenschaften

Learning representations for classification

Training set: $(x^{(i)}, y^{(i)})_{i=1}^{N}$

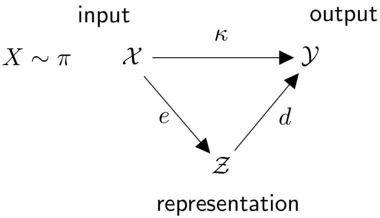




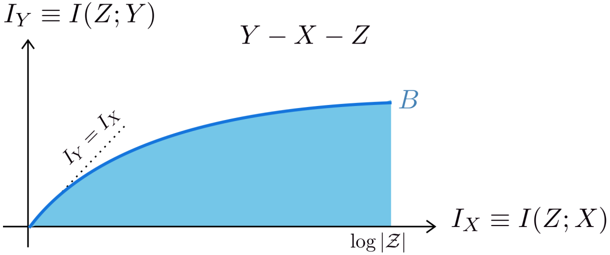
- X is an *input* variable (e.g., image) and Y an *output* variable of interest (e.g., label)
- The *channel* κ is unknown and only accessible through a training set
- Problem: Find a pair of stochastic maps (e, d) so that Z preserves as much relevant information as possible about the output (*sufficiency*) while maximally "compressing" the input (*minimality*)

The Information Bottleneck (IB)

$$(X, Y) \sim \pi \times \kappa$$



$$Q^0 = P_{XY} \times e_{Z|X}$$



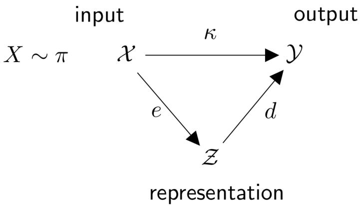
- **Approximate minimal sufficiency**: The IB maximizes

$$I_{Q^0}(Z;Y) \; - \beta I_{Q^0}(Z;X), \;\; \beta \in [0,1]$$

- **IB curve**: B is the (point-wise) smallest function for which

$$I(Z;Y) \leq B(I(Z;X)) \leq I(Z;X) \quad \text{for all } Y - X - Z$$

The Deficiency Bottleneck (DB)



$$(X, Y) \sim \pi \times \kappa$$

$$\widehat{\kappa}(y|x) = \int d(y|z)e(z|x)\, dz$$

- The deficiency of d w.r.t. κ is

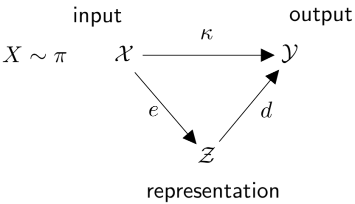
$$\delta^\pi(d, \kappa) := \min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} D_{KL}(\pi \times \kappa \| \pi \times \widehat{\kappa})$$

- The DB minimizes

$$\delta^\pi(d, \kappa) + \beta I(X; Z)$$

over all pairs (e, d), where $\beta \in [0, 1]$ is a regularization parameter

Deficiency and Input Blackwell Sufficiency



$$(X, Y) \sim \pi \times \kappa$$

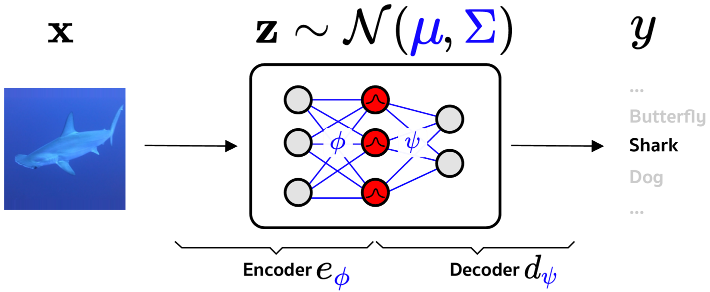
$$\widehat{\kappa}(y|x) = \int d(y|z)e(z|x)\,dz$$

$$\delta^{\pi}(d, \kappa) = \min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} D_{KL}(\pi \times \kappa \| \pi \times \widehat{\kappa})$$

Definition (Input Blackwell sufficiency [Bla53, Nas18])

Given two channels, $\kappa \in \mathsf{M}(\mathcal{X};\mathcal{Y})$ and $d \in \mathsf{M}(\mathcal{Z};\mathcal{Y})$, κ is input-degraded from d, denoted $d \succeq_{\mathcal{Y}} \kappa$, if $\kappa = \int d(y|z)e(z|x)\,dz$ for some $e \in \mathsf{M}(\mathcal{X};\mathcal{Z})$. We say that d is input Blackwell sufficient for κ if $d \succeq_{\mathcal{Y}} \kappa$.

The Variational Deficiency Bottleneck (VDB)



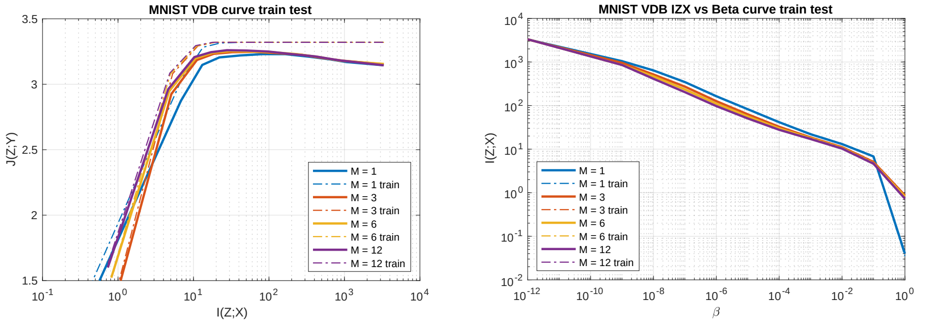
$$\mathcal{L} := \frac{1}{N}\sum_{i=1}^{N}\left[-\log\left(\frac{1}{M}\sum_{j=1}^{M}[d_\psi(y^{(i)}|f(x^{(i)},\epsilon^{(j)}))]\right) + \beta D(e_\phi(Z|x^{(i)})\|r(Z))\right]$$

$$e_\phi(z|x) = \mathcal{N}(z|f_e^\mu(x), f_e^\Sigma(x)) \qquad d_\psi(y|z) = \text{softmax}(y|f_d(z)) \qquad r(z) \sim \mathcal{N}(0, I)$$

For $M = 1$, we recover the Variational Information Bottleneck (VIB) objective [AFDM17]

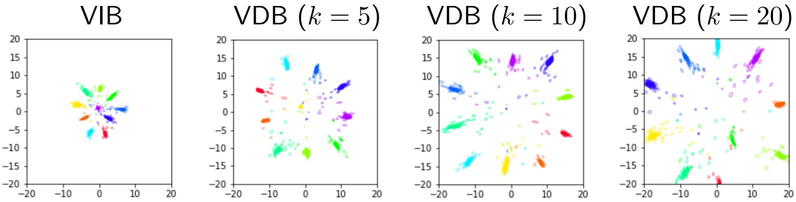
The VDB curve for MNIST

- In the VDB, "more sufficient" means "less deficient"
  $J(Z;Y) := H(Y) - \mathbb{E}_{(x,y)\sim p_{\mathcal{D}}(x)}\left[-\log \widehat{\kappa}(y|x)\right]$
- Use M encoder samples to compute the expectation inside the log. $J(Z;Y) = I(Z;Y)$ for $M = 1$



- For good values of β, higher values of M (our method) lead to a *smaller generalization gap* and *more compression of the input for the same level of sufficiency*

2D representations for MNIST



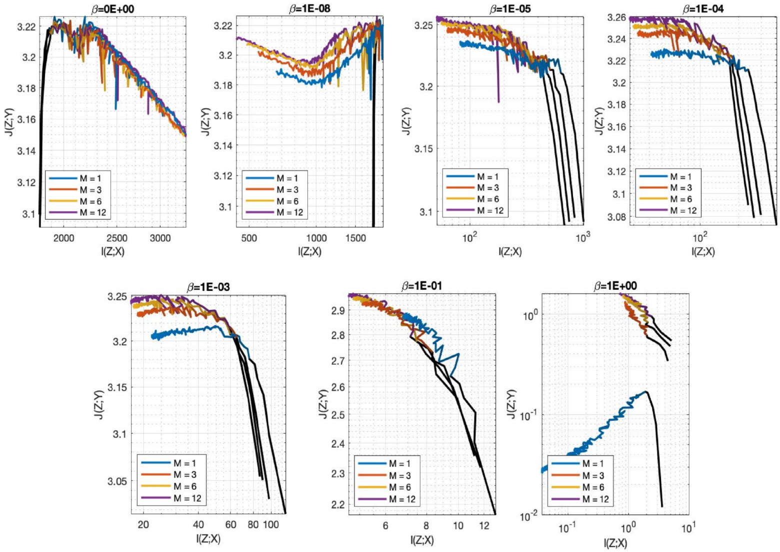
Posterior Gaussian distributions of 1000 test images from MNIST after training with the **VIB**, and the **VDB** with $k = 5, 10, 20$ encoder update steps per decoder update. $\beta = 10^{-3}$, $M = 1$. Colors correspond to the 10 different class labels.

- Nested optimization strategy to approximate the deficiency

$$\min_{d \in \mathsf{M}(\mathcal{Z};\mathcal{Y})} \left[\min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} \left[D(\pi \times \kappa \| \pi \times \widehat{\kappa}) + \beta D(\pi \times e \| \pi \times r) \right] \right], \quad \widehat{\kappa}(y|x) = \int d(y|z) e(z|x)\, dz$$

- Improved out-of-distribution robustness on MNIST-C [MG19] and CIFAR-10-C [HD19]

Information plane learning curves for MNIST



Conclusions

- A new bottleneck method for learning data representations based on *information deficiency*, rather than the more traditional *information sufficiency*
- VDB and VIB coincide in the regime of single-shot Monte Carlo approximations
- Training with the VDB improves out-of-distribution robustness over the VIB on two benchmark datasets, MNIST-C [MG19] and CIFAR-10-C [HD19]
- Unsupervised version of the VDB shares superficial similarities with the Importance Weighted Autoencoder (IWAE) [BGS16]

References

Alexander A. Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy.
Deep variational information bottleneck.
In *International Conference on Learning Representations*, 2017.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov.
Importance weighted autoencoders.
In *International Conference on Learning Representations*, 2016.

David Blackwell.
Equivalent comparisons of experiments.
The Annals of Mathematical Statistics, 24(2):265–272, 1953.

Dan Hendrycks and Thomas Dietterich.
Benchmarking neural network robustness to common corruptions and perturbations.
In *International Conference on Learning Representations*, 2019.

Norman Mu and Justin Gilmer.
MNIST-C: A robustness benchmark for computer vision.
arXiv preprint arXiv:1906.02337, 2019.

Rajai Nasser.
Characterizations of two channel orderings: Input-degradedness and the Shannon ordering.
IEEE Transactions on Information Theory, 64(10):6759–6770, 2018.

Ravid Shwartz-Ziv and Naftali Tishby.
Opening the black box of deep neural networks via information.
arXiv preprint arXiv:1703.00810, 2017.