Deep Learning and the Information Bottleneck Principle

CS396/496 – Deep Learning for Practitioners

Presenter: Alex Tang

Department of Computer Science, Northwestern University

Motivations

- Information theory plays an important part in ML studies on mainly two aspects

1. Stability
   - Replace-one (RO) stability guarantees generalization
   - Can be formalized with information-theoretic terms
2. Information bottleneck in DL
   - Generalizes lossy channel communication studies
   - Has a deep connection with pure statistics
   - Possiblity to be implemented for training a NN

- These two aspects are tightly related as well.

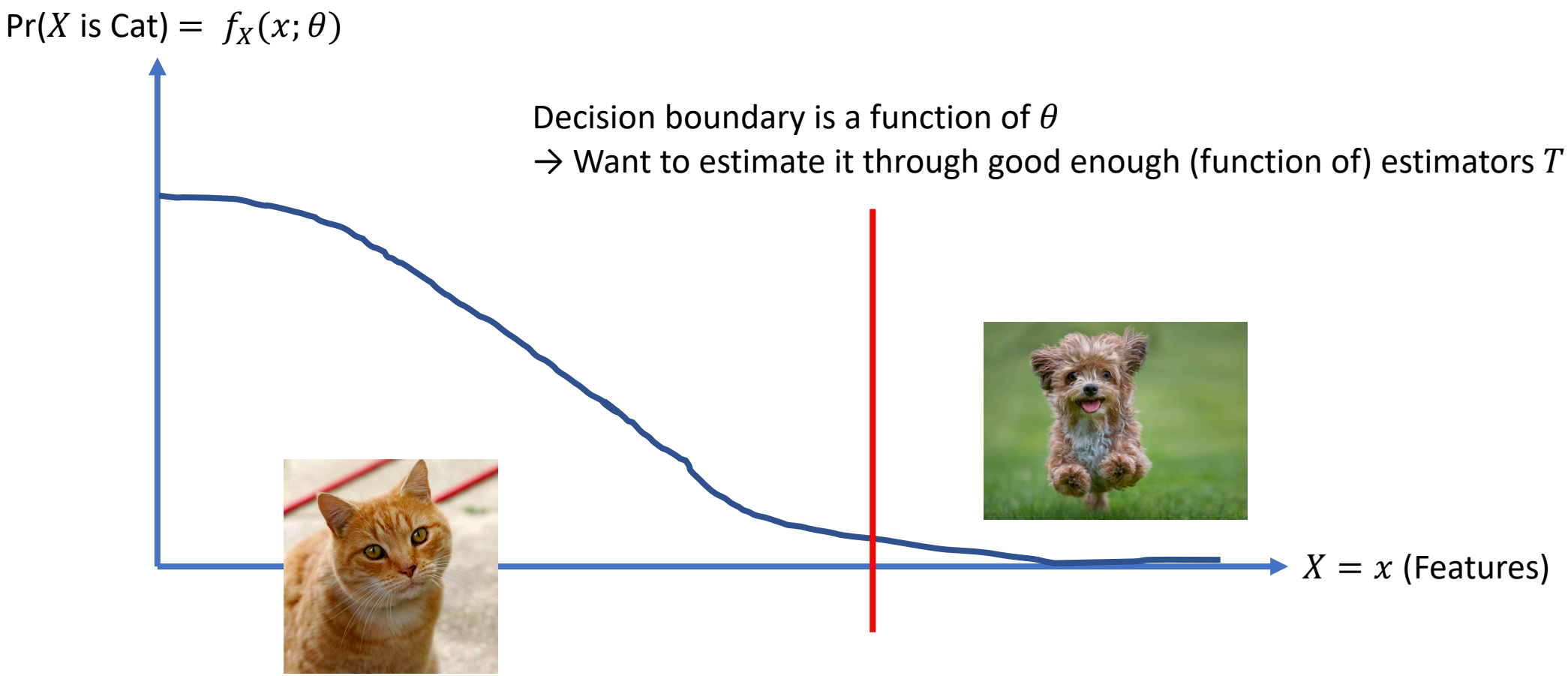
Preliminaries: Methametical Statistics

- Definition (Statistic)
  - For i.i.d. random variables $X_1, X_2, \ldots, X_n$, a statistic is a function $T(X_1, \ldots X_n)$
  - A statistic is free of parameters of the underlying distribution
  - Examples: Sample mean/variance, order statistic (i.e. sorted X_i), etc.

- Definition (Estimator)
  - Suppose $X_1, X_2, \ldots, X_n$ are drawn from a distribution with p.d.f. $f_X(x; \theta)$
  - We are interested in finding statistic T to estimate the true parameter θ
  - T is an estimator of θ
  - For continuous functions τ, $\tau(T)$ is an estimator of $\tau(\theta)$
  - Neural networks with fixed hyperparameters can be viewed as estimators too.

Preliminaries: Mathematical Statistics

- Why functions of parameters?



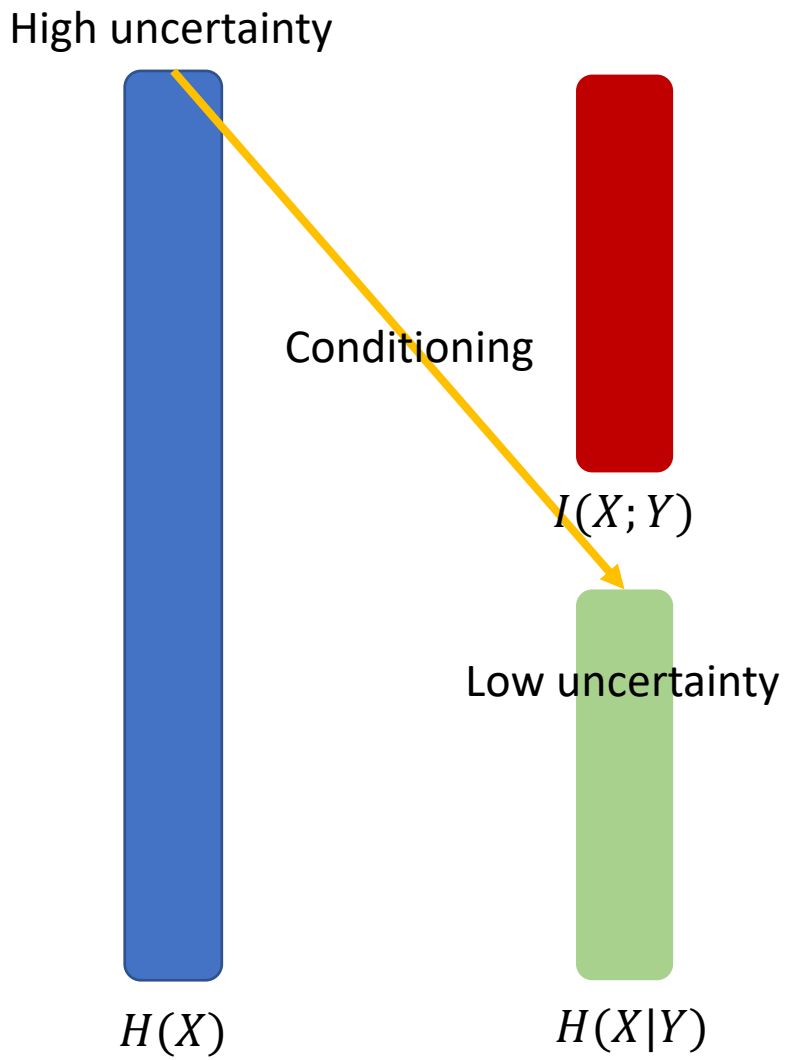
Preliminaries: Mathematical Statistics

- Finding good estimators is important in both statistics and ML.
- How to quantify its quality?
  - Bias, consistency, efficiency, completeness, etc.

- Definition (Sufficiency)
  - For a statistic T, if $f_X(X|T)$ is free of parameters θ, we say T is sufficient for X
  - If T is "most independent" of X, we say T is minimally sufficient for X
  - Interpretation:
    - Estimator T has as much information of θ as X has
    - T can be the result of performing dimensional reduction or compression on X in ML

Preliminaries: Information Theory

- Definition (Entropy)
  - For a discrete random variable X, $H(X) = E\left[\log \frac{1}{P(X)}\right]$
  - Measures the amount of "uncertainty"
- Definition (Mutual information)
  - For two discrete random variables X, Y
  - $I(X;Y) = H(X) - H(X|Y)$
  - Measures the amount of "dependency"



Preliminaries: Information Theory

- Properties of Mutual Information
  - $I(X;Y) = D_{KL}(P_{XY} \,||\, P_X P_Y)$
    - $D_{KL}(P_A \,||\, P_B) = \sum_x P_A(X = x) \log(\frac{P_A(X=x)}{P_B(X=x)})$ is the Kullback-Leibler divergence between P_A and P_B
  - $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$
    - Implies $I(X;Y) = I(Y;X) \rightarrow$ commutative operator
    - $I(X;X) = H(X)$
  - $0 \leq I(X;Y) \leq H(X)$
    - Equals 0 iff X, Y are independent
    - Equals $H(X)$ iff Y is a deterministic function of X

Preliminaries: Information Theory

- Definition (Markov Chain)
  - For sequence of random variables $X_1, X_2, \dots, X_n$, if $P(X_i = x \mid X_1, .., X_{i-1}) = P(X_i = x \mid X_{i-1})$, we say sequence $X_1, \dots, X_n$ is a Markov Chain
  - We denote such Markov Chain as $X_1 - X_2 \dots - X_n$

- Properties of Markov Chain $X - Y - Z$
  - $P_{Z|(X,Y)} = P_{Z|Y} \rightarrow P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$
  - $P_{XYZ} = P_X P_{Y|X} P_{Z|Y} \rightarrow P_{(X,Z)|Y} = P_{X|Y} P_{Z|Y}$

Preliminaries: Information Theory

- Properties of Mutual Information (Conditioning)
  - $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
    - $I(X;Y|Z) = 0$ iff X, Y are independent given Z (i.e. $P_{(X,Y)|Z} = P_{X|Z}P_{Y|Z}$)
      - X, Z, Y forms a Markov Chain $X - Z - Y$
  - $I(X;Y_1, Y_2, \dots, Y_n) = \sum_i I(X;Y_i \mid Y_1, Y_2, \dots, Y_{i-1})$

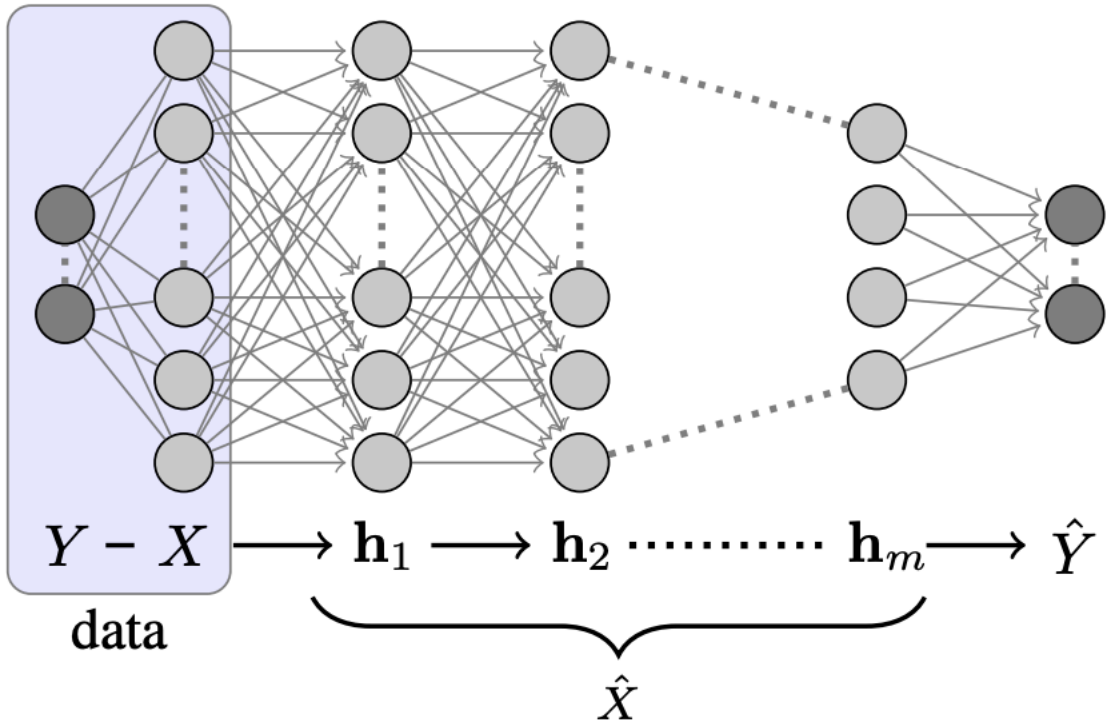
- Theorem (Data Processing Inequality, DPI)
  - For a Markov Chain $X - Y - Z$, it implies $I(X;Z|Y) = 0$
  - $I(X;Z) \leq I(X;Y)$
  - Interpretation: Data processing reduces dependency

Preliminaries: Statistics and Information

- Sufficient Statistics
  - For all statistics T, Markov Chain $\Theta - X - T$ always holds since $P_{T|(\Theta,X)} = P_{T|X}$ by definition of a statistic
  - For a sufficient statistic T, $P_{X|T}$ is indepedent of P_Θ (i.e. $P_{X|(T,\Theta)} = P_{X|T}$)
  - $\Theta - T - X$ is also a Markov Chain $\rightarrow I(\Theta;T) = I(\Theta;X)$ (DPI)
  - $T = \arg max_{T'}\ I(\Theta;T')$
- Minimal Sufficient Statistics
  - If T^* is minimal sufficient for X, then it is chosen from
    - $T^* = \arg min_T\ I(X;T)$ s.t. $I(\Theta;T) = max_T\ I(\Theta;T)$

Information Bottleneck

- Consider Y as $\tau(\Theta)$, $\hat{X}$ as T and $\hat{Y}$ as the predicted label
  - We have $Y - X - \hat{X} - \hat{Y}$
- Neural networks can be viewed as a Markov Chian



Information Bottleneck

- Consider the partial Markov Chain $Y - X - \hat{X}$ first
  - $I(Y;\hat{X}) \leq I(Y;X)$ by DPI
    - Remark: $I(Y;X)$ is an unknown fixed constant!
  - Want $I(Y;\hat{X})$ close to $I(Y;X)$ as much as possible so that $\hat{X}$ represents X well
    - Objective: maximize $I(Y;\hat{X})$
  - But we don't want $\hat{X}$ to be the same as X
    - Otherwise the NN is just memorizing $X \rightarrow$ overfitting!
    - Adjusted objective: minimize $I(X;\hat{X})$ while maximizing $I(Y;\hat{X})$
  - Observations:
    - When $I(Y;\hat{X})$ is maximized to be $I(Y;X)$, Markov Chain $Y - \hat{X} - X$ holds (DPI)
    - $\hat{X}$ is a sufficient statistic for X
    - If $\hat{X} = \arg\min I(X;\hat{X})$, $\hat{X}$ is the minimal sufficient statistic for X

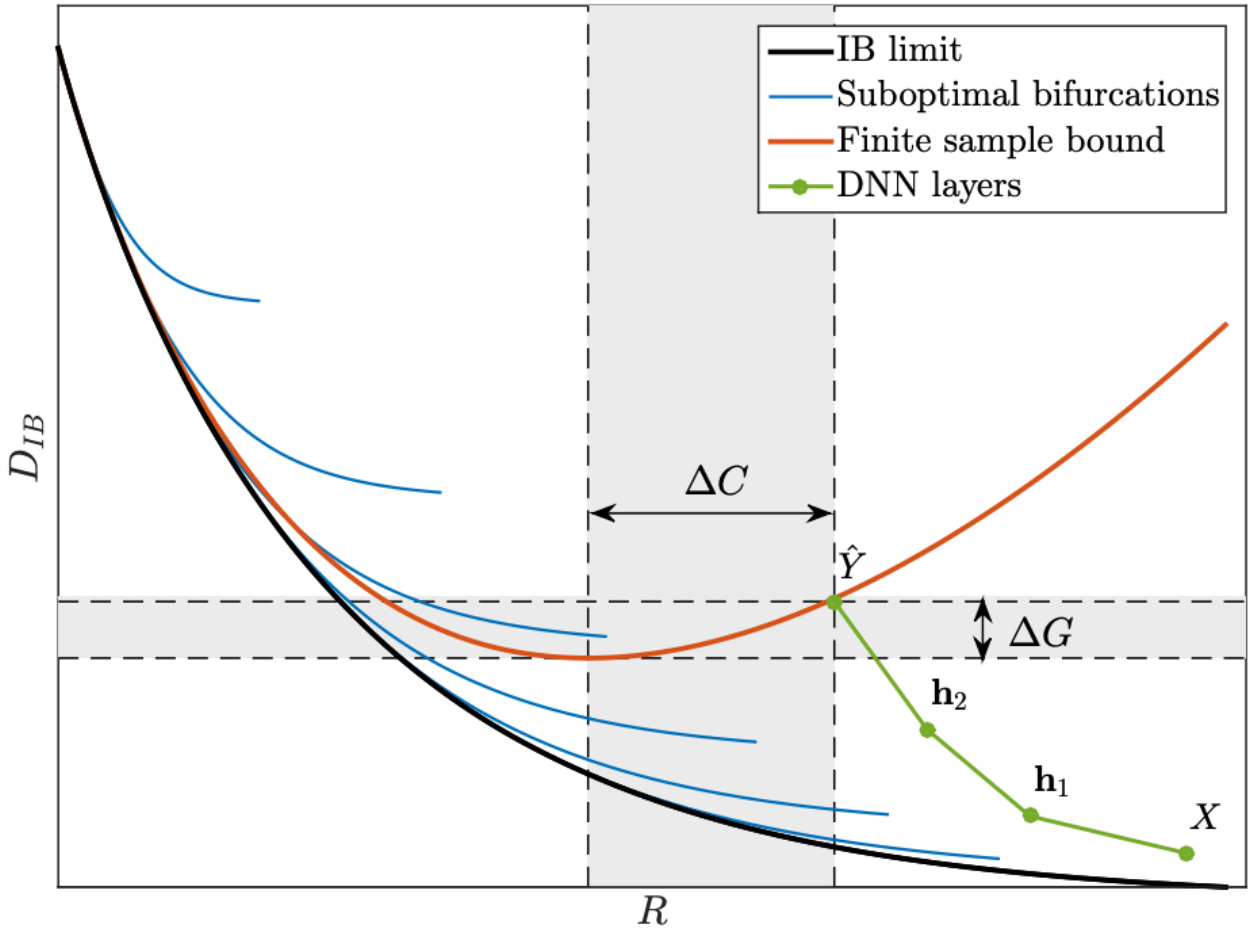
Information Bottleneck

- Goal: $\min_{\hat{X}} \mathrm{I}(\mathrm{X};\hat{X}) - \beta I(Y;\hat{X})$, $\beta \geq 0$ is the Lagrangian multiplier
  - Compare to training objective: $\min_{h \in H} \hat{L}(h) + \lambda R(h)$
  - Small β: minimizes $I(X;\hat{X}) \rightarrow$ Larger regularization term $\lambda R(h)$
  - Large β: maximizes $I(Y;\hat{X}) \rightarrow$ Smaller empirical risk term $\hat{L}(h)$
  - $I(X;\hat{X}) = R$ is the compression rate of X
- Equivalent to $\min_{\hat{X}} I(X;\hat{X}) + \beta' I(X;Y|\hat{X})$, $\beta' \geq 0$
  - Explicitly minimize $I(X;Y|\hat{X})$ so that $\hat{X}$ is almost sufficient for X
  - If $I(X;Y|\hat{X})$ is not 0, $\hat{X}$ doesn't capture all information between X, Y
  - $I(X;Y|\hat{X}) = E[\ell(X,\hat{X})] = D_{IB}$ is the expected distortion

Information Bottleneck

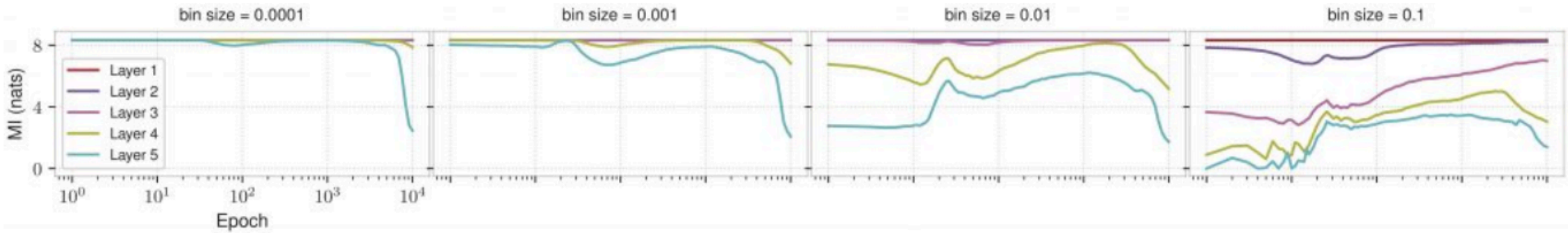
- An alternative way of training NNs
  - $\min_{T_i} I(X;T_i) - \beta_i I(Y;T_i)$
  - T_i is the i'th layer of a NN

- Connections with generalization:
  - $|L(h) - \hat{L}(h)| = O\left(\sqrt{\frac{I((X,Y);\hat{X})}{n}}\right)$
  - Note that $I(X;\hat{X}), I(Y;\hat{X}) \leq I((X,Y);\hat{X})$



Limits of Information Bottleneck

- Mutual information cannot be reliably estimated
  - NN is in practice deterministic instead of layers of random variables T_i
  - Requires finite cardinality $|X|, |Y|, \left|\hat{X}\right| < \infty$ to estimate empirical mutual information (Since true MI depends on the underlying distribution)
  - The gap between empirical MI an true MI scales with $O(\frac{|\hat{X}|}{\sqrt{n}})$



Possible solutions

- Estimating Information Flow in DNN [Goldfeld et al., ICML'19]
  - Add Gaussian noise at each layer of NN in both training and testing
  - New estimator of mutual information w/o quantizing neurons

- The HSIC Bottleneck: DL w/o Back-Propagation [Ma et al., AAAI'20]
  - Replace mutual information terms with normalized Hilbert-Schmidt independence criterion $nHSIC(X;\hat{X})$

Experiments

- Experiment of [Goldfeld'19]