Information-Theoretic Generalization Bounds for Deep Neural Networks

InfoCog Workshop @ NeurIPS 2023

Haiyun He, Christina Lee Yu, and Ziv Goldfeld

Cornell University, Center for Applied Mathematics

Cornell University

Outline

Key Takeaway

★ Goal:

capture the effects of depth in learning via information-theoretic generalization bounds

Key Takeaway

★ Goal:

capture the effects of depth in learning via information-theoretic generalization bounds

- **Result 1:** a hierarchical bound shrinks as the layer index increases

★ Goal:

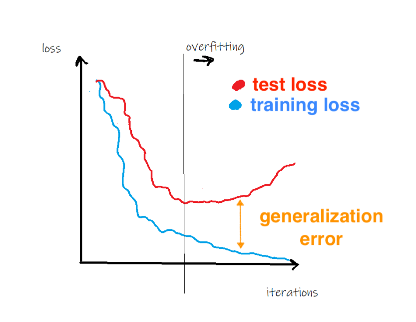
capture the effects of depth in learning via information-theoretic generalization bounds

- **Result 1:** a hierarchical bound shrinks as the layer index increases
- **Result 2:** quantifies the contraction when moving deeper into the network, via the strong data processing inequality (SDPI)

  $\Longrightarrow$ network depth, layer dimension, activation function, stochasticity

Cornell University

♠ Generalization error (in practice):



test loss=training loss+generalization error

- test loss: based on test data
- training loss: based on training data (usually small)

$\longrightarrow$ in theory, population risk/empirical risk

♦ Existing information-theoretic bounds:

not specialized to the DNN setting $\Longrightarrow$ did not capture the effect of depth on the generalization bound

Cornell University

▲ Supervised learning problem:

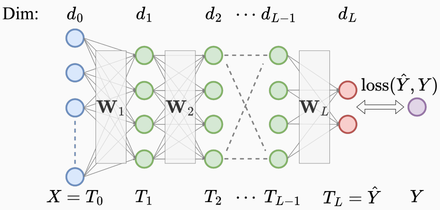


Figure 1: L-layer feedforward network

- Feedforward DNN model with L layers:

  $$\hat{Y} := g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X), \quad g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$$

  where $\phi_l : \mathbb{R} \to \mathbb{R}$ is the activation function;
- l^{th} internal representation: $T_l := g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$

Cornell University

▲ Supervised learning problem:

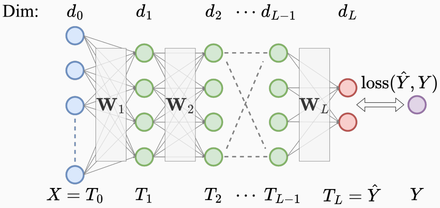


Figure 1: L-layer feedforward network

- Feedforward DNN model with L layers:

  $$\hat{Y} := g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X), \quad g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$$

  where $\phi_l : \mathbb{R} \to \mathbb{R}$ is the activation function;
- l^{th} internal representation: $T_l := g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$
- Loss function $\ell(\mathbf{w}, x, y) = \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$

Cornell University

▲ Supervised learning problem:

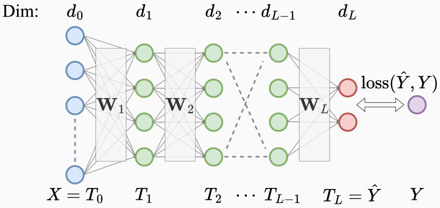


Figure 1: L-layer feedforward network

- Feedforward DNN model with L layers:

  $\hat{Y} := g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X), \quad g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$

  where $\phi_l : \mathbb{R} \to \mathbb{R}$ is the activation function;
- l^{th} internal representation: $T_l := g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$
- Loss function $\ell(\mathbf{w}, x, y) = \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$
- Label set $\mathcal{Y} = [K] \subseteq \mathbb{Z}_+$ (or $\mathbb{R}$ for regression)

Cornell University

▲ Supervised learning problem:

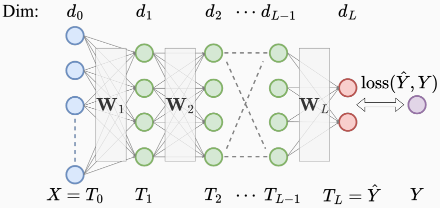


Figure 1: L-layer feedforward network

- Feedforward DNN model with L layers:
  $$\hat{Y} := g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X), \quad g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$$
  where $\phi_l : \mathbb{R} \to \mathbb{R}$ is the activation function;
- l^{th} internal representation: $T_l := g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$
- Loss function $\ell(\mathbf{w}, x, y) = \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$
- Label set $\mathcal{Y} = [K] \subseteq \mathbb{Z}_+$ (or $\mathbb{R}$ for regression)
- Training dataset: $D_n = \{(X_i, Y_i)\}_{i=1}^n$, identically $\sim P_{X,Y}$

Cornell University

▲ Supervised learning problem:

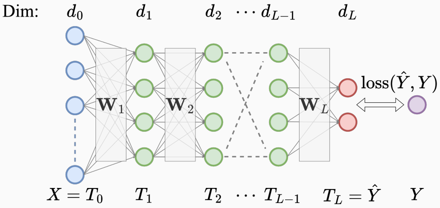


Figure 1: L-layer feedforward network

- Feedforward DNN model with L layers:

$$\hat{Y} := g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X), \quad g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$$

where $\phi_l : \mathbb{R} \to \mathbb{R}$ is the activation function;

- l^{th} internal representation: $T_l := g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$
- Loss function $\ell(\mathbf{w}, x, y) = \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$
- Label set $\mathcal{Y} = [K] \subseteq \mathbb{Z}_+$ (or $\mathbb{R}$ for regression)
- Training dataset: $D_n = \{(X_i, Y_i)\}_{i=1}^n$, identically $\sim P_{X,Y}$

- Expected generalization error:

$$\text{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) := \mathbb{E}[\underbrace{\mathcal{L}_{\mathsf{P}}(\mathbf{W}, P_{X,Y})}_{\text{Population Risk}} - \underbrace{\mathcal{L}_{\mathsf{E}}(\mathbf{W}, D_n)}_{\text{Empirical Risk}}]$$

$$:= \mathbb{E}\left[\mathbb{E}\left[\ell(\mathbf{W}, X, Y) - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{W}, X_i, Y_i) \middle| \mathbf{W}\right]\right]$$

Generalization Bound via DPI

Cornell University

▲ Data-processing inequality (DPI) for f-divergences D_f:

$$P_X,\, Q_X \in \mathcal{P}(\mathcal{X}) \to \boxed{P_{Y|X}} \to P_Y = P_{Y|X} \circ P_X,\, Q_Y = P_{Y|X} \circ Q_X$$

DPI: $\mathsf{D}_f(P_Y \| Q_Y) \leq \mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X) \leq \mathsf{D}_f(P_X \| Q_X)$

▲ Data-processing inequality (DPI) for f-divergences D_f:

$$P_X,\ Q_X \in \mathcal{P}(\mathcal{X}) \to \boxed{P_{Y|X}} \to P_Y = P_{Y|X} \circ P_X,\ Q_Y = P_{Y|X} \circ Q_X$$

DPI: $$\mathsf{D}_f(P_Y \| Q_Y) \le \mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X) \le \mathsf{D}_f(P_X \| Q_X)$$

♠ In DNN:

$$\begin{aligned} \mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) &:= \mathbb{E}\left[\mathbb{E}\left[\ell(\mathbf{W}, X, Y) - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{W}, X_i, Y_i) \middle| \mathbf{W}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\ell}(g_{\mathbf{W}_{l+1}^L}(T_l), Y) - \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(g_{\mathbf{W}_{l+1}^L}(T_{l,i}), Y_i) \middle| \mathbf{W}_1^l\right]\right] \qquad (l = 1, \ldots, L) \end{aligned}$$

▲ Data-processing inequality (DPI) for f-divergences D_f:

$$P_X,\ Q_X \in \mathcal{P}(\mathcal{X}) \to \boxed{P_{Y|X}} \to P_Y = P_{Y|X} \circ P_X,\ Q_Y = P_{Y|X} \circ Q_X$$

DPI: $\mathsf{D}_f(P_Y \| Q_Y) \le \mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X) \le \mathsf{D}_f(P_X \| Q_X)$

♠ In DNN:

$$\begin{aligned}\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) &:= \mathbb{E}\left[\mathbb{E}\left[\ell(\mathbf{W}, X, Y) - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{W}, X_i, Y_i)\middle|\mathbf{W}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\ell}(g_{\mathbf{W}_{l+1}^{L}}(T_l), Y) - \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(g_{\mathbf{W}_{l+1}^{L}}(T_{l,i}), Y_i)\middle|\mathbf{W}_1^l\right]\right] \qquad (l = 1, \dots, L)\end{aligned}$$

Conditioned on $\mathbf{W}_1^l$,

$$T_{l-1,i}, T_{l-1} \to \boxed{g_{\mathbf{W}_l}(\cdot)} \to T_{l,i}, T_l$$

Generalization Bound via DPI

Cornell University

Theorem 1 (Hierarchical bound)

If the loss function $\ell(\mathbf{w}, X, Y)$ *is* σ*-sub-Gaussian under* $P_{X,Y}$*, for all* $\mathbf{w} \in \mathcal{W}$*. We have*

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \mathsf{UB}(L) \leq \mathsf{UB}(L-1) \leq \ldots \leq \underbrace{\mathsf{UB}(0)}_{\textit{existing bound}},$$

where

$$\mathsf{UB}(0) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathsf{I}(X_i, Y_i; \mathbf{W})},$$

$$\mathsf{UB}(l) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L}|\mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}\left(P_{T_{l,i}, Y_i|\mathbf{W}_1^l} \middle\| P_{T_l, Y|\mathbf{W}_1^l} \middle| P_{\mathbf{W}_1^l}\right)}, \quad l = 1, \ldots, L.$$

Cornell University

Theorem 1 (Hierarchical bound)

If the loss function $\ell(\mathbf{w}, X, Y)$ is σ-sub-Gaussian under $P_{X,Y}$, for all $\mathbf{w} \in \mathcal{W}$. We have

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \mathsf{UB}(L) \leq \mathsf{UB}(L-1) \leq \ldots \leq \underbrace{\mathsf{UB}(0)}_{\textit{existing bound}},$$

where

$$\mathsf{UB}(0) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathsf{I}(X_i, Y_i; \mathbf{W})},$$

$$\mathsf{UB}(l) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L}|\mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}\left(P_{T_{l,i}, Y_i|\mathbf{W}_1^l} \middle\| P_{T_l, Y|\mathbf{W}_1^l} \middle| P_{\mathbf{W}_1^l}\right)}, \quad l = 1, \ldots, L.$$

▲ Remarks:

1. **Interpretation:** The model generalizes when
   - Subsequent layers do not strongly depend on the input internal representation
   - Learned posterior of internal representation matches the prior

Cornell University

Theorem 1 (Hierarchical bound)

If the loss function $\ell(\mathbf{w}, X, Y)$ is σ-sub-Gaussian under $P_{X,Y}$, for all $\mathbf{w} \in \mathcal{W}$. We have

$$|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \mathrm{UB}(L) \leq \mathrm{UB}(L-1) \leq \ldots \leq \underbrace{\mathrm{UB}(0)}_{\text{existing bound}},$$

where

$$\mathrm{UB}(0) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathrm{I}(X_i, Y_i; \mathbf{W})},$$

$$\mathrm{UB}(l) = \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\mathrm{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L} | \mathbf{W}_1^l) + \mathrm{D}_{\mathrm{KL}}(P_{T_{l,i}, Y_i | \mathbf{W}_1^l} \| P_{T_l, Y | \mathbf{W}_1^l} | P_{\mathbf{W}_1^l})}, \quad l = 1, \ldots, L.$$

▲ Remarks:

2. **Special cases (Discrete latent space):** when T_l is finite (e.g., the VQ-VAE)

$\min P_{T_l, Y | \mathbf{W}_1^l} \in (0, |\mathcal{T}_l \times \mathcal{Y}|^{-1})$ higher $\Rightarrow$ posterior with higher entropy/variance

$\Rightarrow$ smaller $\mathrm{UB}(l)$ and generalization error $\Rightarrow$ stochasticity helps

Strong Data-Processing Inequality (SDPI)

- Quantify the contraction from $\text{UB}(l-1)$ to $\text{UB}(l)$:

$$\text{UB}(L) \leq \text{coeff}_L \text{UB}(L-1) \leq \cdots \leq \prod_{l=1}^{L} \text{coeff}_l \text{UB}(0)$$

♠ Quantify the contraction from $\mathsf{UB}(l-1)$ to $\mathsf{UB}(l)$:

$$\mathsf{UB}(L) \leq \mathsf{coeff}_L \mathsf{UB}(L-1) \leq \cdots \leq \prod_{l=1}^{L} \mathsf{coeff}_l \mathsf{UB}(0)$$

The SDPI contraction coefficient for $P_{Y|X}$ under some f-divergence ($P_X \ll Q_X$):

$$\eta_f(P_{Y|X}) := \sup_{P_X, Q_X} \frac{\mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X)}{\mathsf{D}_f(P_X \| Q_X)} \in [0,1].$$

Cornell University

♠ Quantify the contraction from $\mathsf{UB}(l-1)$ to $\mathsf{UB}(l)$:

$$\mathsf{UB}(L) \le \mathsf{coeff}_L \mathsf{UB}(L-1) \le \cdots \le \prod_{l=1}^{L} \mathsf{coeff}_l \mathsf{UB}(0)$$

The SDPI contraction coefficient for $P_{Y|X}$ under some f-divergence ($P_X \ll Q_X$):

$$\eta_f(P_{Y|X}) := \sup_{P_X, Q_X} \frac{\mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X)}{\mathsf{D}_f(P_X \| Q_X)} \in [0,1].$$

⋆ **Properties:**

- $\eta_f(P_{Y|X}) \le \eta_{\mathsf{TV}}(P_{Y|X}) = \sup_{x,x' \in \mathcal{X}} \|P_{Y|X=x} - P_{Y|X=x'}\|_{\mathsf{TV}}$ *(Dobrushin's coefficient)*

Strong Data-Processing Inequality (SDPI)

Cornell University

♠ Quantify the contraction from $\mathrm{UB}(l-1)$ to $\mathrm{UB}(l)$:

$$\mathrm{UB}(L) \leq \mathrm{coeff}_L \mathrm{UB}(L-1) \leq \cdots \leq \prod_{l=1}^{L} \mathrm{coeff}_l \mathrm{UB}(0)$$

The SDPI contraction coefficient for $P_{Y|X}$ under some f-divergence ($P_X \ll Q_X$):

$$\eta_f(P_{Y|X}) := \sup_{P_X, Q_X} \frac{\mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X)}{\mathsf{D}_f(P_X \| Q_X)} \in [0,1].$$

⋆ **Properties:**

- $\eta_f(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X}) = \sup_{x,x' \in \mathcal{X}} \|P_{Y|X=x} - P_{Y|X=x'}\|_{\mathsf{TV}}$ *(Dobrushin's coefficient)*
- $g(\cdot)$ deterministic: $\eta_f(P_{g(X)|X}) = \eta_{\mathsf{TV}}(P_{g(X)|X}) = 1$

Strong Data-Processing Inequality (SDPI)

Cornell University

♠ Quantify the contraction from $\mathsf{UB}(l-1)$ to $\mathsf{UB}(l)$:

$$\mathsf{UB}(L) \le \mathsf{coeff}_L \mathsf{UB}(L-1) \le \cdots \le \prod_{l=1}^{L} \mathsf{coeff}_l \mathsf{UB}(0)$$

The SDPI contraction coefficient for $P_{Y|X}$ under some f-divergence ($P_X \ll Q_X$):

$$\eta_f(P_{Y|X}) := \sup_{P_X, Q_X} \frac{\mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X)}{\mathsf{D}_f(P_X \| Q_X)} \in [0,1].$$

⋆ **Properties:**

- $\eta_f(P_{Y|X}) \le \eta_{\mathsf{TV}}(P_{Y|X}) = \sup_{x,x' \in \mathcal{X}} \|P_{Y|X=x} - P_{Y|X=x'}\|_{\mathsf{TV}}$ *(Dobrushin's coefficient)*
- $g(\cdot)$ deterministic: $\eta_f(P_{g(X)|X}) = \eta_{\mathsf{TV}}(P_{g(X)|X}) = 1$

$\Longrightarrow$ if all the feature maps $g_{\mathbf{w}_l}$ in the DNN are deterministic $\longrightarrow$ the SDPI coeff $= 1$

Tighter Bound via Contraction

♠ Train neural network with noise $\Rightarrow$ enhance generalization, improve robustness:

Tighter Bound via Contraction

♠ Train neural network with noise ⇒ enhance generalization, improve robustness:

- Additive noise: Gaussian/Laplace/Salt and Pepper/...
- Dropout
- DropConnect
- Data Augmentation: rotating/flipping/scaling/cropping images/MixUp...
- Label Smoothing: adding noise to label
- . . .

♠ Train neural network with noise $\Rightarrow$ enhance generalization, improve robustness:

- Additive noise: Gaussian /Laplace/Salt and Pepper/...
- Dropout
- DropConnect
- Data Augmentation: rotating/flipping/scaling/cropping images/MixUp...
- Label Smoothing: adding noise to label
- . . .

▲ **Noisy DNN model:** feature map at each layer is perturbed by isotropic Gaussian noise, i.e.,

$$\widetilde{T}_l = T_l + \epsilon_l Z_l = \phi_l(\mathbf{W}_l \widetilde{T}_{l-1}) + \epsilon_l Z_l, \quad l = 1, \ldots, L,$$

where $\phi_l(\cdot)$ is the activation function, $Z_l \sim N(0, \mathbf{I}_{d_l})$ is independent and $\epsilon_l \in \mathbb{R}_+$ is a constant.

[1] Goldfeld et al., Estimating information flow in deep neural network. ICML 2019

Tighter Bound via Contraction

▲ **Noisy DNN model:** feature map at each layer is perturbed by isotropic Gaussian noise, i.e.,

$$\widetilde{T}_l = T_l + \epsilon_l Z_l = \phi_l(\mathbf{W}_l \widetilde{T}_{l-1}) + \epsilon_l Z_l, \quad l = 1, \ldots, L,$$

where $\phi_l(\cdot)$ is the activation function, $Z_l \sim N(0, \mathbf{I}_{d_l})$ is independent and $\epsilon_l \in \mathbb{R}_+$ is a constant.

$\Rightarrow$ stochastic approximation of deterministic DNN [1]

[1] Goldfeld et al., Estimating information flow in deep neural network. ICML 2019

▲ **Noisy DNN model:** feature map at each layer is perturbed by isotropic Gaussian noise, i.e.,

$$\widetilde{T}_l = T_l + \epsilon_l Z_l = \phi_l(\mathbf{W}_l \widetilde{T}_{l-1}) + \epsilon_l Z_l, \quad l = 1, \dots, L,$$

where $\phi_l(\cdot)$ is the activation function, $Z_l \sim N(0, \mathbf{I}_{d_l})$ is independent and $\epsilon_l \in \mathbb{R}_+$ is a constant.

$\Rightarrow$ stochastic approximation of deterministic DNN [1]

Lemma 1 (SDPI coefficient bound)

Let $X \sim P_X \in \mathcal{P}(\mathbb{R}^{d_x})$ and consider a bounded function $g : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$. Set $Y = g(X) + \epsilon N$, where $\epsilon > 0$ and $N \sim \mathcal{N}(0, \mathbf{I}_{d_y})$ is independent of X. The SDPI coefficient of the induced channel $P_{Y|X}$ satisfies

$$\eta_f(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X}) \leq 1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_y}\|g\|_\infty}{2\epsilon}\right),$$

where $\mathsf{Q}(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t$ is the Gaussian complimentary CDF.

[1] Goldfeld et al., Estimating information flow in deep neural network. ICML 2019

Tighter Bound via Contraction

Cornell University

Theorem 2 (Noisy DNN generalization bound)

Consider the noisy DNN model with bounded activation functions $\phi_l, l = 1, \ldots, L$.

$$\left|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})\right| \le \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\left(1 - 2\mathrm{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right)\mathsf{I}(X_i;\mathbf{W}|Y_i) + \underbrace{\mathsf{I}(Y_i;\mathbf{W})}_{\text{no SDPI}}}.$$

Cornell University

Theorem 2 (Noisy DNN generalization bound)

Consider the noisy DNN model with bounded activation functions $\phi_l, l = 1, \ldots, L$.

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\left(1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right)\mathsf{I}(X_i;\mathbf{W}|Y_i) + \underbrace{\mathsf{I}(Y_i;\mathbf{W})}_{\textit{no SDPI}}}.$$

▲ **Remarks:**

1. $\mathsf{I}(Y_i;\mathbf{W})$ factored out $\leftarrow$ the label is not processed by the noisy DNN. $(\mathsf{I}(Y_i;\mathbf{W}) \leq \log K)$

Tighter Bound via Contraction

Cornell University

Theorem 2 (Noisy DNN generalization bound)

Consider the noisy DNN model with bounded activation functions $\phi_l, l = 1, \ldots, L$.

$$|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\left(1 - 2\mathrm{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right)\mathrm{I}(X_i;\mathbf{W}|Y_i) + \underbrace{\mathrm{I}(Y_i;\mathbf{W})}_{\textit{no SDPI}}}.$$

▲ Remarks:

1. $\mathrm{I}(Y_i;\mathbf{W})$ factored out $\leftarrow$ the label is not processed by the noisy DNN. $(\mathrm{I}(Y_i;\mathbf{W}) \leq \log K)$
2. $\|\phi_l\|_\infty = 1$ if $\phi_l \in \{\mathrm{sigmoid}, \mathrm{softmax}, \mathrm{tanh}\}$

Theorem 2 (Noisy DNN generalization bound)

Consider the noisy DNN model with bounded activation functions $\phi_l, l = 1, \ldots, L$.

$$|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\left(1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right)\mathsf{I}(X_i;\mathbf{W}|Y_i) + \underbrace{\mathsf{I}(Y_i;\mathbf{W})}_{\textit{no SDPI}}}.$$

▲ **Remarks:**

1. $\mathsf{I}(Y_i;\mathbf{W})$ factored out $\leftarrow$ the label is not processed by the noisy DNN. $(\mathsf{I}(Y_i;\mathbf{W}) \leq \log K)$
2. $\|\phi_l\|_\infty = 1$ if $\phi_l \in \{\mathrm{sigmoid}, \mathrm{softmax}, \tanh\}$
3. **Observation:** with fixed noise level,

   $d_l \downarrow$ & $L \uparrow \Rightarrow$ SDPI coeff $\downarrow$ from 1 to $0 \Rightarrow$ generalization error $\downarrow$

Cornell University

Theorem 2 (Noisy DNN generalization bound)

Consider the noisy DNN model with bounded activation functions $\phi_l, l = 1, \ldots, L$.

$$|\text{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\left(1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right)\mathsf{I}(X_i;\mathbf{W}|Y_i) + \underbrace{\mathsf{I}(Y_i;\mathbf{W})}_{\text{no SDPI}}}.$$

▲ **Remarks:**

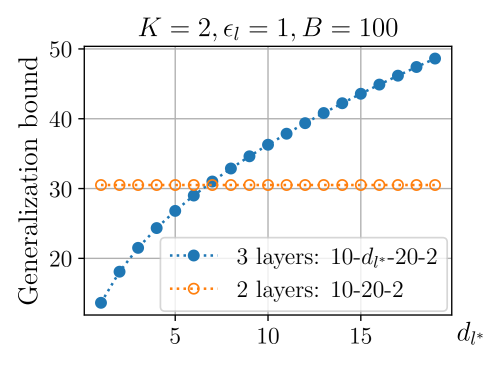
1. $\mathsf{I}(Y_i;\mathbf{W})$ factored out $\leftarrow$ the label is not processed by the noisy DNN. $(\mathsf{I}(Y_i;\mathbf{W}) \leq \log K)$
2. $\|\phi_l\|_\infty = 1$ if $\phi_l \in \{\text{sigmoid}, \text{softmax}, \tanh\}$
3. **Observation:** with fixed noise level,

   $d_l \downarrow$ & $L \uparrow \Rightarrow$ SDPI coeff $\downarrow$ from 1 to $0 \Rightarrow$ generalization error $\downarrow \Rightarrow$ benefit of depth and stochasticity

Tighter Bound via Contraction

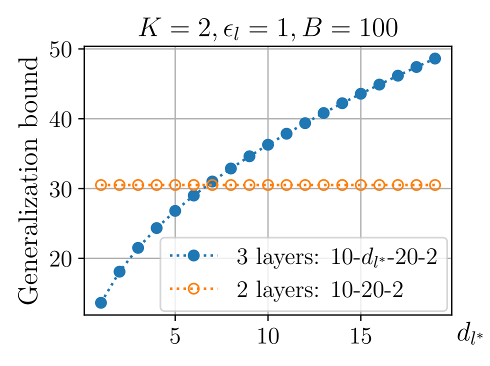
Simple example: Finite DNN parameter space $\mathcal{W} = [B]^{d_1 \times d_0} \times \cdots \times [B]^{d_L \times d_{L-1}}$ for some $B \in \mathbb{Z}_+$. Then

$$\mathsf{I}(X_i; \mathbf{W} | Y_i) \leq \mathsf{H}(\mathbf{W}) \leq \sum_{l=1}^{L} d_l d_{l-1} \log B.$$



Simple example: Finite DNN parameter space $\mathcal{W} = [B]^{d_1 \times d_0} \times \cdots \times [B]^{d_L \times d_{L-1}}$ for some $B \in \mathbb{Z}_+$. Then

$$\mathsf{I}(X_i; \mathbf{W}|Y_i) \leq \mathsf{H}(\mathbf{W}) \leq \sum_{l=1}^{L} d_l d_{l-1} \log B.$$



A deep but narrower network may generalize better.

(Requires further explorations.)

Tighter Bound via Contraction

▲ **DNN with Dropout:** l^{th} layer Dropout prob $\delta_l \in [0, 1] \Rightarrow$ activation output of the $(l+1)^{\text{th}}$ layer

$$T_{l+1} = \phi_{l+1}(\mathbf{W}_{l+1}(T_l \odot E_l)) =: \phi_{l+1}(\mathbf{W}_{l+1}\widetilde{T}_l), \quad l = 0, \ldots, L$$

where $E_l \sim \text{Bern}(1 - \delta_l)^{d_l}$ is independent and $\odot$ denotes the elementwise product operation.

Tighter Bound via Contraction

▲ **DNN with Dropout:** l^{th} layer Dropout prob $\delta_l \in [0,1] \Rightarrow$ activation output of the $(l+1)^{\text{th}}$ layer

$$T_{l+1} = \phi_{l+1}(\mathbf{W}_{l+1}(T_l \odot E_l)) =: \phi_{l+1}(\mathbf{W}_{l+1}\widetilde{T}_l), \quad l = 0, \ldots, L$$

where $E_l \sim \text{Bern}(1-\delta_l)^{d_l}$ is independent and $\odot$ denotes the elementwise product operation.

The Markov chain $T_l \to \widetilde{T}_l \to T_{l+1}$:

$$P_{T_{l+1}|\widetilde{T}_l,\mathbf{W}} \text{ -- deterministic,} \quad P_{\widetilde{T}_l|T_l} \text{ -- } d_l \text{ parallel } Z\text{-channel}$$

Cornell University

Lemma 2 Dropout SDPI coefficient

SDPI coefficient for Dropout channel with parameter δ_l and dimension d_l is $\eta_{\mathsf{KL}}(P_{\widetilde{T}_l|T_l}) = 1 - \delta_l^{d_l}$, for $l = 0, \ldots, L$.

Cornell University

Lemma 2 Dropout SDPI coefficient

SDPI coefficient for Dropout channel with parameter δ_l and dimension d_l is $\eta_{\mathsf{KL}}(P_{\widetilde{T}_l|T_l}) = 1 - \delta_l^{d_l}$, for $l = 0, \ldots, L$.

Theorem 3 (DNN with Dropout generalization bound)

Consider the DNN model with Dropout rate $\delta_l \in [0,1]$, $l = 0, \ldots, L-1$. *If the loss function* $\ell(\mathbf{w}, X, Y)$ *is* σ*-sub-Gaussian, we have*

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n} \sum_{i=1}^{n} \sqrt{\prod_{l=0}^{L-1} (1 - \delta_l^{d_l}) \mathsf{I}(X_i; \mathbf{W}|Y_i) + \mathsf{I}(Y_i; \mathbf{W})}.$$

Tighter Bound via Contraction

Cornell University

Lemma 2 Dropout SDPI coefficient

SDPI coefficient for Dropout channel with parameter δ_l and dimension d_l is $\eta_{\mathsf{KL}}(P_{\widetilde{T}_l|T_l}) = 1 - \delta_l^{d_l}$, for $l = 0, \ldots, L$.

Theorem 3 (DNN with Dropout generalization bound)

Consider the DNN model with Dropout *rate* $\delta_l \in [0, 1]$, $l = 0, \ldots, L-1$. *If the loss function* $\ell(\mathbf{w}, X, Y)$ *is* σ*-sub-Gaussian, we have*

$$|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=0}^{L-1}(1-\delta_l^{d_l})\mathsf{I}(X_i;\mathbf{W}|Y_i) + \mathsf{I}(Y_i;\mathbf{W})}.$$

In addition:

$\mathsf{I}(X_i;\mathbf{W}|Y_i) + \mathsf{I}(Y_i;\mathbf{W})$ monotonically shrink to 0 as the input Dropout rate δ_0 increases from 0 to 1.

▲ Wasserstein Generalization Bound

for $p \in \mathbb{Z}_+$ and $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$:   (no DPI)

$$\mathbb{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')^p] \right)^{1/p},$$

where $\Pi(\mu, \nu)$ denotes the set of couplings on $\mathcal{X}^2$ with marginals μ and ν.

▲ Wasserstein Generalization Bound

for $p \in \mathbb{Z}_+$ and $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$: (no DPI)

$$\mathbb{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x') \sim \pi} [c(x, x')^p] \right)^{1/p},$$

where $\Pi(\mu, \nu)$ denotes the set of couplings on $\mathcal{X}^2$ with marginals μ and ν.

⋆ Property: $p \leq q$, $\mathbb{W}_p(\cdot, \cdot) \leq \mathbb{W}_q(\cdot, \cdot) \Rightarrow$ here we consider $\mathbb{W}_1$

▲ Wasserstein Generalization Bound

for $p \in \mathbb{Z}_+$ and $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$: (no DPI)

$$\mathbb{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')^p] \right)^{1/p},$$

where $\Pi(\mu, \nu)$ denotes the set of couplings on $\mathcal{X}^2$ with marginals μ and ν.

⋆ Property: $p \leq q$, $\mathbb{W}_p(\cdot, \cdot) \leq \mathbb{W}_q(\cdot, \cdot) \Rightarrow$ here we consider $\mathbb{W}_1$

Theorem 4 (Min Wasserstein generalization bound)

Suppose that the loss function $\tilde{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$ is ρ_0-Lipschitz and the activation function $\phi_l(\cdot)$ is ρ_l-Lipschitz, $l = 1, \ldots, L$. Let $\tilde{\rho}_l = \max\{\rho_0, \rho_0 \prod_{j=l+1}^{L} \rho_j\}$. We have

$$|\mathrm{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \min_{l=0,\ldots,L} \frac{\tilde{\rho}_l}{n} \sum_{i=1}^{n} \mathbb{W}_1(P_{T_{l,i}, Y_i|\mathbf{W}}, P_{T_{l,Y}|\mathbf{W}} | P_{\mathbf{W}}).$$

where $\mathbb{W}_1(P_{T_{l,i}, Y_i|\mathbf{W}}, P_{T_{l,Y}|\mathbf{W}} | P_{\mathbf{W}}) = \mathbb{E}[\mathbb{W}_1(P_{T_{l,i}, Y_i|\mathbf{W}}, P_{T_{l,Y}|\mathbf{W}})]$.

Thanks for listening!

Q & A