

Information Theoretic Analysis For Transfer Learning

Xuetong Wu¹ Jonathan H. Manton¹ Uwe Aickelin² Jingge Zhu¹

¹Department of Electrical and Electronic Engineering

²Department of Computing and Information Systems
University of Melbourne, Australia

IEEE International Symposium on Information Theory
June, 2020

Generalization in Machine Learning

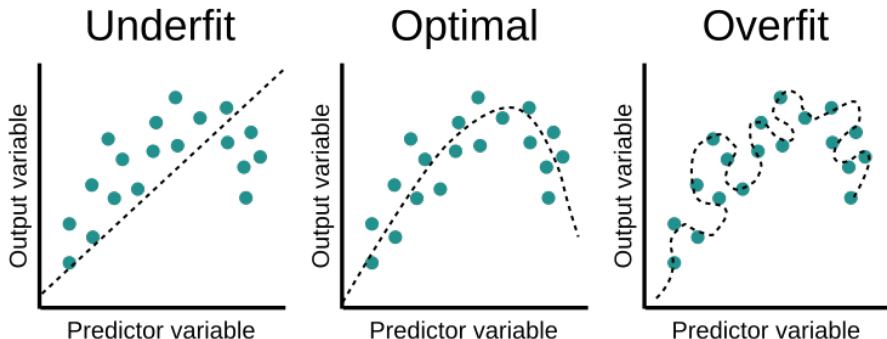


Figure: Overfitting and Underfitting¹

¹Source: www.educative.io/edpresso

Traditional Learning Theory

- Dataset $\mathcal{S}_n := \{z_1, z_2, \dots, z_n\}$
- Instance: $z_i \in \mathcal{Z} : \mathcal{X} \times \mathcal{Y}(\text{Supervised})$ or $\mathcal{X}(\text{Unsupervised})$
- Distributions: z_i i.i.d. drawn from μ
- Output hypothesis: $w \in \mathcal{W}$
- Loss function $\ell : (\mathcal{W}, \mathcal{Z}) \rightarrow \mathbb{R}^+$
- Optimal hypothesis $w^* := \arg \min_{w \in \mathcal{W}} \mathbb{E}_{z \sim \mu}[\ell(w^*, z)]$
- Excess risk:

$$\text{ER}(w) := \mathbb{E}_{z \sim \mu}[\ell(w, z)] - \mathbb{E}_{z \sim \mu}[\ell(w^*, z)]$$

Generalization Error for ERM algorithm

For ERM algorithm such that $w_{\text{ERM}} = \arg \min_w \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$, we have

Generalization Error for ERM algorithm

For ERM algorithm such that $w_{\text{ERM}} = \arg \min_w \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$, we have

$$\begin{aligned} \text{ER}(w_{\text{ERM}}) &= \mathbb{E}_{z \sim \mu}[\ell(w_{\text{ERM}}, z)] - \mathbb{E}_{z \sim \mu}[\ell(w^*, z)] \\ &= \text{gen}(w_{\text{ERM}}) + \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(w_{\text{ERM}}, z_i) - \frac{1}{n} \sum_{i=1}^n \ell(w^*, z_i)}_{< 0} + \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(w^*, z_i) - \mathbb{E}_{z \sim \mu}[\ell(w^*, z)]}_{LLN} \\ &\leq |\text{gen}(w_{\text{ERM}})| + \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(w^*, z_i) - \mathbb{E}_{z \sim \mu}[\ell(w^*, z)]}_{LLN} \end{aligned}$$

where we define the generalization error as

$$\text{gen}(W) := \left| \mathbb{E}_{Z \sim \mu}[\ell(W, Z)] - \frac{1}{n} \sum_{i=1}^n \ell(W, Z_i) \right|$$

Information Theory for generalization error

- Russo & Zou [TIT'19], Xu & Raginsky [NIPS'17]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

Information Theory for generalization error

- Russo & Zou [TIT'19], Xu & Raginsky [NIPS'17]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

- Raginsky et al. [ITW'16]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n I(W; Z_i | S^{-i})}$$

Information Theory for generalization error

- Russo & Zou [TIT'19], Xu & Raginsky [NIPS'17]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

- Raginsky et al. [ITW'16]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n I(W; Z_i | S^{-i})}$$

- Bu et al. [ISIT'19]:

$$\mathbb{E}[\text{gen}(W)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

Information Theory for generalization error

- Russo & Zou [TIT'19], Xu & Raginsky [NIPS'17]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

- Raginsky et al. [ITW'16]:

$$\mathbb{E}[\text{gen}(W)] \leq \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n I(W; Z_i | S^{-i})}$$

- Bu et al. [ISIT'19]:

$$\mathbb{E}[\text{gen}(W)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

- Negrea et al. [NIPS'19]:

$$\mathbb{E}[\text{gen}(W)] \leq \mathbb{E}[\sqrt{D_{\text{KL}}(Q(S) \| P(S_J)) / 2}]$$

What is Transfer Learning?

Example:



Figure: Transfer learning of digit recognition

Notations

- Target Data: $(z_1, z_2, \dots, z_{\beta n})$ for $\beta \in [0, 1)$ with

$$S' \in \mathcal{S}' := \underbrace{\mathcal{Z} \times \mathcal{Z} \cdots \times \mathcal{Z}}_{\beta n \text{ times}}$$

- Target Data distribution:

$$P_{S'} := \underbrace{\mu' \times \mu' \cdots \times \mu'}_{\beta n \text{ times}} := \mu'^{\otimes \beta n}$$

Notations

- Target Data: $(z_1, z_2, \dots, z_{\beta n})$ for $\beta \in [0, 1)$ with

$$S' \in \mathcal{S}' := \underbrace{\mathcal{Z} \times \mathcal{Z} \cdots \times \mathcal{Z}}_{\beta n \text{ times}}$$

- Target Data distribution:

$$P_{S'} := \underbrace{\mu' \times \mu' \cdots \times \mu'}_{\beta n \text{ times}} := \mu'^{\otimes \beta n}$$

- Source Data: $(z_{\beta n+1}, z_{\beta n+2}, \dots, z_n)$ with

$$S \in \mathcal{S} := \underbrace{\mathcal{Z} \times \mathcal{Z} \cdots \times \mathcal{Z}}_{(1-\beta)n \text{ times}}$$

- Source Data distribution:

$$P_S := \underbrace{\mu \times \mu \cdots \times \mu}_{(1-\beta)n \text{ times}} := \mu^{\otimes (1-\beta)n}$$

- Optimal hypothesis:

$$w^* = \arg \min_w \mathbb{E}_{Z \sim \mu'} [\ell(w, Z)]$$

- Empirical Loss [Ben-David et al., ML'10]:

$$\begin{aligned} L_\alpha(w) &= \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \ell(w, z_i) + \frac{1 - \alpha}{(1 - \beta)n} \sum_{i=\beta n+1}^n \ell(w, z_i) \\ &:= \alpha \hat{\mathbb{E}}_{S'} [\ell(w, z_i)] + (1 - \alpha) \hat{\mathbb{E}}_S [\ell(w, z_i)] \end{aligned}$$

- ERM hypothesis:

$$w_{\text{ERM}} = \arg \min_w L_\alpha(w, S')$$

Excess Risk:

$$ER(W) := \mathbb{E}_{Z' \sim \mu'}[\ell(w, Z')] - \mathbb{E}_{Z' \sim \mu'}[\ell(w^*, Z')]$$

Excess Risk:

$$ER(W) := \mathbb{E}_{Z' \sim \mu'}[\ell(w, Z')] - \mathbb{E}_{Z' \sim \mu'}[\ell(w^*, Z')]$$

We can then decompose the excess risk as

$$ER(W_{\text{ERM}}) \leq |\text{gen}(W_{\text{ERM}}) + \underbrace{L_{\alpha}(w^*, S, S') - L_{P_S, P_{S'}}(w^*)}_{LNN} + \underbrace{(1 - \alpha)(L_{\mu}(w^*) - L_{\mu'}(w^*))}_{\text{constant}}|$$

Generalization error for transfer learning:

$$\text{gen}(W_{\text{ERM}}) = L_{\mu'}(W_{\text{ERM}}) - L_{\alpha}(W_{\text{ERM}}, S, S')$$

Main results

If $\ell(W, Z) - \mathbb{E}\{\ell(W, Z)\}$ is upper bounded by $\psi(\lambda)$ in the interval (b_-, b_+) under the product distribution $P_W \otimes \mu'$ for some $b_- < 0$ and $b_+ > 0$:

$$\begin{aligned}\mathbb{E}\{\text{gen}(W_{\text{ERM}})\} &\leq \underbrace{\alpha \hat{\mathbb{E}}_{S'}[\psi_-^{*-1}(I(W_{\text{ERM}}; Z_i))]}_{\text{Target}} + \underbrace{(1 - \alpha) \hat{\mathbb{E}}_S[\psi_-^{*-1}(I(W_{\text{ERM}}; Z_i) + D(\mu||\mu')))]}_{\text{Source}} \\ -\mathbb{E}\{\text{gen}(W_{\text{ERM}})\} &\leq \alpha \hat{\mathbb{E}}_{S'}[\psi_+^{*-1}(I(W_{\text{ERM}}; Z_i))] + (1 - \alpha) \hat{\mathbb{E}}_S[\psi_+^{*-1}(I(W_{\text{ERM}}; Z_i) + D(\mu||\mu')))]\end{aligned}$$

where we define

$$\begin{aligned}\psi_-^{*-1}(x) &:= \inf_{\lambda \in [0, -b_-)} \frac{x + \psi(-\lambda)}{\lambda} \\ \psi_+^{*-1}(x) &:= \inf_{\lambda \in [0, b_+)} \frac{x + \psi(\lambda)}{\lambda}\end{aligned}$$

Sub-gaussian upper bound

If the loss function $\ell(W, Z)$ is r^2 -subgaussian, namely

$$\log \mathbb{E}\{e^{\lambda(\ell(W, Z) - \mathbb{E}\{\ell(W, Z)\})}\} \leq \frac{r^2 \lambda^2}{2}$$

Sub-gaussian upper bound

If the loss function $\ell(W, Z)$ is r^2 -subgaussian, namely

$$\log \mathbb{E}\{e^{\lambda(\ell(W, Z) - \mathbb{E}\{\ell(W, Z)\})}\} \leq \frac{r^2 \lambda^2}{2}$$

For any source samples and target samples

$$\begin{aligned} |\mathbb{E}\{\text{gen}(W_{\text{ERM}})\}| &\leq \alpha \hat{\mathbb{E}}_{S'} \left[\sqrt{2r^2 I(W_{\text{ERM}}; Z_i)} \right] \\ &\quad + (1 - \alpha) \hat{\mathbb{E}}_S \left[\sqrt{2r^2 (I(W_{\text{ERM}}; Z_i) + D(\mu || \mu'))} \right] \end{aligned}$$

Sub-gaussian upper bound

If the loss function $\ell(W, Z)$ is r^2 -subgaussian, namely

$$\log \mathbb{E}\{e^{\lambda(\ell(W, Z) - \mathbb{E}\{\ell(W, Z)\})}\} \leq \frac{r^2 \lambda^2}{2}$$

For any source samples and target samples

$$\begin{aligned} |\mathbb{E}\{\text{gen}(W_{\text{ERM}})\}| &\leq \alpha \hat{\mathbb{E}}_{S'} \left[\sqrt{2r^2 I(W_{\text{ERM}}; Z_i)} \right] \\ &\quad + (1 - \alpha) \hat{\mathbb{E}}_S \left[\sqrt{2r^2 (I(W_{\text{ERM}}; Z_i) + D(\mu || \mu'))} \right] \end{aligned}$$

For $\beta = 0$, under generic $\hat{W}$

$$|\mathbb{E}_{\hat{W}S} [L_{\mu'}(\hat{W}) - \hat{L}(\hat{W}, S)]| \leq \frac{\sqrt{2r^2}}{n} \sum_{i=1}^n \sqrt{I(\hat{W}; Z_i) + D(\mu || \mu')}$$

Excess Risk Bound

Excess Risk w.r.t. μ' :

$$\text{ER}(W) = \mathbb{E}_{Z \sim \mu'}[\ell(W, Z_i)] - \mathbb{E}_{Z \sim \mu'}[\ell(W^*, Z_i)]$$

With probability at least $1 - \delta$:

$$\begin{aligned} \text{ER}(W_{\text{ERM}}) &\leq \alpha \hat{\mathbb{E}}_{S'} \left[\sqrt{2r^2 I(W_{\text{ERM}}; Z_i)} \right] \\ &\quad + (1 - \alpha) \hat{\mathbb{E}}_S \left[\sqrt{2r^2 (I(W_{\text{ERM}}; Z_i) + D(\mu || \mu'))} \right] \\ &\quad + \sqrt{\frac{\alpha^2}{\beta} + \frac{(1 - \alpha)^2}{(1 - \beta)}} \sqrt{\frac{2r^2 \ln \frac{2}{\delta}}{n}} + (1 - \alpha) D_{\mathcal{W}}(\mu, \mu') + \epsilon \end{aligned}$$

where $D_{\mathcal{W}}(\mu, \mu') = \sup_{w \in \mathcal{W}} |L_{\mu}(w) - L_{\mu'}(w)|$.

Noisy iterative algorithm

We consider the following noisy gradient descent algorithm

$$W(t) = W(t-1) - \eta_t \nabla \hat{L}_\alpha(W(t-1), S, S') + n(t)$$

Noisy iterative algorithm

We consider the following noisy gradient descent algorithm

$$W(t) = W(t-1) - \eta_t \nabla \hat{L}_\alpha(W(t-1), S, S') + n(t)$$

After running T iterations, we can bound

$$\mathbb{E}_{wSS'}[\text{gen}(W(T))] \leq \alpha \sqrt{\frac{2r^2}{\beta n} \hat{I}(S')} + (1-\alpha) \sqrt{2r^2 \left(\frac{\hat{I}(S)}{(1-\beta)n} + D(\mu \parallel \mu') \right)}$$

Where we define

$$\hat{I}(S') = \hat{I}(S) := \frac{d}{2} \sum_{t=1}^T \log \left(2\pi e \frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d} \right) - \sum_{t=1}^T h(n(t))$$

- ϕ divergence bounds [Jiao et al., ISIT'17]

$$\begin{aligned}\mathbb{E}[\text{gen}(W_{\text{ERM}})] &\leq \frac{\alpha \|\sigma\|_{\infty}}{\beta n} \sum_{i=1}^{\beta n} I_{\phi}(W_{\text{ERM}}; z_i) \\ &\quad + \frac{(1-\alpha) \|\sigma\|_{\infty}}{(1-\beta)n} \sum_{i=\beta n+1}^n (I_{\phi}(W_{\text{ERM}}; z_i) + 2TV(\mu, \mu'))\end{aligned}$$

- ϕ divergence bounds [Jiao et al., ISIT'17]

$$\begin{aligned}\mathbb{E}[\text{gen}(W_{\text{ERM}})] &\leq \frac{\alpha \|\sigma\|_{\infty}}{\beta n} \sum_{i=1}^{\beta n} I_{\phi}(W_{\text{ERM}}; z_i) \\ &\quad + \frac{(1-\alpha) \|\sigma\|_{\infty}}{(1-\beta)n} \sum_{i=\beta n+1}^n (I_{\phi}(W_{\text{ERM}}; z_i) + 2TV(\mu, \mu'))\end{aligned}$$

- Wasserstein distance bounds [Wang et al., ISIT'19]

$$\begin{aligned}\mathbb{E}[\text{gen}(W_{\text{ERM}})] &\leq \frac{\alpha \mathcal{L}}{\beta n} \sum_{i=1}^{\beta n} \mathbb{E}_{\mu'}[\mathbb{W}_1(P_{W_{\text{ERM}}}, P_{W_{\text{ERM}}|Z_i})] \\ &\quad + \frac{(1-\alpha) \mathcal{L}}{(1-\beta)n} \sum_{i=\beta n+1}^n (\mathbb{E}_{\mu}[\mathbb{W}_1(P_{W_{\text{ERM}}}, P_{W_{\text{ERM}}|Z_i})] + \mathbb{W}_1(\mu, \mu'))\end{aligned}$$

Gaussian Transfer

Data Distribution: $\mu = \mathcal{N}(m, \sigma^2)$, $\mu' = \mathcal{N}(m', \sigma^2)$, $m \neq m'$

Loss function: $\ell(w, z) = (w - z)^2$.

ERM solution & Mutual Information::

$$W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i, \quad I(W_{\text{ERM}}; Z_i) = \frac{1}{2} \log \frac{n}{n-1}$$

Gaussian Transfer

Data Distribution: $\mu = \mathcal{N}(m, \sigma^2)$, $\mu' = \mathcal{N}(m', \sigma^2)$, $m \neq m'$

Loss function: $\ell(w, z) = (w - z)^2$.

ERM solution & Mutual Information::

$$W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i, \quad I(W_{\text{ERM}}; Z_i) = \frac{1}{2} \log \frac{n}{n-1}$$

True generalization error:

$$\mathbb{E}\{\hat{L}(W_{\text{ERM}}, S) - L_{\mu'}(W_{\text{ERM}})\} = \frac{2\sigma^2}{n} + 2\sigma^2 D(\mu \parallel \mu')$$

Gaussian Transfer

Data Distribution: $\mu = \mathcal{N}(m, \sigma^2)$, $\mu' = \mathcal{N}(m', \sigma^2)$, $m \neq m'$

Loss function: $\ell(w, z) = (w - z)^2$.

ERM solution & Mutual Information::

$$W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i, \quad I(W_{\text{ERM}}; Z_i) = \frac{1}{2} \log \frac{n}{n-1}$$

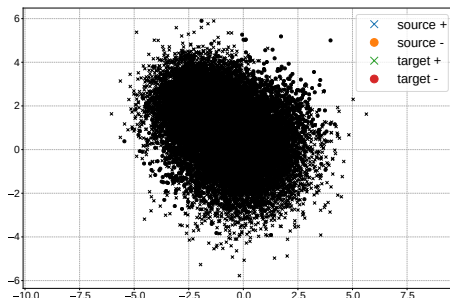
True generalization error:

$$\mathbb{E}\{\hat{L}(W_{\text{ERM}}, S) - L_{\mu'}(W_{\text{ERM}})\} = \frac{2\sigma^2}{n} + 2\sigma^2 D(\mu \parallel \mu')$$

Upper-Bound:

$$\mathbb{E}\{\text{gen}(W_{\text{ERM}})\} \leq 2 \left(\frac{n+1}{n} \right) \sigma^2 \sqrt{\frac{1}{2} \log \frac{n}{n-1}} + 2\sigma^2 D(\mu \parallel \mu')$$

Logistic Transfer



Bernoulli distribution with probability:

$$p(1) = \frac{1}{1 + e^{-w^T x}}$$

Loss function:

$$\ell(w, z_i) := -(y_i \log(\sigma(w^T x_i)) + (1 - y_i) \log(1 - \sigma(w^T x_i)))$$

Comparison with existing works

We compare the bound with the Rademacher Complexity bound [Zhang et al., NIPS'12] under the same settings:

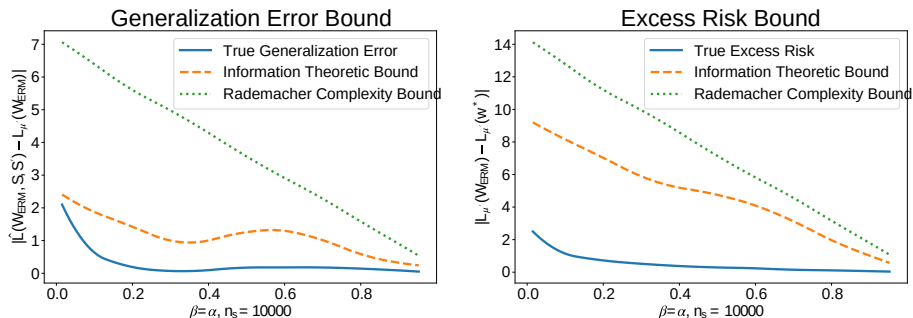


Figure: Comparisons for generalization error and excess risk

Conclusion

- Generalization error and excess risk bounds for transfer learning using mutual information.

Conclusion

- Generalization error and excess risk bounds for transfer learning using mutual information.
- KL divergence (and other types of metrics) can be used to capture the domain divergence under this framework.

Conclusion

- Generalization error and excess risk bounds for transfer learning using mutual information.
- KL divergence (and other types of metrics) can be used to capture the domain divergence under this framework.
- Our bound is tighter in specific classification problems than the bound derived using Rademacher complexity.

Conclusion

- Generalization error and excess risk bounds for transfer learning using mutual information.
- KL divergence (and other types of metrics) can be used to capture the domain divergence under this framework.
- Our bound is tighter in specific classification problems than the bound derived using Rademacher complexity.
- Future work: Gibbs algorithm.

- Generalization error and excess risk bounds for transfer learning using mutual information.
- KL divergence (and other types of metrics) can be used to capture the domain divergence under this framework.
- Our bound is tighter in specific classification problems than the bound derived using Rademacher complexity.
- Future work: Gibbs algorithm.

Thank You!