PAC-Bayesian Domain Adaptation Bounds for Multiclass Learners

Anthony Sicilia[†] Kate Atwell[†] Malihe Alikhani[†] Seong Jae Hwang[‡]
{anthonysicilia, kaa139, malihe}@pitt.edu, seongjae@yonsei.ac.kr

[†]University of Pittsburgh, Pittsburgh, PA, USA
[‡]Yonsei University, Seoul, South Korea

August 2, 2022

Primary Objective of the Paper

Provide tools for theoretical and empirical analysis of multiclass neural networks in domain adaptation

Multiclass Neural Networks are Common in Adaptation

Multiclass neural networks are frequently used in implementation of unsupervised domain adaptation algorithms:

- invariant feature learning algorithms [Ganin and Lempitsky, 2015, Long et al., 2017, 2018, Zhang et al., 2019],
- importance weighting algorithms [Lipton et al., 2018],
- or combinations of both techniques [Tachet des Combes et al., 2020].

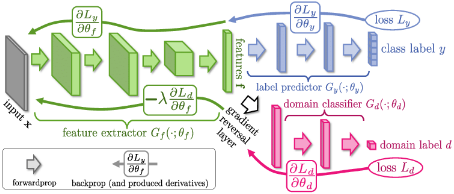

Figure: (**Case Study**) The DANN Algorithm [Ganin and Lempitsky, 2015]

DANN and its Theoretical Motivations

DANN is theoretically motivated by an error bound. In particular, DANN is a map $(S, T_X) \mapsto h$ and for all solution models h:

$$\underbrace{\mathbf{R}_{\mathbb{T}}(h)}_{\text{target error}} \leq \underbrace{\mathbf{R}_{S}(h)}_{\text{source error}} + \underbrace{\lambda(\mathbb{S}, \mathbb{T})}_{\text{adaptability}} + \underbrace{\mathbf{d}(S_X, T_X, h)}_{\text{divergence}} + \underbrace{\Gamma(n, m, h)}_{\text{sample complexity}} \tag{1}$$

- **Target Error** $\mathbf{R}_{\mathbb{T}}(h)$: error on the goal distribution $\mathbb{T}$
- **Source Error** $\mathbf{R}_{S}(h)$: error on the sample in-hand $S \overset{iid}{\sim} \mathbb{S}$
- **Adaptability** $\lambda(\mathbb{S}, \mathbb{T})$: change in labeling functions from $\mathbb{S}$ to $\mathbb{T}$
- **Divergence** $\mathbf{d}(S_X, T_X, h)$: change in feature distributions from $\mathbb{S}$ to $\mathbb{T}$
- **Sample Complexity** $\Gamma(n, m, h)$: data hunger (efficiency) of the solution h, dependent on sample size n of S and/or m of T_X

DANN and its Theoretical Motivations

DANN is theoretically motivated by an error bound. In particular, DANN is a map $(S, T_X) \mapsto h$ and for all solution models h:

$$\underbrace{\mathbf{R}_{\mathbb{T}}(h)}_{\text{target error}} \leq \underbrace{\mathbf{R}_S(h)}_{\text{source error}} + \underbrace{\lambda(\mathbb{S}, \mathbb{T})}_{\text{adaptability}} + \underbrace{\mathbf{d}(S_X, T_X, h)}_{\text{divergence}} + \underbrace{\Gamma(n, m, h)}_{\text{sample complexity}} \quad (2)$$

- **Target Error** $\mathbf{R}_{\mathbb{T}}(h)$: error on the goal distribution $\mathbb{T}$
- **Source Error** $\mathbf{R}_S(h)$: error on the sample in-hand $S \overset{iid}{\sim} \mathbb{S}$
- **Adaptability** $\lambda(\mathbb{S}, \mathbb{T})$: change in labeling functions from $\mathbb{S}$ to $\mathbb{T}$
- **Divergence** $\mathbf{d}(S_X, T_X, h)$: change in feature distributions from $\mathbb{S}$ to $\mathbb{T}$
- **Sample Complexity** $\Gamma(n, m, h)$: data hunger (efficiency) of the solution h, dependent on sample size n of S and/or m of T_X

DANN minimizes the source error and divergence

DANN Ignores Some Important Terms

Problem: This ignores some important terms

$$\underbrace{\mathbf{R}_{\mathbb{T}}(h)}_{\text{target error}} \leq \underbrace{\mathbf{R}_S(h)}_{\text{source error}} + \underbrace{\lambda(\mathbb{S}, \mathbb{T})}_{\text{adaptability}} + \underbrace{\mathbf{d}(S_X, T_X)}_{\text{divergence}} + \underbrace{\Gamma(n, m, h)}_{\text{sample complexity}} \tag{3}$$

DANN can behave unexpectedly b/c it ignores similarity of the labeling functions [Johansson et al., 2019, Wu et al., 2019, Zhao et al., 2019].

This Paper: Sample complexity is ignored as well, and adaptability is understudied, empirically [Redko et al., 2020]:

> *Can adaptability be empirically estimated to provide insight on the behavior of domain adaptation algorithms?*
>
> *Can non-uniform sample complexity provide additional empirical insight on the behavior of domain adaptation algorithms?*

What Tools are Available in Adaptation?

Unfortunately, the literature lacks suitable theoretical tools.

To answer these questions, we require:

- **A Multiclass Setting**: Most algorithms (e.g., DANN) are employed on multiclass datasets like MNIST
- **Practicality**: All terms should be easily empirically estimable
- **PAC-Bayes**: Demonstrated accuracy in measuring the *non-uniformity* of neural network sample complexity [Jiang et al., 2019, Dziugaite et al., 2020, Pérez-Ortiz et al., 2021]

Absent of suitable candidates, we propose some, and further, propose approximation techniques for all of their contained terms.

Proposed Bounds

Theorem

For any $\mathbb{P}$ *over* $\mathcal{H}$*, all* $\delta > 0$*, w.p. at least* $1 - \delta$*, for all* $\mathbb{Q}$ *over* $\mathcal{H}$

$$\mathbf{R}_{\mathbb{T}}(\mathbb{Q}) \leq \tilde{\lambda}_{S,T} + \mathbf{R}_S(\mathbb{Q}) + \mathbf{E}_{H\sim\mathbb{Q}}[\mathbf{d}_{\mathcal{C}_H}(S_X, T_X)] + \sqrt{\frac{\mathrm{KL}(\mathbb{Q}||\mathbb{P}) + \ln\sqrt{4m} - \ln(\delta)}{2m}} \tag{4}$$

where $\tilde{\lambda}_{S,T} = \min_{\eta\in\mathcal{H}} \mathbf{R}_S(\eta) + \mathbf{R}_T(\eta)$ *and we may choose either* $\mathcal{C}_h = \mathcal{H}\Delta\mathcal{H}$ *for all* h *as before or* $\mathcal{C}_h = h\Delta\mathcal{H}$*.*

What we will talk about?

1. Key empirical results on some elements of the bound
   - 12K+ trained models, 5 datasets, diverse adaptation scenarios
2. Application of the bound to DANN on MNIST

What (else) is in the paper?

1. Technical details on estimation methodology, more detailed results

Estimating Adaptability

$$\underbrace{\mathbf{R}_{\mathbb{T}}(h)}_{\text{target error}} \leq \underbrace{\mathbf{R}_S(h)}_{\text{source error}} + \underbrace{\lambda(\mathbb{S},\mathbb{T})}_{\text{adaptability}} + \underbrace{\mathbf{d}(S_X, T_X, h)}_{\text{divergence}} + \underbrace{\Gamma(n, m, h)}_{\text{sample complexity}} \tag{5}$$

- Bound replaces population stat. $\lambda(\mathbb{S},\mathbb{T})$ by sample stat. $\lambda(S, T)$
- Removes generalization penalty for more interpretable results
- *First-ever* large-scale empirical study of λ: **confirms it is often small**

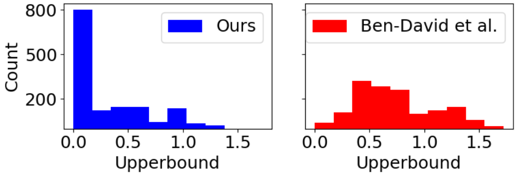

Figure: Older definition (right) requires generalization penalty that mars results

Non-Uniform Sample Complexity Provides Useful Insight

$$\underset{\text{target error}}{\mathbf{R}_{\mathbb{T}}(h)} \leq \underset{\text{source error}}{\mathbf{R}_S(h)} + \underset{\text{adaptability}}{\lambda(\mathbb{S}, \mathbb{T})} + \underset{\text{divergence}}{\mathbf{d}(S_X, T_X, h)} + \underset{\text{sample complexity}}{\Gamma(n, m, h)} \tag{6}$$

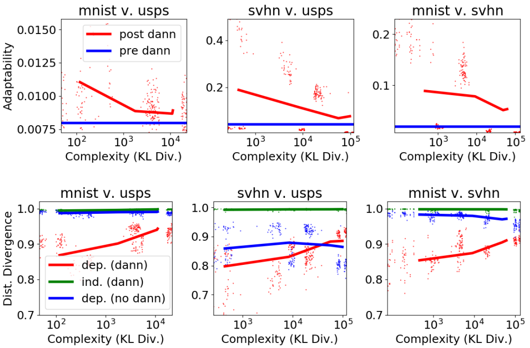

Figure: (**DANN** on **Digits**) Divergence and adaptability compete, leading to imperfect minimization. As divergence increases, adaptability decreases and vice-versa. Sample complexity is a modulating factor.

Approximation of Multiclass Divergence

$$\underbrace{\mathbf{R}_{\mathbb{T}}(h)}_{\text{target error}} \leq \underbrace{\mathbf{R}_{S}(h)}_{\text{source error}} + \underbrace{\lambda(\mathbb{S}, \mathbb{T})}_{\text{adaptability}} + \underbrace{\mathbf{d}(S_X, T_X, h)}_{\text{divergence}} + \underbrace{\Gamma(n, m, h)}_{\text{sample complexity}} \quad (7)$$

- PAC-Bayes divergences limited to binary setting [Germain et al., 2013, 2020]
- Build on strategies in binary setting [Ben-David et al., 2010, Kuroki et al., 2019]
- Multiclass setting requires removal of a symmetry assumption, proposal of a novel surrogate loss, and additional constraints
- Propose replacement of inefficient MC sampling by flatness penalty

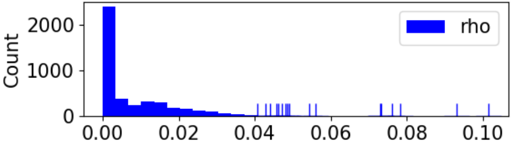

Figure: Flatness penalty is typically small.

Conclusion

We propose new PAC-Bayesian adaptation bounds and approximation techniques for the statistics within these bounds

- First multiclass adaptation bound (non-uniform sample complexity)
- Design focuses on empirical practicality
- Demonstrate utility in empirical analysis of adaptability and DANN

Some useful links:

OpenReview: openreview.net/pdf?id=S0lx6I8j9xq
Concurrent Work (ACL 2022): arxiv.org/abs/2203.11317
Code: github.com/anthonysicilia/pacbayes-adaptation-UAI2022
Package: github.com/anthonysicilia/classifier-divergence

Contact Us:

{anthonysicilia, kaa139, malihe}@pitt.edu, seongjae@yonsei.ac.kr

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *NeurIPS*, 33, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, pages 738–746. PMLR, 2013.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. Pac-bayes and domain adaptation. *Neurocomputing*, 379: 379–397, 2020.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*, 2019.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *AISTATS*, pages 527–536. PMLR, 2019.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *AAAI*, volume 33, pages 4122–4129, 2019.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, pages 3122–3130. PMLR, 2018.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217. PMLR, 2017.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018.

Marıa Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *JMLR*, 22, 2021.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory. *ArXiv*, abs/2004.11829, 2020.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *NeurIPS*, 33, 2020.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, pages 6872–6881. PMLR, 2019.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, pages 7404–7413. PMLR, 2019.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *ICML*, pages 7523–7532. PMLR, 2019.