Sharper Generalization Bounds for Pairwise Learning

Yunwen Lei[1,2], Antoine Ledent[2] and Marius Kloft[2]

[1]University of Birmingham
[2]University of Kaiserslautern

y.lei@bham.ac.uk {ledent, kloft}@cs.uni-kl.de

December, 2020

Pairwise Learning

Data: $S = \{z_i = (x_i, y_i)\}_{i=1}^n \sim \rho$ defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

We learn a model $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$ or $h_{\mathbf{w}} : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}, \mathbf{w} \in \mathcal{W}$

Pairwise loss: $\ell(\mathbf{w}; z, z')$ measures behavior of $h_{\mathbf{w}}$ over z, z'

Population risk and Empirical risk

$$R(\mathbf{w}) = \mathbb{E}_{z,\tilde{z}}\big[\ell(\mathbf{w}; z, \tilde{z})\big], \quad R_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \ell(\mathbf{w}; z_i, z_j).$$

Algorithm: $A : \mathcal{Z}^n \mapsto \mathcal{W}$ (output $A(S)$ when applied to S)

We study generalization gap $R(A(S)) - R_S(A(S))$!

Algorithmic Stability

Uniform Stability

We say $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is γ-uniformly stable if for any training datasets $S, S' \in \mathcal{Z}^n$ that differ by at most a single example

$$\sup_{z,\tilde{z}\in\mathcal{Z}} \left|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z})\right| \leq \gamma.$$

On-average stability

Let $S = \{z_1, \dots, z_n\}, S' = \{z'_1, \dots, z'_n\}$. For any $i < j$ let

$$S_{i,j} = \{z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_{j-1}, z'_j, z_{j+1}, \dots, z_n\}. \tag{1}$$

We say a deterministic algorithm A is γ-on-average stable if

$$\frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,S'}\Big[\ell\big(A(S_{i,j}); z_i, z_j\big) - \ell\big(A(S); z_i, z_j\big)\Big] \leq \gamma.$$

(Bousquet and Elisseeff, 2002; Elisseeff et al., 2005; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Feldman and Vondrak, 2019)

Generalization by Stability

Generalization by Uniform Stability

If $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is γ-uniformly stable, then with high probability

$$|R_S(A(S)) - R(A(S))| = \widetilde{O}\Big(\gamma + n^{-1/2}\Big).$$

- Improves the existing bound $O(\sqrt{n}\gamma + n^{-1/2})$ by a factor of $\sqrt{n}$ (Agarwal and Niyogi, 2009; Wang et al., 2019)
- Uses novel decomposition to address dependency of $n(n-1)$ terms in R_S

Generalization by On-average Stability

If A is γ-on-average stable, then

$$\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \le \gamma.$$

Application

Regularized Risk Minimization (RRM): with a regularizer $r : \mathcal{W} \mapsto \mathbb{R}$

$$\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathcal{W}} \Big[F_S(\mathbf{w}) := \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \ell(\mathbf{w}; z_i, z_j) + r(\mathbf{w})\Big]. \tag{2}$$

SGD: at t-th iteration, SGD randomly selects (i_t, j_t) and

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}).$$

Let $\mathbf{w}_R^* = \arg\inf_{\mathbf{w}} R(\mathbf{w})$ and A be RRM/SGD with appropriate parameters.

- We get excess risk bound $R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-1/2})$
- Existing stability analysis shows $R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-1/4})$ (Agarwal and Niyogi, 2009; Wang et al., 2019)
- We remove bounded loss assumption (Bousquet et al., 2020)

References I

S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(Feb):441–474, 2009.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6 (Jan):55–79, 2005.

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

B. Wang, H. Zhang, P. Liu, Z. Shen, and J. Pineau. Multitask metric learning: Theory and algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 3362–3371, 2019.