Sharper Generalization Bounds for Pairwise Learning

Yunwen Lei\textsuperscript{1,2}, Antoine Ledent\textsuperscript{2} and Marius Kloft\textsuperscript{2}

\textsuperscript{1}University of Birmingham
\textsuperscript{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 $h_w : \mathcal{X} \mapsto \mathcal{Y}$ or $h_w : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}$, $w \in \mathcal{W}$

Pairwise loss: $\ell(w; z, z')$ measures behavior of $h_w$ over $z, z'$

Population risk and Empirical risk

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

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 $\gamma$-uniformly stable if for any training datasets $S, S' \in \mathcal{Z}^n$ that differ by at most a single example

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

### On-average stability

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

$$S_{i,j} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n\}.$$ \hfill (1)

We say a deterministic algorithm $A$ is $\gamma$-on-average stable if

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

(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 $\gamma$-uniformly stable, then with high probability

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

- 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 $\gamma$-on-average stable, then

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

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

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

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

$$w_{t+1} = w_t - \eta_t \ell'(w_t; z_{i_t}, z_{j_t}).$$

Let $w_R^* = \arg \inf_w R(w)$ and $A$ be RRM/SGD with appropriate parameters.

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


