Generalization Guarantees via Algorithm-dependent Rademacher Complexity

Tim van Erven

University of Amsterdam

Joint work with:

Sarah Sachs  Liam Hodgkinson  Rajiv Khanna  Umut Şimşekli
Standard Batch Setting

Given:

- Data: $S^n = (Z_1, \ldots, Z_n) \overset{i.i.d.}{\sim} \mathcal{D}$
- Bounded loss: $\ell: \Theta \times \mathcal{Z} \rightarrow [a, a + b]$
- Algorithm: $\hat{\theta} \equiv \text{Alg}(S^n) \in \Theta$

Want to control the generalization error:

$$R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)$$

Where:

- Risk: $R(\theta) = \mathbb{E}_{Z \sim \mathcal{D}}[\ell(\theta, Z)]$
- Empirical risk: $\hat{R}(\theta, S^n) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i)$
Control via Mutual Information

Bound with mutual information [Catoni, 2007, Russo and Zou, 2016]:

$$\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \lesssim \sqrt{\frac{I(\hat{\theta}; S^n)}{n}}$$
Control via Mutual Information

Bound with mutual information [Catoni, 2007, Russo and Zou, 2016]:

$$
\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \lesssim \sqrt{\frac{I(\hat{\theta}; S^n)}{n}}
$$

Refined to conditional mutual information via symmetrization with a ghost sample [Steinke and Zakynthinou, 2020]:

$$
\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \lesssim \sqrt{\frac{\text{CMI}(\text{Alg})}{n}}
$$

Known limitations:

- No high probability bounds possible for CMI [Steinke and Zakynthinou, 2020]

- Bounds do not depend on loss function, so Steinke and Zakynthinou [2020] have variant of CMI to take advantage of e.g. smoothness of $\ell(\theta, z)$ in $\theta$. 
Standard Control via Rademacher Complexity

\[ R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq \sup_{\theta \in \Theta} \left( R(\theta) - \hat{R}(\theta, S^n) \right) \]  

(*)

Lemma (Algorithm-independent upper bound)

\[ \mathbb{E} \left[ \sup_{\theta \in \Theta} (R(\theta) - \hat{R}(\theta, S^n)) \right] \leq 2 \mathbb{E}_{S^n}[\text{Rad}(\Theta, S^n)] \]

and, with probability at least 1 − δ,

\[ \sup_{\theta \in \Theta} (R(\theta) - \hat{R}(\theta, S^n)) \leq 2 \mathbb{E}_{S^n}[\text{Rad}(\Theta, S^n)] + b \sqrt{\frac{\log(2/\delta)}{2n}} \]

Empirical Rademacher complexity:

\[ \text{Rad}(\Theta, S^n) = \frac{1}{n} \mathbb{E}_{\sigma} [\sup_{\theta \in \Theta} \sum_{i=1}^{n} \sigma_i \ell(\theta, Z_i)], \]

where \( \sigma = (\sigma_1, \ldots, \sigma_n) \) with \( \Pr(\sigma_i = -1) = \Pr(\sigma_i = +1) = 1/2. \)
Control via Algorithm-dependent Rademacher Complexity

Lemma (Algorithm-dependent upper bound)

\[ \mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \leq 2 \mathbb{E}_{S^n, \hat{\Theta}^n, S^n_+}[\text{Rad}(\hat{\Theta}^n, S^n_+)] \]
Control via Algorithm-dependent Rademacher Complexity

\[ \hat{\Theta}^n := \{ \text{Alg}(S^n_\sigma) : \sigma \in \{-1, +1\}^n \} \subset \Theta. \]

\[ S^n_- = (Z_1^{-1}, \ldots, Z_n^{-1}) \]
\[ S^n_+ = (Z_1^{+1}, \ldots, Z_n^{+1}) \]
\[ S^n_\sigma = (Z_1^{\sigma_1}, \ldots, Z_n^{\sigma_n}) \]

Lemma (Algorithm-dependent upper bound)

\[ \mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \leq 2 \mathbb{E}_{S^n_- \cdot S^n_+} [\text{Rad}(\hat{\Theta}^n, S^n_+)] \]

- Like normal Rademacher bound, but with \( \hat{\Theta}^n \) instead of \( \Theta \)
- Symmetrization with ghost sample \( S^n_- \) like CMI
- Proof: similar to standard proof, but upper bound \( \hat{\theta} \) by supremum over \( \theta \) later, after symmetrization
Control via Algorithm-dependent Rademacher Complexity

\[ \Theta^n := \{ \text{Alg}(S^n_{\sigma}) : \sigma \in \{-1, +1\}^n \} \subset \Theta. \]

\[
\begin{align*}
S^n_- &= (Z_1^{-1}, \ldots, Z_n^{-1}) \\
S^n_+ &= (Z_1^{+1}, \ldots, Z_n^{+1}) \\
S^n_{\sigma} &= (Z_{1}^{\sigma_1}, \ldots, Z_{n}^{\sigma_n})
\end{align*}
\]

Lemma (Algorithm-dependent upper bound)

\[
\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \leq 2 \mathbb{E}_{S^n_-, S^n_+} [\text{Rad}(\hat{\Theta}^n, S^n_+)]
\]

and, with probability at least \( 1 - \delta \),

\[
R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq 4 \text{ess sup}_{S^n_-, S^n_+} \text{Rad}(\hat{\Theta}^n, S^n_+) + b\sqrt{\frac{8 \log(2/\delta)}{n}}
\]

▶ Refines special case of a result by Foster et al. [2019]
Consequences 1: Topological Bounds

Define the (random) set \( \hat{\Theta} := \bigcup_{n=1}^{\infty} \hat{\Theta}^n \)

Minkowski dimension: \( \overline{\text{dim}}_M(\hat{\Theta}) = \limsup_{\delta \to 0^+} \frac{\log \text{Cover}(\hat{\Theta}, \| \cdot \|, \delta)}{\log(1/\delta)} \)

Theorem

Suppose \( \ell(\theta, z) \) is Lipschitz continuous in \( \theta \). Then

\[
\limsup_{n \to \infty} \frac{\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)]}{\sqrt{\log(n)/n}} \leq b \sqrt{2 \mathbb{E}[\overline{\text{dim}}_M(\hat{\Theta})]}.
\]

- Avoids bad \( l_\infty \) term (much larger than regular mutual information) from previous topological bounds [Simsekli et al., 2020]
- Non-asymptotic result at the poster
Consequences 2: Generalization for SGD

Greatly simplified proof of result by Park et al. [2022]:

Suppose $z \mapsto \ell(\theta, z)$:

- $\alpha$-strongly convex
- $\beta$-smooth
- $L$-Lipschitz

+ Other standard assumptions

**Theorem**

Then, for $T$ iterations of stochastic optimization by stochastic gradient descent with constant step size $\eta \in (0, \beta)$, w.p. $\geq 1 - \delta$

\[
R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) = O\left(\sqrt{\log n \over \log(1/\gamma)n} + \sqrt{\log(1/\delta) \over n} + {L \over n}\right),
\]

where $\gamma = \sqrt{1 - 2\alpha \eta + \alpha \beta \eta^2}$. 
Consequences 3: Properties Like CMI

Generalization for VC Classes:
For binary classification with $V = \text{VCdim}(\Theta)$:

$$\text{Rad}(\hat{\Theta}^n, S^n_+) \leq \text{Rad}(\Theta, S^n_+) = O\left(\sqrt{\frac{V \log n}{n}}\right)$$

Generalization for compression schemes:
If Alg is a $k$-compression scheme, then

$$\text{Rad}(\hat{\Theta}^n, S^n_+) = O\left(\sqrt{\frac{k \log n}{n}}\right)$$
Summary

Algorithm-dependent Rademacher complexity:
- Rademacher complexity of algorithm- and data-dependent set $\hat{\Theta}^n$ controls generalization error

Consequences:
1. New topological generalization bounds
2. Greatly simplified proof of a generalization bound for SGD
3. Generalization for VC classes and compression schemes (like CMI)
References


