COLT 2023

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: $\ell : \Theta \times \mathcal{Z} \to [a, a+b]$
- Algorithm: $\hat{\theta} \equiv \mathrm{Alg}(S^n) \in \Theta$

Want to control the **generalization error**:

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

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]:

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

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 θ.

Standard Control via Rademacher Complexity

$$R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq \sup_{\theta \in \Theta} \Big(R(\theta) - \hat{R}(\theta, S^n) \Big) \qquad (*)$$

Lemma (Algorithm-independent upper bound)

$$\mathbb{E}\Big[\sup_{\theta \in \Theta} \big(R(\theta) - \hat{R}(\theta, S^n)\big)\Big] \leq 2 \mathop{\mathbb{E}}_{S^n}[\mathrm{Rad}(\Theta, S^n)]$$

and, with probability at least $1 - \delta$,

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

Empirical Rademacher complexity:

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

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)

$$\mathbb{E}[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)] \leq 2 \mathop{\mathbb{E}}_{S_-^n, S_+^n}[\mathrm{Rad}(\hat{\boldsymbol{\Theta}}^{\boldsymbol{n}}, S_+^n)]$$

Control via Algorithm-dependent Rademacher Complexity

$$\hat{\Theta}^n := \left\{\mathrm{Alg}(S^n_\sigma) : \sigma \in \{-1,+1\}^n\right\} \subset \Theta.$$

$$\left.\begin{array}{l} S^n_- = (Z_1^{-1}, \ldots, Z_n^{-1}) \\ S^n_+ = (Z_1^{+1}, \ldots, Z_n^{+1}) \end{array}\right\} \quad S^n_\sigma = (Z_1^{\sigma_1}, \ldots, Z_n^{\sigma_n})$$

Lemma (Algorithm-dependent upper bound)

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

- Like normal Rademacher bound, but with $\hat{\Theta}^n$ instead of Θ
- Symmetrization with ghost sample S^n_- like CMI
- Proof: similar to standard proof, but upper bound $\hat{\theta}$ by supremum over θ later, after symmetrization

Control via Algorithm-dependent Rademacher Complexity

$$\hat{\Theta}^n := \left\{ \mathrm{Alg}(S^n_\sigma) : \sigma \in \{-1,+1\}^n \right\} \subset \Theta.$$

$$\left.\begin{array}{l} S^n_- = (Z_1^{-1}, \ldots, Z_n^{-1}) \\ S^n_+ = (Z_1^{+1}, \ldots, Z_n^{+1}) \end{array}\right\} \quad S^n_\sigma = (Z_1^{\sigma_1}, \ldots, Z_n^{\sigma_n})$$

Lemma (Algorithm-dependent upper bound)

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

and, with probability at least $1-\delta$,

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

- 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: $\overline{\dim}_{\mathcal{M}}(\hat{\Theta}) = \limsup_{\delta \to 0^+} \frac{\log \mathrm{Cover}(\hat{\Theta}, \|\cdot\|, \delta)}{\log(1/\delta)}$

Theorem

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

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

- Avoids bad I_∞ 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 $z \mapsto \ell(\theta, z)$:

- **α-strongly convex**
- **β-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. $\geq 1 - \delta$*

$$R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) = O\Big(\sqrt{\frac{\log n}{\log(\frac{1}{\gamma})n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{L}{n}\Big),$$

where $\gamma = \sqrt{1 - 2\alpha\eta + \alpha\beta\eta^2}$.

Consequences 3: Properties Like CMI

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

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

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

$$\mathrm{Rad}(\hat{\Theta}^n, S_+^n) = O\left(\sqrt{\frac{k \log n}{n}}\right)$$

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

O. Catoni. Pac-Bayesian supervised classification: The thermodynamics of statistical learning. Lecture Notes — Monograph Series, Volume 56, 2007.

D. J. Foster, S. Greenberg, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems 32*, 2019.

S. Park, U. Şimşekli, and M. A. Erdogdu. Generalization bounds for stochastic gradient descent via localized ε-covers. In *Advances in Neural Information Processing Systems*, 2022.

D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, 2016.

U. Simsekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In *Advances in Neural Information Processing Systems 33*, 2020.

T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *COLT*, 2020.