Sharp uniform convergence bounds through empirical centralization

http://cs.brown.edu/~ccousins/ecentralization/
†cyrus_cousins@brown.edu

Cyrus Cousins† and Matteo Riondato

Brown University Amherst College

Neural Information Processing Systems 2020

Tools for Uniform Convergence

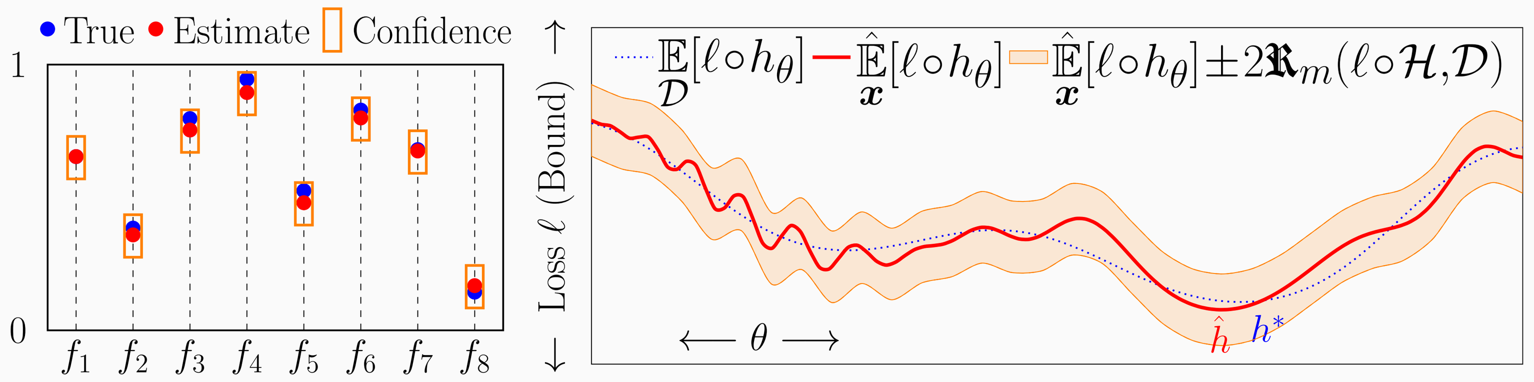
Family $\mathcal{F} \subseteq \mathcal{X} \to [0, r] \subseteq \mathbb{R}$ Sample $\boldsymbol{x} \sim \mathcal{D}^m$ Rademacher $\boldsymbol{\sigma} \sim \mathcal{U}(\pm 1)$

Empirical Rade Avg $\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right]$

Rademacher Avg $\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right]\right]$

The Wimpy Variance:

	Empirical	**True**
Raw	$\hat{v}^{\mathrm{raw}}_{\sup} \doteq \sup_{f \in \mathcal{F}} \hat{\mathbb{E}}_{\boldsymbol{x}}[f^2]$	$v^{\mathrm{raw}}_{\sup} \doteq \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[f^2]$
Centralized	$\hat{v}_{\sup} \doteq \sup_{f \in \mathcal{F}} \hat{\mathbb{V}}_{\boldsymbol{x}}[f]$	$v_{\sup} \doteq \sup_{f \in \mathcal{F}} \mathbb{V}_{\mathcal{D}}[f]$



(Empirical) Centralization

The Symmetrization Inequality:

$$\tfrac{1}{2}\mathfrak{R}_m(\mathcal{F} - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathcal{F}], \mathcal{D}) \leq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\sup_{f \in \mathcal{F}}\left|\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\right|\right]}_{\text{Supremum Deviation}} \leq \overbrace{2\mathfrak{R}_m(\mathcal{F} - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathcal{F}], \mathcal{D})}^{\text{Centralized}} \quad \underbrace{2\mathfrak{R}_m(\mathcal{F}, \mathcal{D})}_{\text{Noncentralized}}$$

The *Khintchine* and *Massart* inequalities tell us that

Noncentralized: $\sqrt{\frac{\hat{v}^{\mathrm{raw}}_{\sup}}{2m}} \leq \hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \sqrt{\frac{2\hat{v}^{\mathrm{raw}}_{\sup}\ln(2|\mathcal{F}|)}{m}}$

Centralized: $\sqrt{\frac{\hat{v}_{\sup}}{2m}} \leq \hat{\mathfrak{R}}_m(\mathcal{F} - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathcal{F}], \boldsymbol{x}) \lessapprox \sqrt{\frac{2\hat{v}_{\sup}\ln(2|\mathcal{F}|)}{m}}$

Problem: need to know $\mathbb{E}_{\mathcal{D}}[\mathcal{F}]$ to centralize!
Idea: Use ERA of *empirically centralized family* $\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})$

Lemma: Empirical Centralization Bias

$$\frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})\right]}{1 + \boldsymbol{\Theta}(1/\sqrt{m})} \leq \mathfrak{R}_m(\mathcal{F} - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathcal{F}], \mathcal{D}) \leq \frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})\right]}{1 - \boldsymbol{\Theta}(1/\sqrt{m})}$$

Perceived Problems with the Rademacher Average

- Common criticisms of Rademacher averages
  - "Highly theoretical tool"
  - "Not useful in practice"
  - "Fail to explain" generalization of popular ML methods
- Where do these claims come from?
  - Is the symmetrization inequality tight?
  - Do we have sharp tail bounds on the empirical SD?
  - Are data-dependent bounds on $\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$ tight?
- In standard practice: **no, no, and no!**
  - To "fix" the bounds, we must repair *every part* of them
  - Can we match lower-bounds at every step?
    - ∗ Minimax mean-estimation bound: $m \in \boldsymbol{\Omega}\frac{v_{\sup}\ln\frac{1}{\delta}}{2\varepsilon^2}$
  - The loosest bound is the bottleneck!

Standard McDiarmid Bounds

$$\underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right]}_{\text{RA}} \leq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right]}_{\text{Empirical RA}} + r\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

$$\underbrace{\sup_{f \in \mathcal{F}}\left|\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\right|}_{\text{Supremum Deviation}} \leq 2\underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right]}_{\text{RA}} + r\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

Can we break the

McDiarmid bottleneck?

Target: Bousquet's Inequality

We want dependence on $v_{\sup}$, not $v^{\mathrm{raw}}_{\sup}$ or r^2
In particular, we want to match *Bousquet's inequality*:

$$\sup_{f \in \mathcal{F}}\left|\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\right| \leq 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) + \frac{r\ln\frac{1}{\delta}}{3m} + \sqrt{\frac{2\left(v_{\sup} + 4r\mathfrak{R}_m(\mathcal{F}, \mathcal{D})\right)\ln\frac{1}{\delta}}{m}}$$

We show that $\hat{v}_{\sup}$ is sufficient to *sharply bound* $v_{\sup}$

With High Probability

Theorem: Concentration of Empirically Centralized RAs

With probability at least $1 - \delta$ over choice of $\boldsymbol{x}$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[\mathcal{F}], \boldsymbol{x})\right] \leq \hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[\mathcal{F}], \boldsymbol{x}) + \mathbf{O}\left(\frac{r\ln\frac{1}{\delta}}{m} + \sqrt{\frac{4r\left(\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x}) + r/\sqrt{m}\right)\ln\frac{1}{\delta}}{m}}\right)$$

The *Monte-Carlo* Method: Given $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$, define

$$\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) \doteq \frac{1}{n}\sum_{j=1}^n \sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \boldsymbol{\sigma}_{j,i} f(\boldsymbol{x}_i)\right|$$

Theorem: Monte-Carlo Error Bounds

With probability at least $1 - \delta$ over choice of $\boldsymbol{\sigma} \sim \mathcal{U}^{n \times m}(\pm 1)$:

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2r\ln\frac{1}{\delta}}{3nm} + \sqrt{\frac{4\hat{v}^{\mathrm{raw}}_{\sup}\ln\frac{1}{\delta}}{nm}}$$

$$\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[\mathcal{F}], \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\mathcal{F} - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[\mathcal{F}], \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{4r\ln\frac{1}{\delta}}{3nm} + \sqrt{\frac{4\hat{v}_{\sup}\ln\frac{1}{\delta}}{nm}}$$

Comparative Analysis of Tail Bounds

- Prior Work: with probability at least $1 - \delta$, we may bound the SD as

$$2\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) + \mathbf{O}\left(\frac{r\ln\frac{1}{\delta}}{m} + \sqrt{\frac{(v_{\sup} + r\mathfrak{R}_m(\mathcal{F}, \mathcal{D}))\ln\frac{1}{\delta}}{m}}\right)$$

  - Implicit dependence of $\boldsymbol{\Omega}\sqrt{\frac{v^{\mathrm{raw}}_{\sup}}{m}}$
  - Prior work: Monte-Carlo requires $n \in \boldsymbol{\Omega}\frac{r^2}{v^{\mathrm{raw}}_{\sup}}$ trials
  - This work: $n = 1$ is *asymptotically optimal*
- This Work: With probability at least $1 - \delta$, we may bound the SD as

$$\frac{2\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})}{1 - \boldsymbol{\Theta}\sqrt{\frac{1}{m}}} + \mathbf{O}\left(\frac{r\ln\frac{1}{\delta}}{m} + \sqrt{\frac{(v_{\sup} + r\mathfrak{R}_m(\mathcal{F} - \mathbb{E}_{\mathcal{D}}[\mathcal{F}], \mathcal{D}) + \frac{r^2}{\sqrt{m}})\ln\frac{1}{\delta}}{m}}\right)$$

  - All $v^{\mathrm{raw}}_{\sup}$ dependence moved to $v_{\sup}$ dependence
  - Can substitute $\hat{v}_{\sup}$ and Monte-Carlo ERAs
- Versus Localization
  - Complementary methods: *first* and *second* moment corrections
  - Localization often conflates *raw* and *centralized* variances