Learning with Sample-Dependent Hypothesis Sets

Joint work with
Dylan Foster (MIT)
Spencer Greenberg (Spark Wave)
Satyen Kale (Google Research)
Haipeng Luo (USC)
Karthik Sridharan (Cornell).

MEHRYAR MOHRI MOHRI@

GOOGLE RESEARCH & COURANT INSTITUTE

Motivation

- Common scenario in ML practice:
  - hypothesis set selected **after** receiving training sample.
  - original family restricted after observations.
  - ensemble family decided after receiving sample.
  - regularization chosen using labeled sample.
  - feature transformation or data normalization based on sample.

Motivation

- Standard learning bounds:
  - fixed hypothesis set.
  - selected **before** receiving training sample.
  - guarantees depend on the complexity of hypothesis set.
- Questions:
  - can we derive learning guarantees for sample-dependent hypothesis sets?
  - existing techniques cannot be used; what tools and concepts should we use?

Related Work

- Luckiness framework (Shawe-Taylor et al., 1998): analysis of SRM over data-dependent hierarchies based on concept of luckiness.
  - can be viewed as a study of data-dependent hypothesis sets using luckiness functions and ω-smallness.
  - algorithm-specific guarantees (Herbrich and Williamson, 2002): show some connection with stability, at the price of a strong condition on stability parameter, $\beta = o(\frac{1}{m})$.

Related Work

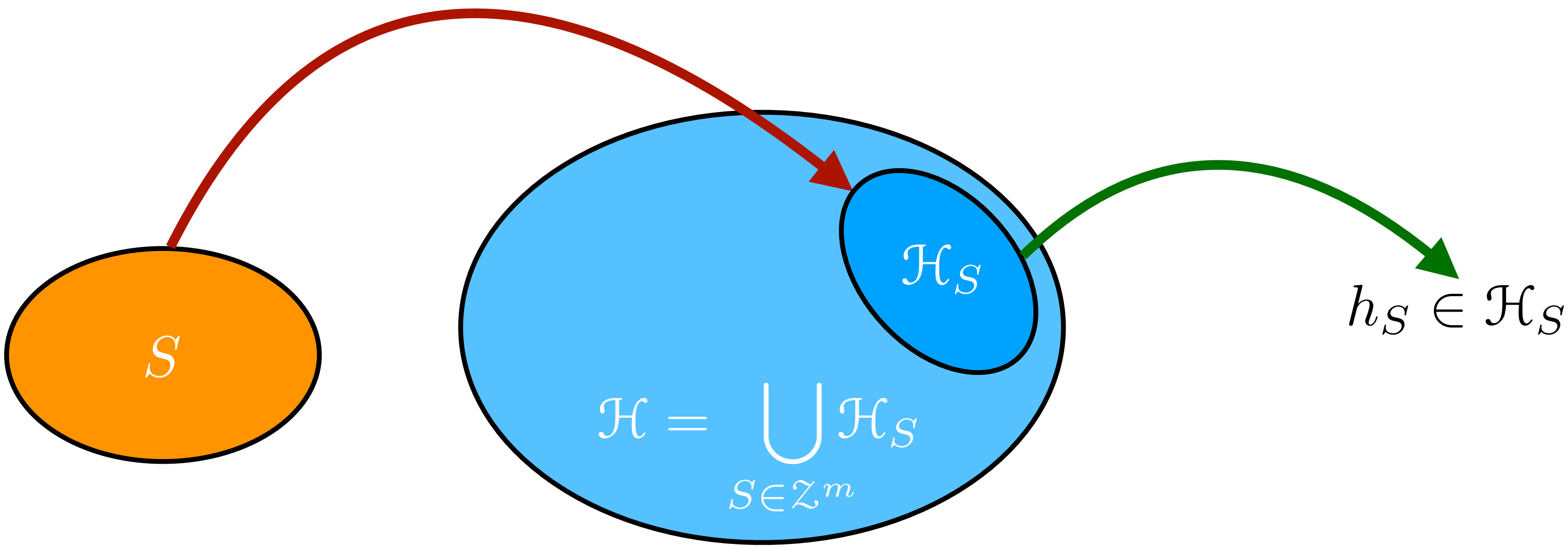
- General bounds for binary classification (Gat 2001; Cannon et al., 2002): expressed in terms a notion of shattering coefficients adapted to data-dependent setting.
- PAC-Bayes bounds (Dziugate and Roy, 2018): prior selected using training sample via a differentially private algorithm.

This Talk

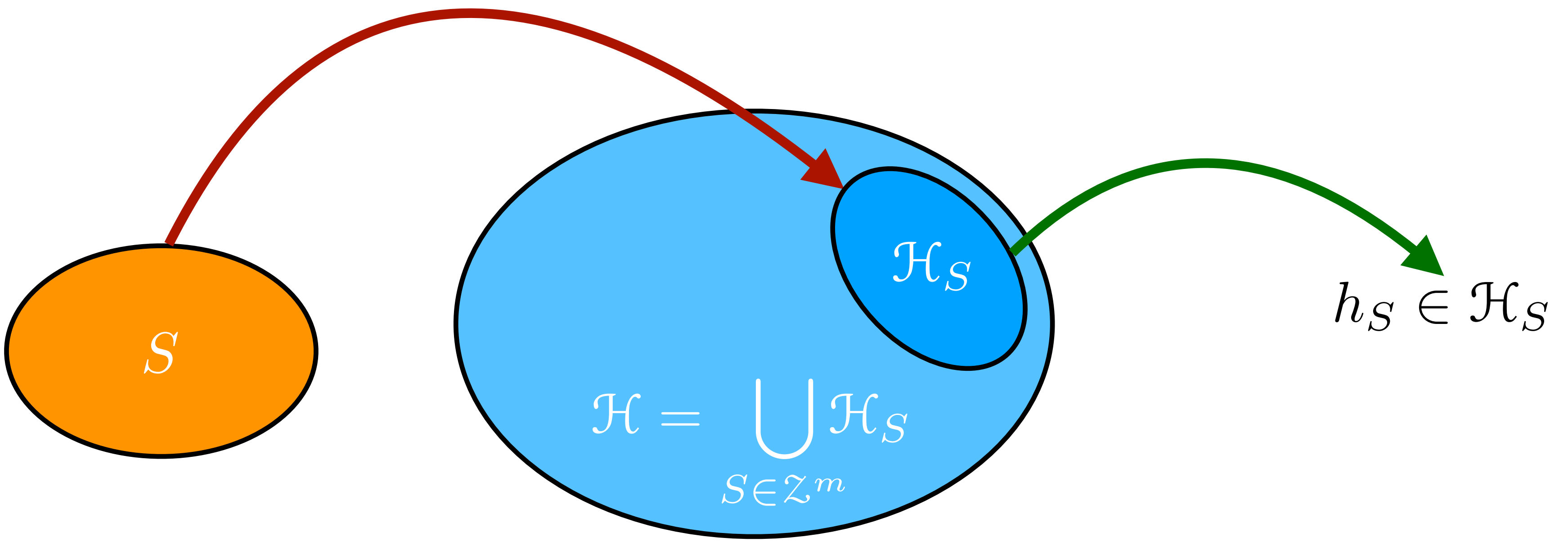
- Setup.
- General sample-dependent guarantees.
- Hypothesis set stability guarantees.
- Applications.

Setup

Learning Stages



Learning Stages



- Special cases:
  - standard generalization: $\mathcal{H}_S = \mathcal{H}$.
  - algorithmic stability: $\mathcal{H}_S = \{h_S\}$.

Definitions

- $\mathcal{X}$ input space, $\mathcal{Y}$ output space, $\mathcal{D}$ distribution over $\mathcal{X} \times \mathcal{Y}$.
- Loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, loss of $h \colon \mathcal{X} \to \mathcal{Y}$ on $z = (x, y)$ denoted $L(h, z) = \ell(h(x), y)$.
- Expected and empirical losses:

$$R(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h, z)]$$

$$\widehat{R}_S(h) = \mathop{\mathbb{E}}_{z \sim S}[L(h, z)] = \frac{1}{m} \sum_{i=1}^{m} L(h, z_i).$$

- Family of losses of hypotheses $\mathcal{G} = (\mathcal{G}_S)_{S \in \mathcal{Z}^m}$:

$$\mathcal{G}_S = \{z \mapsto L(h, z) \colon h \in \mathcal{H}_S\}.$$

General Sample-Dep. Guarantee

Setup

- How can we derive learning bounds for data-dependent hypothesis sets?
  - straightforward idea: use $\overline{\mathcal{H}}_m = \bigcup_{S \in \mathcal{Z}^m} \mathcal{H}_S$; but the family can be very rich and the bound uninformative.
  - alternative: for some supersample U of size $m + n$, consider the family $\overline{\mathcal{H}}_{U,m} = \bigcup_{\substack{S \in \mathcal{Z}^m \\ S \subseteq U}} \mathcal{H}_S$;
  - learning guarantees based on the maximum transductive Rademacher complexity.

Transductive Rad. Complexity

- Definition: transductive Rademacher complexity,

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h \in \overline{\mathcal{H}}_{U,m}} \frac{1}{m+n} \sum_{i=1}^{m+n} \sigma_i L(h, z_i^U)\right],$$

with σ_is independent random variables taking
value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$;
value $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$.

General Learning Bound

- Theorem: let $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ be a family of data-dependent hypothesis sets and let $\mathcal{G}$ be the corresponding family of loss functions. Then, for any $\delta > 0$, with probability $1 - \delta$ over the draw of a sample $S \in \mathcal{Z}^m$, the following holds for all $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + \max_{U \in \mathcal{Z}^{m+n}} 2\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) + 3\sqrt{\left(\tfrac{1}{m} + \tfrac{1}{n}\right)\log\left(\tfrac{2}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m} + \tfrac{1}{n}\right)^3 mn}.$$

Proof Sketch

- Symmetrization lemma (extends to data-dependent case, as observed by Gat (2001)), for $m\epsilon^2 \geq 2$:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S} R(h)-\widehat{R}_S(h)>\epsilon\right]\leq 2\mathop{\mathbb{P}}_{\substack{S\sim\mathcal{D}^m\\T\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}_S}\widehat{R}_T(h)-\widehat{R}_S(h)>\frac{\epsilon}{2}\right].$$

- Concentration bound: upper bound RHS in terms of

$$\mathop{\mathbb{P}}_{\substack{(S,T)\sim U\\|S|=m,|T|=n}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}}\widehat{R}_T(h)-\widehat{R}_S(h)>\frac{\epsilon}{2}\right],$$

  - use extension of McDiarmid's inequality to sampling without replacement (Cortes et al., 2008).
  - bound expectation in terms of Rademacher complexity.

Hypothesis Set Stability Guarantee

Algorithmic Stability

- Definition: for any two samples S and S' differing by one point,

$$\forall z \in \mathcal{Z}, |L(h, z) - L(h', z)| \leq \beta.$$

- Generalization bounds:
  - i.i.d. setting:
    - (Bousquet and Elisseeff, 2002): $O(\beta\sqrt{m} + \frac{1}{\sqrt{m}})$;
    - (Feldman and Vondrak, 2018, 2019): $O(\beta \log^2(m) + \frac{1}{\sqrt{m}})$.
    - (Bousquet et al., 2019): $O(\beta \log(m) + \frac{1}{\sqrt{m}})$.
  - non-i.i.d. stationay (Rostamizadeh and MM, 2010);
  - non-stationary phi- and beta-mixing bounds (Kuznetsov and MM, 2017).

Hypothesis Set Stability

- Definition: a family $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ of data-dependent hypothesis sets is uniformly β-stable if for any two samples S and S' differing by one point,

$$\forall h \in \mathcal{H}_S, \exists h' \in \mathcal{H}_{S'} \colon \ \forall z \in \mathcal{Z}, |L(h, z) - L(h', z)| \leq \beta.$$

Diameter

- Definition: the average diameter, diameter, and maximum diameter of a family $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ of data-dependent hypothesis sets are defined by

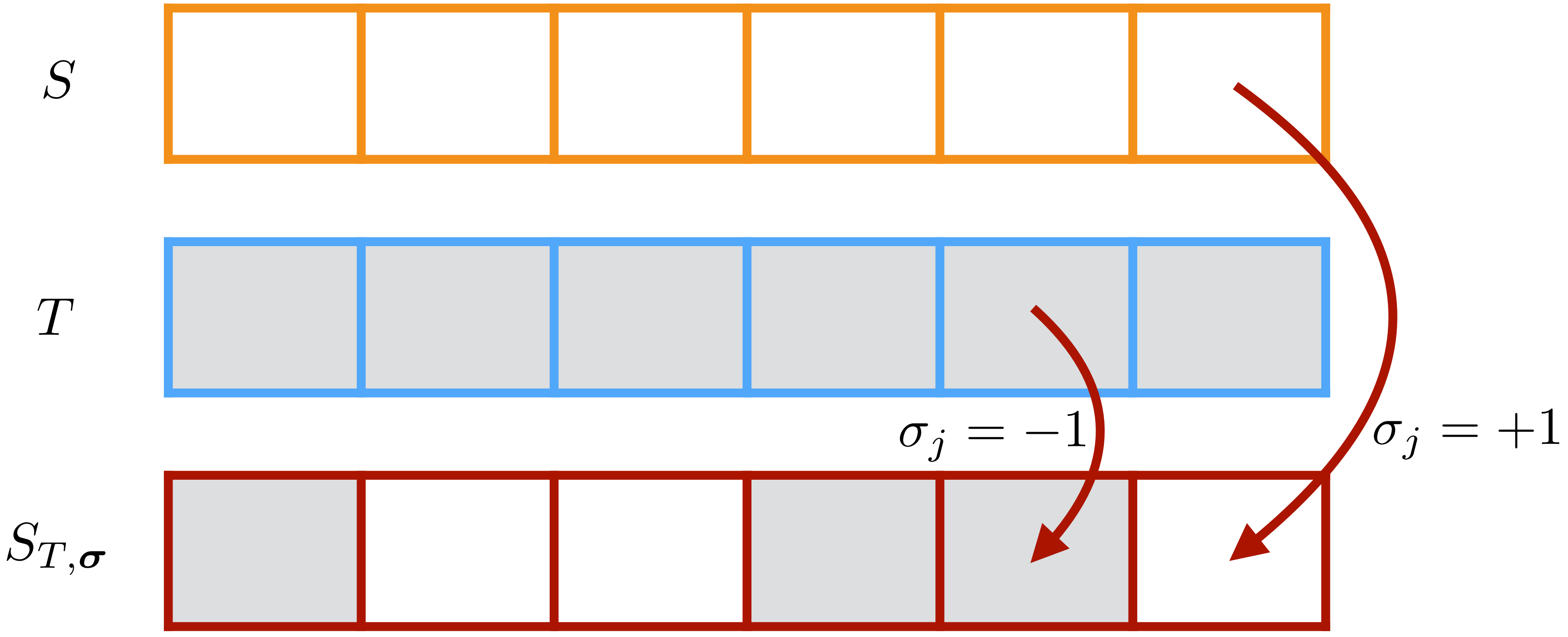
$$\mathop{\mathbb{E}}_{\substack{S \sim \mathcal{Z}^m \\ z \sim S}} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \overline{\Delta}$$

$$\sup_{S \in \mathcal{Z}^m} \mathop{\mathbb{E}}_{z \sim S} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \Delta$$

$$\sup_{\substack{S \in \mathcal{Z}^m \\ z \in S}} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \Delta_{\max}.$$

Rademacher Complexity

- Notation: for samples $S, T \sim \mathcal{Z}^m$ and vector of Rademacher variables $\boldsymbol{\sigma}$, $S_{T,\boldsymbol{\sigma}}$ is defined as follows, and $\mathcal{H}^{\boldsymbol{\sigma}}_{S,T} = \mathcal{H}_{S_{T,\boldsymbol{\sigma}}}$.



Rademacher Complexity

- Empirical Rademacher complexity of $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$:

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right].$$

- Rademacher complexity of $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$:

$$\mathfrak{R}^{\diamond}_{m}(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\substack{S,T \sim \mathcal{D}^m \\ \boldsymbol{\sigma}}} \left[\sup_{h \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right].$$

Properties

- Concentration: for a β-stable family $\mathcal{H}$ with $\beta = O(1/m)$, with high probability,

$$\left| \widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) - \mathfrak{R}^{\diamond}_{m}(\mathcal{H}) \right| \leq O(1/\sqrt{2m}).$$

- Upper bound: let $\mathcal{H}_{S,T} = \bigcup_{\substack{U \subseteq S \cup T \\ U \in \mathcal{Z}^m}} \mathcal{H}_U$, then,

$$\mathfrak{R}^{\diamond}_{m}(\mathcal{H}) \leq \frac{1}{m} \mathop{\mathbb{E}}_{\substack{S,T \sim \mathcal{D}^m \\ \boldsymbol{\sigma}}} \left[\sup_{h \in \mathcal{H}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right] = \mathop{\mathbb{E}}_{S,T \sim \mathcal{D}^m} \left[\widehat{\mathfrak{R}}_T(\mathcal{H}_{S,T}) \right].$$

Example

- For $\mathcal{H}_S$ defined by

$$\mathcal{H}_S = \left\{ x \mapsto w^S \cdot x \colon \;\; w^S = \sum_{i=1}^m \alpha_i x_i^S, \|\alpha\|_1 \leq \Lambda_1 \right\},$$

and $r_T = \sqrt{\frac{\sum_{i=1}^m \|x_i^T\|_2^2}{m}}$ $\quad r_{S \cup T} = \max_{x \in S \cup T} \|x\|_2,$

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) \leq r_T \, r_{S \cup T} \Lambda_1 \sqrt{\frac{2 \log(4m)}{m}} \leq r^2_{S \cup T} \Lambda_1 \sqrt{\frac{2 \log(4m)}{m}}.$$

Hypothesis Stability Bound

- Theorem: let $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ be a β-stable family and let $\mathcal{G}$ be the corresponding family of loss functions. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S \in \mathcal{Z}^m$, the following holds for all $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + \min\{2\mathfrak{R}^{\diamond}_m(\mathcal{G}), \beta + \overline{\Delta}\} + [1 + 2\beta m]\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof Sketch

- Mc-Diarmid's inequality applied to $\Psi(S, S)$ where

$$\Psi(S, S') = \sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_{S'}(h).$$

  - proof of $(\frac{1}{m} + \Delta)$-sensitivity of $\Psi(S, S)$.
  - upper bound on $\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[\Psi(S, S)]$ in terms of Rademacher complexity.

Hypothesis Stability Bound

- **Theorem**: let $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ be a β-stable family and let $\mathcal{G}$ be the corresponding family of loss functions. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S \in \mathcal{Z}^m$, the following holds for all $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + \min \begin{cases} 2\mathfrak{R}^{\diamond}_m(\mathcal{G}) + (1 + 2\beta m)\sqrt{\frac{1}{2m} \log(\frac{1}{\delta})}, \\ \sqrt{e}\,(\beta + \Delta) + 4\sqrt{(\frac{1}{m} + 2\beta) \log(\frac{6}{\delta})}, \\ 47(3\beta + \Delta_{\max}) \log(m) \log(\frac{5m^3}{\delta}) + \sqrt{\frac{4}{m} \log(\frac{4}{\delta})} \end{cases}.$$

Proof

- Proof of second statement: uses a differential privacy-based technique, as in (Feldman and Vondrak, 2018). A key part consists of bounding $\underset{\substack{\mathsf{S}\sim\mathcal{D}^{pm} \\ k=\mathcal{A}(\mathsf{S})}}{\mathbb{E}}[\Psi(S_k, S_k)]$ in terms of χ .
- Proof of third statement: uses the observation that an algorithm choosing a predictor in $\mathcal{H}_S$ is $(\beta + \Delta_{\max})$-stable, and the stability bound of (Feldman and Vondrak, 2018).

Applications

Bagging

- Description:
  - k batches $B_1, \ldots, B_k$ each of size p by sampling with replacement from S.
  - algorithm $\mathcal{A}$ trained on each sample → $\mathcal{A}(B_j)$.
  - $w_i \leq C/k$, for some $C \geq 1$.
  - return convex combination $\sum_{i=1}^k w_i \mathcal{A}(B_i)$; thus,

$$\mathcal{H}_S := \left\{ \sum_{i=1}^k w_i \mathcal{A}(B_i) \colon \; w \in \Delta_k^{C/k} \right\}.$$

Bagging

- Analysis:
  - loss assumed μ-Lipschitz.
  - sampling without replacement.
  - learning bound: whp, for all $h \in \mathcal{H}_S$,

$$R(h) \leq \widehat{R}_S(h) + 2\mu\sqrt{\frac{2p\log(4m)}{m}} + \left[1 + 2\left[p + \sqrt{\frac{2pm\log(\frac{1}{\delta})}{k}}\right] \cdot C\mu\beta_{\mathcal{A}}\right]\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

  - For $p = o(\sqrt{m})$ and $k = \omega(p)$, bound converging regardless of the stability of algorithm $\mathcal{A}$.
  - Somewhat similar but not comparable bound by (Elisseeff et al., 2005).

Stochastic Strongly-Convex Opt.

- Description:
  - uniform convergence bounds do not hold for the stochastic convex optimization problem in general (Shalev-Shwartz et al., 2010).
  - 1st stage: K stochastic strongly-convex optimization algorithms each returning $\widehat{w}_j^S$, $j \in [K]$; these algorithms are $\beta = O(\frac{1}{m})$-sensitive (Shalev-Shwartz et al., 2010).
  - 2nd stage: choose ensemble from

$$\mathcal{H}_S = \left\{ \sum_{j=1}^{K} \alpha_j \widehat{w}_j^S \colon \ \alpha \in \Delta_K \cap \mathsf{B}_1(\alpha_0, r) \right\},$$

with $r = \frac{1}{2\mu D \sqrt{m}}$.

Stochastic Strongly-Convex Opt.

- Analysis:
  - loss assumed μ-Lipschitz.
  - $\mathcal{H}_S$ is shown to be $\mu\beta$-stable.
  - average diameter bound: $\overline{\Delta} \leq \frac{1}{\sqrt{m}}$.
  - learning bound: whp, for all $h \in \mathcal{H}_S$,

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}} \left[L\left(\sum_{j=1}^{K} \alpha_j \widehat{w}_j^S, z \right) \right]$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} L\left(\sum_{j=1}^{K} \alpha_i \widehat{w}_j^S, z_i^S \right) + \sqrt{\frac{e}{m}} + \sqrt{e}\mu\beta + 4\sqrt{\left[\tfrac{1}{m} + 2\mu\beta\right] \log\left[\tfrac{6}{\delta}\right]}.$$

Δ-Sensitive Mappings

- Description:
  - 1st stage: learning mapping $\Phi_S \colon \mathcal{X} \to \mathbb{R}^N$ that is Δ -sensitive with $\Delta = O(\frac{1}{m})$.
  - 2nd stage: select hypothesis from

$$\mathcal{H}_S = \big\{x \mapsto w \cdot \Phi_S(x) \colon \|w\| \leq \gamma\big\}.$$

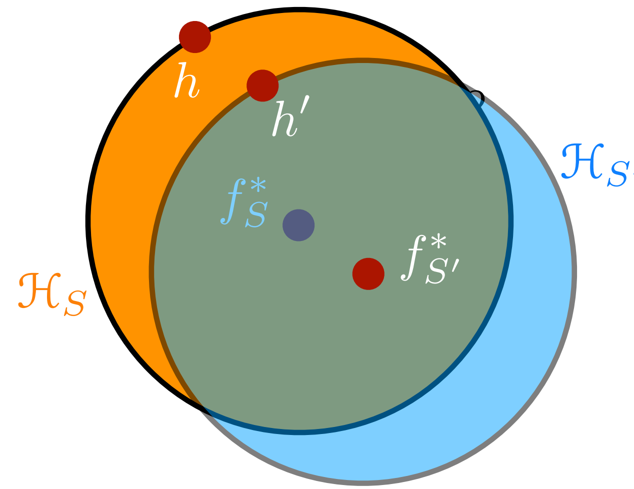
- Analysis:
  - loss assumed μ-Lipschitz.
  - then $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ is $(\mu\gamma\Delta)$-stable, with $\mu\gamma\Delta = O(\frac{1}{m})$.
  - learning bound: whp, for all $h \in \mathcal{H}_S$,

$$R(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}^\diamond_m(\mathcal{G}) + (1 + 2\mu\gamma\Delta m)\sqrt{\frac{1}{2m}\log(\frac{1}{\delta})}.$$

Distillation

- Description:
  - 1st stage: train a very complex model on the training sample S returning $f^*_S\colon \mathcal{X} \to \mathbb{R}$; algorithm assumed β-sensitive:

$$\|f^*_S - f^*_{S'}\| \leq \beta = O(\tfrac{1}{m}).$$

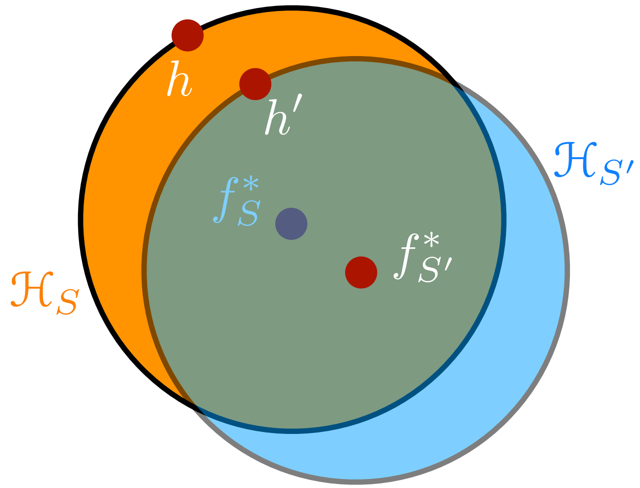


  - 2nd stage: select hypothesis from a less complex family $\mathcal{H}$ with

$$\mathcal{H}_S = \big\{h \in \mathcal{H}\colon \|(h - f^*_S)\|_\infty \leq \gamma\big\}.$$

Distillation

- Analysis:
  - $f^*_{S'} - f^*_S$ assumed in $\mathcal{H} \Rightarrow h' \in \mathcal{H}_{S'}$.
  - loss assumed μ-Lipschitz.
  - → $\mathcal{H}_S$ is $\mu\beta$-stable.
  - learning bound: whp, for all $h \in \mathcal{H}_S$,



$$R(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}^{\diamond}_m(\mathcal{G}) + (1 + 2\mu\beta m)\sqrt{\frac{1}{2m}\log(\frac{1}{\delta})}.$$

Extensions

- Almost everywhere hypothesis set stability.
- Randomized algorithms.
- Data-dependent priors.
- Many other applications.

Conclusion

- Broad analysis of generalization with data-dependent hypothesis sets:
  - hypothesis set stability learning guarantees.
  - applications to many scenarios in practice.
  - other extensions: local Rademacher complexity bouds, model selection bounds.
  - non-i.i.d. learning bounds: stationary beta-mixing processes, discrepancy-based bounds for non-stationary processes.
  - general learning bound for data-dependent hypothesis sets.