Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 9:

Real-Valued Loss:

Covering Numbers, VC-Subgraph Dimension and Rademacher Complexities

Beyond Binary Classification

$$\min_h L(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[loss(h(x); y)]$$

- Until now: $h: \mathcal{X} \to \{\pm 1\}$, $loss(\hat{y}; y) = \left[\left[\hat{y} \neq y\right]\right]$
- More generally:

$$\mathcal{D}(\mathcal{X}, \mathcal{Y}), \qquad h: \mathcal{X} \to \hat{\mathcal{Y}}, \qquad loss: \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$$

- Regression:

$$\mathcal{Y} = \hat{\mathcal{Y}} = \mathbb{R}, \qquad loss(\hat{y}, y) = (y - \hat{y})^2 \text{ or } loss(\hat{y}, y) = |\hat{y} - y|$$

- Multiclass:

$$\mathcal{Y} = \hat{\mathcal{Y}} = \{1, \dots, 10\}, \qquad loss(\hat{y}, y) = \left[\left[\hat{y} \neq y\right]\right] \text{ or cost-sensitive loss}$$

- More complex label spaces, $\mathcal{Y} =$ sequences, graphs, images, sentences…
- Even for binary classification $\mathcal{Y} = \{\pm 1\}$, might want $\hat{\mathcal{Y}} = \mathbb{R}$
  - Hinge loss: $loss(\hat{y}, y) = [1 - \hat{y} \cdot y]_+$
  - Logistic loss: $loss(\hat{y}, y) = \log\left(1 + e^{-\hat{y} \cdot y}\right)$
  - Exp-loss: $loss(\hat{y}, y) = e^{-\hat{y} \cdot y}$

General Learning

$$\min_{h \in \overline{\mathcal{H}}} L(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)]$$

for a given $\ell: \overline{\mathcal{H}} \times \mathcal{Z} \to \mathbb{R}$ and an unknown $\mathcal{D}(\mathcal{Z})$, based on an i.i.d. sample $z_1, \ldots, z_m \sim \mathcal{D}$

- Supervised learning:
  - $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \{z = (x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$
  - $\overline{\mathcal{H}} = \{h: \mathcal{X} \to \hat{\mathcal{Y}}\}$
  - $\ell(h, z) = loss(h(x); y)$
- Proper Learning: compete with $\inf_{h \in \mathcal{H}} L(h)$ restricted to $\overline{\mathcal{H}} = \mathcal{H}$
- Improper Learning: compete with $\inf_{h \in \mathcal{H}} L(h)$ using $\mathcal{H} = \hat{\mathcal{Y}}^{\mathcal{X}}$

General Learning—Examples

$$\min_{h \in \mathcal{H}} L(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)] \text{ based on } z_1, \ldots, z_m \sim iid\ \mathcal{D}$$

- Supervised learning:
  - $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \{z = (x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$
  - $h: \mathcal{X} \to \mathcal{Y}$
  - $\ell(h, z) = loss(h(x); y)$
  - Improper learning: compete with $\mathcal{H} \subset \overline{\mathcal{H}} = \mathcal{Y}^{\mathcal{X}}$
  - Proper learning: $\mathcal{H} = \overline{\mathcal{H}} \subset \mathcal{Y}^{\mathcal{X}}$
- Unsupervised learning, e.g. k-means clustering:
  - $z = x \in \mathbb{R}^d$,
  - $h = (\mu[1], \mu[2], \ldots, \mu[k]) \in \mathbb{R}^{d \times k}$ specified k cluster centers
  - $\ell\big((\mu[1], \mu[2], \ldots, \mu[k]), x\big) = \min_i \|\mu[i] - x\|^2$
- Density estimation:
  - $z = x$ in some measurable space $\mathcal{Z}$ (e.g. $\mathbb{R}^d$)
  - h specifies probability density $p_h(z)$
  - $\ell(h, z) = -\log p_h(z)$
  - Proper learning: fit density model of certain form
  - Improper learning: $\overline{\mathcal{H}} = \{\text{ all probability densities } p(z) \text{ over } \mathcal{Z}\ \}$
- Learning a good route with random traffic:
  - $z =$ traffic delays on each road segment
  - $h =$ route chosen (indicator over road segments in route)
  - $\ell(h, z) = \langle h, z \rangle =$total delay along route

Learning in the General Setting

$$\min_{h \in \overline{\mathcal{H}}} L(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)] \text{ based on } z_1, \ldots, z_m \sim iid\ \mathcal{D}$$

- Generalization of **(agnostic) PAC learning**: Hypothesis class $\mathcal{H} \subseteq \overline{\mathcal{H}}$ agnostically learnable using learning rule $A: \mathcal{Z}^* \to \overline{\mathcal{H}}$ with sample complexity $m(\epsilon, \delta)$ if for all $\mathcal{D}$ and all $\epsilon, \delta > 0$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}} L\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L(h) + \epsilon$
- **ERM**: $ERM_{\mathcal{H}}(S) = \arg\min_{h \in \mathcal{H}} L_S(h) \qquad L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$
- **Uniform Law of Large Numbers (ULLN)**:

$$\Pr_{S \sim \mathcal{D}^m}[\ \forall_{h \in \mathcal{H}} |L(h) - L_S(h)| \leq \epsilon] \geq 1 - \delta$$

- Basic ingredients:
  - LLN: $\forall_h \Pr_{S \sim \mathcal{D}^m}[\ |L(h) - L_S(h)| \leq \epsilon] \geq 1 - \delta$
  - Bound number of different behaviors on S
  - Union bound over different behaviors
  - Symmetrization

Law of Large Numbers— Real Valued Random Variables

- **Theorem** (**Hoefding Bound**): Let $V_1, V_2, \ldots, V_m$ be independent random variables, with $\forall_i \Pr[a \leq V_i \leq b] = 1$, then for $\overline{V} = \frac{1}{m}\sum_{i=1}^{m} V_i$:

$$\Pr\left[\left|\overline{V} - \mathbb{E}\left[\overline{V}\right]\right| > (b-a)\sqrt{\frac{\log 2/\delta}{2m}}\right] < \delta$$

- In our case: $V_i = \ell(h, z_i)$, $\overline{V} = L_S(h)$, $\mathbb{E}\left[\overline{V}\right] = \mathbb{E}[V_i] = L(h)$
- Conclusion: $\forall_h \forall^{\delta}_{S\sim\mathcal{D}^m} |L_s(h) - L(h)| \leq a\sqrt{\frac{\log^2/\delta}{2m}}$
- Applying union bound we can get (same as binary case):
  - Cardinality-based bound for finite classes
  - Description-length bounds
  - PAC-Bayes bounds

Counting Real-Valued Behaviors

- For a class of functions $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\} \subset \mathbb{R}^{\mathbb{Z}}$, and a sample $S \in \mathcal{Z}^m$, consider all possible behaviors

$$\mathcal{F}[S] = \left\{ \left(f(z_1), f(z_2), \ldots, f(z_m)\right) \in \mathbb{R}^m \mid f \in \mathcal{F} \right\}$$

- Maybe infinitely many behaviors, but how many up to some resolution α?
- More specifically: how many behaviors $V \subset \mathbb{R}^m$ can capture all behaviors $\mathcal{F}[S]$ up to error α?
- **Empirical covering numbers**:

$$\mathcal{N}_p(\mathcal{F}, \alpha, S) = \min_{V \subseteq \mathbb{R}^m} |V| \;\; s.t \; \forall_{f \in \mathcal{F}} \exists_{v \in V} \underbrace{\left(\frac{1}{m} \sum_{i=1}^{m} |f(z_i) - v_i|^p \right)^{1/p}}_{\text{For } p = \infty: \sup_i |f(z_i) - v_i|} \leq \alpha$$

$$\mathcal{N}_p(\mathcal{F}, \alpha, m) = \sup_{S \in \mathcal{Z}^m} \mathcal{N}_p(\mathcal{F}, \alpha, S)$$

- At scale $\alpha = 0$, this is the growth function: $\mathcal{N}_p(\mathcal{F}, 0, S) = \Gamma_{\mathcal{F}}(S)$
- But even if $\mathcal{N}_p(\mathcal{F}, 0, S) = \infty$, might be finite for $\alpha > 0$

$$\mathcal{N}_p(\mathcal{F}, \alpha', S) \leq \mathcal{N}_p(\mathcal{F}, \alpha, S) \text{ for } \alpha' > \alpha$$

- Also: $\mathcal{N}_{p'}(\mathcal{F}, \alpha, S) \geq \mathcal{N}_p(\mathcal{F}, \alpha, S)$ for $p' > p$, and in particular:

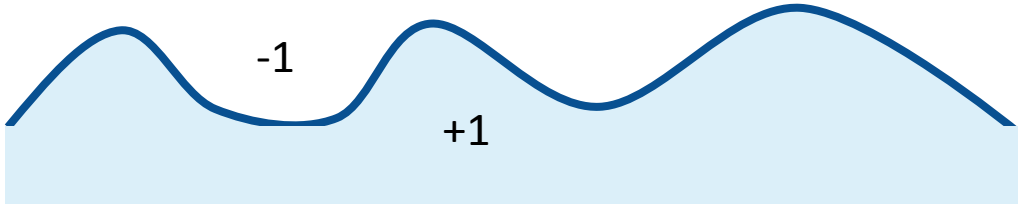
$$\mathcal{N}_1(\mathcal{F}, \alpha, S) \leq \mathcal{N}_2(\mathcal{F}, \alpha, S) \leq \mathcal{N}_\infty(\mathcal{F}, \alpha, S)$$

Controlling Empirical Covering Numbers

- **Definition**: $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ **shatters** $S = \{z_1, \ldots, z_m\}$ if $\exists_{\theta_1,\theta_2,\ldots,\theta_m \in \mathbb{R}}$ s.t. $\forall_{y_1,y_2,\ldots,y_m \in \pm 1} \exists_{f \in \mathcal{F}}$ s.t. $\forall_i$:

$$y_i = +1 \Leftrightarrow f(z_i) > \theta_i$$

- **Definition**: The **VC-subgraph dimension** $\mathrm{VCdim}(\mathcal{F})$ (aka **Pollard's pseudo-dimension**) of $\mathcal{F}$ is the largest m, s.t. there exists $S \in \mathcal{Z}^m$ that it shattered by $\mathcal{F}$
- Equivalently: $\mathrm{VCdim}(\mathcal{F}) = \mathrm{VCdim}\left(\left\{(z,\theta) \mapsto \left[\left[f(z) \leq \theta\right]\right] \mid f \in \mathcal{F}\right\}\right)$



- **Theorem**: For $\mathcal{F} = \{f : \mathcal{Z} \to [-a, a]\}$ with $\mathrm{VCdim}(\mathcal{F}) \leq D$:

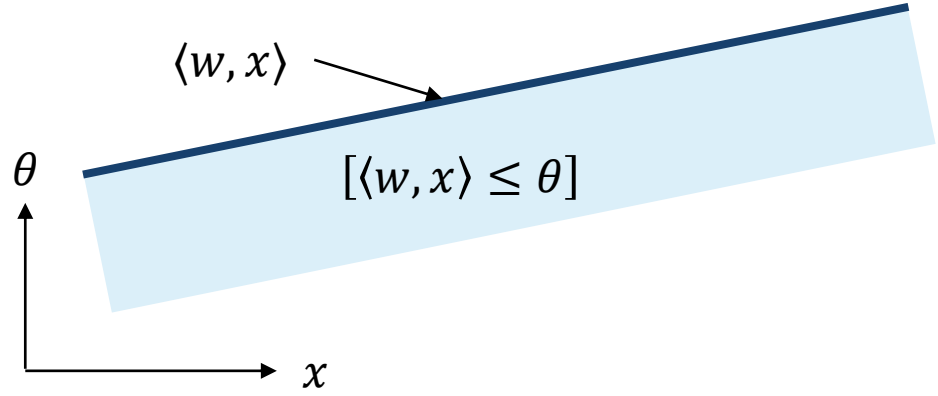
$$\mathcal{N}_p(\mathcal{F}, \alpha, m) \leq \mathcal{N}_\infty(\mathcal{F}, \alpha, m) \leq \sum_{k=1}^{D} \binom{m}{k} \left(\frac{a}{\alpha}\right)^k \leq \left(\frac{em}{D}\frac{a}{\alpha}\right)^D$$

VC Subgraph Dimension: Example

$$\mathcal{H} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d\}$$

- Can shatter the d standard basis vectors, with thresholds $\theta_i = 0$
- Why can't shatter more?
- Consider subgraph class:

$$\{(x, \theta) \mapsto [\langle w, x \rangle \leq \theta]\, w \in \mathbb{R}^d\} \quad \subset \quad \text{half-spaces in } \mathbb{R}^{d+1}$$



- Conclusion: $\mathrm{VCdim}(\mathcal{H}) \leq d + 1$
- In fact, $\mathrm{VCdim}(\mathcal{H}) = d$

What Function Class Should We Be Covering?

$$\min_{h \in \mathcal{H}} L(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)]$$

Supervised learning: $\ell\big(h, z = (x, y)\big) = loss(h(x); y)$, $h: \mathcal{X} \to \mathcal{Y}$

- $\mathcal{N}_p(\mathcal{H}, \alpha, S)$?
  - $h: \mathcal{X} \to \mathcal{Y}$ might not be real valued (e.g. $\mathcal{Y}$ =sentences)
  - In fact, for general learning problem, h might not be a function at all
  - Even if $h: \mathcal{X} \to \mathbb{R}$, we care about scale of loss, which might be different from scale of predictions
- Loss class:

$$\mathcal{F} = \{z \mapsto \ell(h, z) \mid h \in \mathcal{H}\}$$

Supervised learning: $\mathcal{F} = \{(x, y) \mapsto loss(h(x); y) \mid h \in \mathcal{H}\}$

- Claim: For supervised learning with $\mathcal{H} = \{h: \mathcal{X} \to \mathbb{R}\}$,
  - If $loss(\hat{y}, y)$ is monotone in $\hat{y}$ (for each y): $\mathrm{VCdim}(\mathcal{F}) \leq \mathrm{VCdim}(\mathcal{H})$
  - If $loss(\hat{y}, y)$ is unimodal in $\hat{y}$ (for each y): $\mathrm{VCdim}(\mathcal{F}) \leq 2\mathrm{VCdim}(\mathcal{H})$

Back to the Plan

ULLN: $\forall_S^{\delta} \forall_{h \in \mathcal{H}} |L(h) - L_S(h)| \leq \epsilon$

$\Downarrow$

Learning: $\forall_S^{\delta} L\left(ERM_{\mathcal{H}}(S)\right) \leq \inf_{h \in \mathcal{H}} L(h) + 2\epsilon$

- Plan for establishing ULLN:
  - LLN: $\forall_h \forall_S^{\delta} |L(h) - L_S(h)| \leq \epsilon$ ✓
  - Counting behaviors ✓
  - Union bond
  - Symmetrization

- Old-School approach:
  - Use $\mathcal{N}_{\infty}(\mathcal{F}, \alpha = \epsilon/4\,, 2m)$
  - Approximate $\ell(h, z_i)$ with v in cover, incurring $\leq \epsilon/2$ error
  - Union bound over v in cover
  - Symmetrization as in binary case
- Contemporary approach: use Rademacher Complexity

Rademacher Complexity

- Recall loss class $\mathcal{F} = \{z \mapsto \ell(h, z) \mid h \in \mathcal{H}\}$
- For $f \in \mathcal{F}$, denote: $\mathbb{E}_{\mathcal{D}}[f] = \mathbb{E}_{z\sim\mathcal{D}}[f(z)]$ and $\mathbb{E}_S[f] = \frac{1}{m}\sum_{i=1}^m f(z_i)$
- Symmetrization: as a surrogate for $\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f]$, use $\mathbb{E}_{S_1}[f] - \mathbb{E}_{S_2}[f]$
- Starting with $S = \{z_1, \ldots, z_m\}$, use $\xi_1, \ldots, \xi_m \in \pm 1$ to define $S_1 = \{z_i | \xi_i = +1\}$ and $S_2 = \{z_i | \xi_i = -1\}$, then if $|S_1| = |S_2|$:

$$\mathbb{E}_{S_1}[f] - \mathbb{E}_{S_2}[f] = \frac{2}{m}\sum_{i=1}^m \xi_i f(z_i)$$

- **Empirical Rademacher Complexity of $\mathcal{F}$:**

$$\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_{\xi_1,\xi_2,\ldots,\xi_m\sim\text{iid unif}\pm 1}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \xi_i f(z_i)\right]$$

- **Theorem**:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{F}}(\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[S])\right] \leq 2\mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{R}_S(\mathcal{F})]$$

Rademacher Complexity ➔ ULLN

$$\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_{\xi_1,\xi_2,\ldots,\xi_m \sim \text{iid unif} \pm 1}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \xi_i f(z_i)\right]$$

$$\mathcal{R}_m(\mathcal{F}) = \sup_{S\in\mathcal{Z}^m} \mathcal{R}_S(\mathcal{F}) \quad \mathcal{R}_{\mathcal{D}^m}(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{R}_S(\mathcal{F})]$$

- **Theorem**: $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{F}}(\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f])\right] \le 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F})$

Proof: $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_f(\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f])\right] = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_f\left(\mathbb{E}_{S'\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_i f(z_i')\right] - \frac{1}{m}\sum_i f(z_i)\right)\right]$

$= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_f \mathbb{E}_{S'\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_i\left(f(z_i') - f(z_i)\right)\right]\right]$

$\le \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_f \frac{1}{m}\sum_i\left(f(z_i') - f(z_i)\right)\right] = \mathbb{E}_{S,S'\sim\mathcal{D}^m,\xi}\left[\sup_f \frac{1}{m}\sum_i \xi_i\left(f(z_i') - f(z_i)\right)\right]$

$\le \mathbb{E}_{S'\sim\mathcal{D}^m,\xi}\left[\sup_f \frac{1}{m}\sum_i \xi_i f(z_i')\right] + \mathbb{E}_{S\sim\mathcal{D}^m,\xi}\left[\sup_f \frac{1}{m}\sum_i(-\xi_i) f(z_i)\right] = 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F})$

- Corollary: $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{F}}(\mathbb{E}_S[f] - \mathbb{E}_{\mathcal{D}}[f])\right] \le 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F})$

Proof: consider $(-\mathcal{F})$ and note that $\mathcal{R}(-\mathcal{F}) = \mathcal{R}(\mathcal{F})$

- What about high probability guarantees?

McDiarmid Inequality

- Definition: $g: \mathcal{Z}^m \to \mathbb{R}$ satisfies the **bounded difference property** with $c_1, \ldots, c_m \in \mathbb{R}$ if

$$\forall_i \forall_{z_1, z_2, \ldots, z_m, z_i'} |g(z_1, \ldots, z_m) - g(z_1, \ldots, z_i', \ldots, z_m)| \leq c_i$$

- **Theorem (McDiarmid)**: Let $Z, \ldots, Z_m$ be independent random variables over $\mathcal{Z}$. If g satisfies the bounded difference property then $\forall^{\delta}_{z_1, \ldots, z_m}$

$$|g(z_1, \ldots, z_m) - \mathbb{E}[g(Z_1, \ldots, Z_m)]| \leq \sqrt{\frac{1}{2} \sum_{i=1}^{m} c_i^2 \log {}^2/_{\delta}}$$

- Generalizes Hoeffding, where $g(z_1, \ldots, z_m) = \frac{1}{m} \sum_i z_i$
  - If $a \leq z_i \leq b$ then g satisfies bounded difference with $c_i = \frac{1}{m}(b - a)$
  - $\sqrt{\frac{1}{2} \sum_{i=1}^{m} c_i^2 \log {}^2/_{\delta}} = \sqrt{\frac{1}{2} m \left(\frac{b-a}{m}\right)^2 \log {}^2/_{\delta}} = (b - a) \sqrt{\frac{\log {}^2/_{\delta}}{2m}}$

Rademacher ➔ High Probability ULLN

- Use $g(S) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f])$
- If $0 \leq f(z) = \ell(h, z) \leq a$ then $g(S)$ satisfies bounded diff with $c_i = \frac{a}{m}$
- Recall $\mathbb{E}_{S \sim \mathcal{D}^m} \left[\sup_{f \in \mathcal{F}} (\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f]) \right] \leq 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F})$
- Conclusion: $\forall_S^{\delta} \sup_{f \in \mathcal{F}} (\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f]) \leq 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F}) + a\sqrt{\frac{\log^2/\delta}{2m}}$
- And similarly for $\mathbb{E}_S[f] - \mathbb{E}_{\mathcal{D}}[f]$, yielding ULLN:

$$\forall_{S \sim \mathcal{D}^m}^{\delta} \sup_{f \in \mathcal{F}} |\mathbb{E}_{\mathcal{D}}[f] - \mathbb{E}_S[f]| \leq 2\mathcal{R}_{\mathcal{D}^m}(\mathcal{F}) + a\sqrt{\frac{\log 4/\delta}{2m}}$$

Bounding the Rademacher Complexity

- How do we bound $\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \xi_i f(z_i) \right]$?
- For a finite $\mathcal{F}$, apply Hoefding to the tail of each $\sum_{i=1}^{m} \xi_i f(z_i)$ and take a union bound, yielding (Massart's Finite Class Lemma):

$$\mathcal{R}_m(\mathcal{F}) \leq a \sqrt{\frac{\log|\mathcal{F}|}{2m}}$$

- Since $\mathcal{R}_S(\mathcal{F})$ only depends on behavior inside S, can take union bound over behaviors:

$$\mathcal{R}_m(\mathcal{F}) \leq a \sqrt{\frac{\log \Gamma_{\mathcal{F}}(m)}{2m}} = a \sqrt{\frac{\log \mathcal{N}_p(\mathcal{F}, \alpha = 0, m)}{2m}}$$

- Or over behaviors up to resolution α:

$$\mathcal{R}_m(\mathcal{F}) \leq \inf_{\alpha} \left(\alpha + a \sqrt{\frac{\log \mathcal{N}_1(\mathcal{F}, \alpha, m)}{2m}} \right)$$

Bounding the Rademacher Complexity

- Theorem: For $0 \le f \le a$:

$$\mathcal{R}_S(\mathcal{F}) \le \inf_{\alpha}\left(\alpha + a\sqrt{\frac{\log \mathcal{N}_1(\mathcal{F},\alpha,S)}{2m}}\right)$$

Proof: Consider a cover V, $|V| = \log \mathcal{N}_1(\mathcal{F},\alpha,S)$. For every f, let $v^f \in V$ be s.t. $\frac{1}{\mathrm{m}}\sum_i |f(z_i) - v_i| \le \alpha$

$$\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_\xi\left[\sup_f \frac{1}{m}\sum_i \xi_i f(z_i)\right] = \mathbb{E}_\xi\left[\sup_f \frac{1}{m}\sum_i \xi_i\left(v_i^f + f(z_i) - v_i^f\right)\right]$$

$$\le \mathbb{E}_\xi\left[\sup_f \frac{1}{m}\sum_i \xi_i v_i^f\right] + \frac{1}{m}\sum_i |f(z_i) - v_i^f| \le a\sqrt{\frac{\log|V|}{2m}} + \alpha$$

- Conclusion:

$$\mathcal{R}_m(\mathcal{F}) \le \inf_{\alpha}\left(\alpha + a\sqrt{\frac{\mathrm{VC}(\mathcal{F})\log\left(\frac{em}{\mathrm{VC}(\mathcal{F})}\frac{a}{\alpha}\right)}{2m}}\right) \le a\sqrt{\frac{2\mathrm{VC}(\mathcal{F})\log\left(\frac{em}{\mathrm{VC}(\mathcal{F})}\right)}{m}}$$

$$\alpha = a\sqrt{\frac{\mathrm{VCd}(\mathcal{F})}{2m}}$$

Dudley's Integral

- Can improve over

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha}\left(\alpha + a\sqrt{\frac{\log \mathcal{N}_1(\mathcal{F},\alpha,S)}{2m}}\right)$$

- Instead of using cover at a single scale, integrate over all scales:

$$\mathcal{R}_S(\mathcal{F}) \leq \inf_{\alpha_0}\left(4\alpha_0 + 10\int_{\alpha_0}^{a}\sqrt{\frac{\log \mathcal{N}_2(\mathcal{F},\alpha,S)}{m}}\,d\alpha\right)$$

- Allows avoiding log-factor and obtaining:

$$\mathcal{R}_m(\mathcal{F}) = O\left(a\sqrt{\frac{\mathrm{VCdim}(\mathcal{F})}{m}}\right)$$

- Even more important in infinite-dim classes

Putting It All Together

- For supervised learning, $\ell\big(h,(x,y)\big) = loss(h(x);y)$, when
  - $0 \le loss(\hat{y};y) \le a$
  - $loss(\hat{y};y)$ is monotone or unimodal in $\hat{y}$

  then with probability $\ge 1-\delta$ over $S \sim \mathcal{D}^m$:

$$L_{\mathcal{D}}\big(ERM_{\mathcal{H}}(S)\big) \le \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + O\left(a\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}\right)$$

- Need loss to be bounded
  - OK for non-convex loss, e.g. 0/1, or truncated squared loss
  - Can bound targets y and "response" $h(x)$
  - Only needed for high-probability bound (unavoidable!)---can get small error in expectation even without bounded loss.