INTRODUCTION TO STATISTICAL LEARNING THEORY

J. Saketha Nath (IIT Bombay)

What is STL?

"The goal of statistical learning theory is to study, in a statistical framework, the properties of learning algorithms"

– [Bousquet *et.al.*, 04]

Supervised Learning Setting

- Given:
  - ***Training data:*** $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
  - ***Model:*** *set of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - ***Loss function:*** $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
  - *Model: set of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: ?? Pick a candidate that does **well on** new data ??

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
  - *Model: set of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: ?? Pick a candidate that does well on new data ??

- Assumptions:
  - *There exists* $\boldsymbol{F_{XY}}$ *that* ***generates*** $\boldsymbol{D}$ *as well as* ***"new data"*** *(Stochastic framework)*
  - *iid samples and bounded, Lipschitz loss*

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
  - *Model: set* $\mathcal{G}$ *of candidate predictors of the form* $g\colon \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l\colon \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: $\mathrm{g}^* = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, \boldsymbol{E}[\boldsymbol{l}(\boldsymbol{Y}, \boldsymbol{g}(\boldsymbol{X}))]$

- Assumptions:
  - *There exists* $\boldsymbol{F}_{\boldsymbol{XY}}$ *that generates* D *as well as "new data"*
  - *iid samples and bounded, Lipschitz loss*

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$
  - *Model: set* $\mathcal{G}$ *of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: $g^* = \operatorname*{argmin}_{g \in \mathcal{G}} E[l(Y, g(X))]$

Minimize expected loss
(a.k.a. risk $R_l[g]$ minimization)

- Assumptions:
  - *There exists* F_{XY} *that generates* D *as well as "new data"*
  - *iid samples and bounded, Lipschitz loss*

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
  - *Model: set* $\mathcal{G}$ *of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: $\mathrm{g}^* = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, E[l(Y, g(X))]$

Well-defined, but un-realizable.

- Assumptions:
  - *There exists* F_{XY} *that generates* D *as well as "new data"*
  - *iid samples and bounded, Lipschitz loss*

Supervised Learning Setting

- Given:
  - *Training data:* $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$
  - *Model: set* $\mathcal{G}$ *of candidate predictors of the form* $g: \mathcal{X} \mapsto \mathcal{Y}$
  - *Loss function:* $l: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

- Goal: $g^* = \operatorname*{argmin}_{g \in \mathcal{G}} E[l(Y, g(X))]$

How well can we approximate?

- Assumptions:
  - *There exists* F_{XY} *that generates* D *as well as "new data"*
  - *iid samples and bounded, Lipschitz loss*

Skyline ?

- Case of $|\mathcal{G}| = 1$ (estimate error rate)
  - *Law of large numbers:* $\left\{\frac{1}{m}\sum_{i=1}^{m} l\big(Y_i, g(X_i)\big)\right\}_{m=1}^{\infty} \xrightarrow{\text{p}} E[l(Y, g(X))]$

With high probability,
average loss (a.k.a. empirical risk)
on (a large) training set is a good
approximation for risk

Skyline ?

- Case of $|\mathcal{G}| = 1$ (estimate error rate)
  - *Law of large numbers:* $\left\{\frac{1}{m}\sum_{i=1}^{m} l\big(Y_i, g(X_i)\big)\right\}_{m=1}^{\infty} \xrightarrow{\text{p}} E[l(Y, g(X))]$

For given (but any) $F_{XY}, \delta > 0, \epsilon > 0$, we have that:

There exists $m_0(\delta, \epsilon) \in \mathbb{N}$, such that

$$P\left[\left|\frac{1}{m}\sum_{i=1}^{m} l\big(Y_i, g(X_i)\big) - E[l(Y, g(X))]\right| > \epsilon\right] \leq \delta$$

for all $m \geq m_0(\delta, \epsilon)$.

Some Definitions

- A problem $(\mathcal{G}, l)$ is learnable iff there exists an algorithm that selects $\hat{g}_m \in \mathcal{G}$ such that for any $\mathrm{F}_{XY}, \delta > 0, \epsilon > 0$, we have that there exists $m_0(\delta, \epsilon) \in \mathbb{N}$, such that

$$\boldsymbol{P[R_l[\hat{g}_m] - R_l[g^*] > \epsilon] \leq \delta} \textbf{ for all } \boldsymbol{m \geq m_0(\delta, \epsilon)}.$$

  - *g^* is the (true) risk minimizer*

Some Definitions

- A problem $(\mathcal{G}, l)$ is **learnable** iff there exists an algorithm that selects $\hat{g}_m \in \mathcal{G}$ such that for any $\mathrm{F}_{XY}, \delta > 0, \epsilon > 0$, we have that there exists $m_0(\delta, \epsilon) \in \mathbb{N}$, such that

$$\boldsymbol{P[R_l[\hat{g}_m] - R_l[g^*] > \epsilon] \leq \delta} \textbf{ for all } \boldsymbol{m \geq m_0(\delta, \epsilon)}.$$

  - *g^* is the (true) risk minimizer*

- Such an algorithm is called universally consistent $m_0(\delta, \epsilon)$ may depend on F_{XY}
- (Smallest) m_0 is called sample complexity of the problem
  - *Analogously sample complexity of algorithm*

Some Algorithms

SAMPLE AVERAGE APPROXIMATION
(a.k.a Empirical Risk Minimization)

1. $\min_{g \in \mathcal{G}} E[l(Y, g(X))] \approx \min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} l\big(y_i, g(x_i)\big)$

   (consistent estimator approximation)

2. Bounds based on concentration of mean
3. Indirect bounds (choice optimization alg.)

[Vapnik, 92]

Some Algorithms

SAMPLE AVERAGE APPROXIMATION
(a.k.a Empirical Risk Minimization)

1. $\min_{g \in \mathcal{G}} E[l(Y, g(X))] \approx \min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} l(y_i, g(x_i))$

   Minimize error on training set $\widehat{\boldsymbol{R}}_{\boldsymbol{m}}[\boldsymbol{g}]$

   (consistent estimator approximation)

2. Bounds based on concentration of mean
3. Indirect bounds (choice optimization alg.)

[Vapnik, 92]

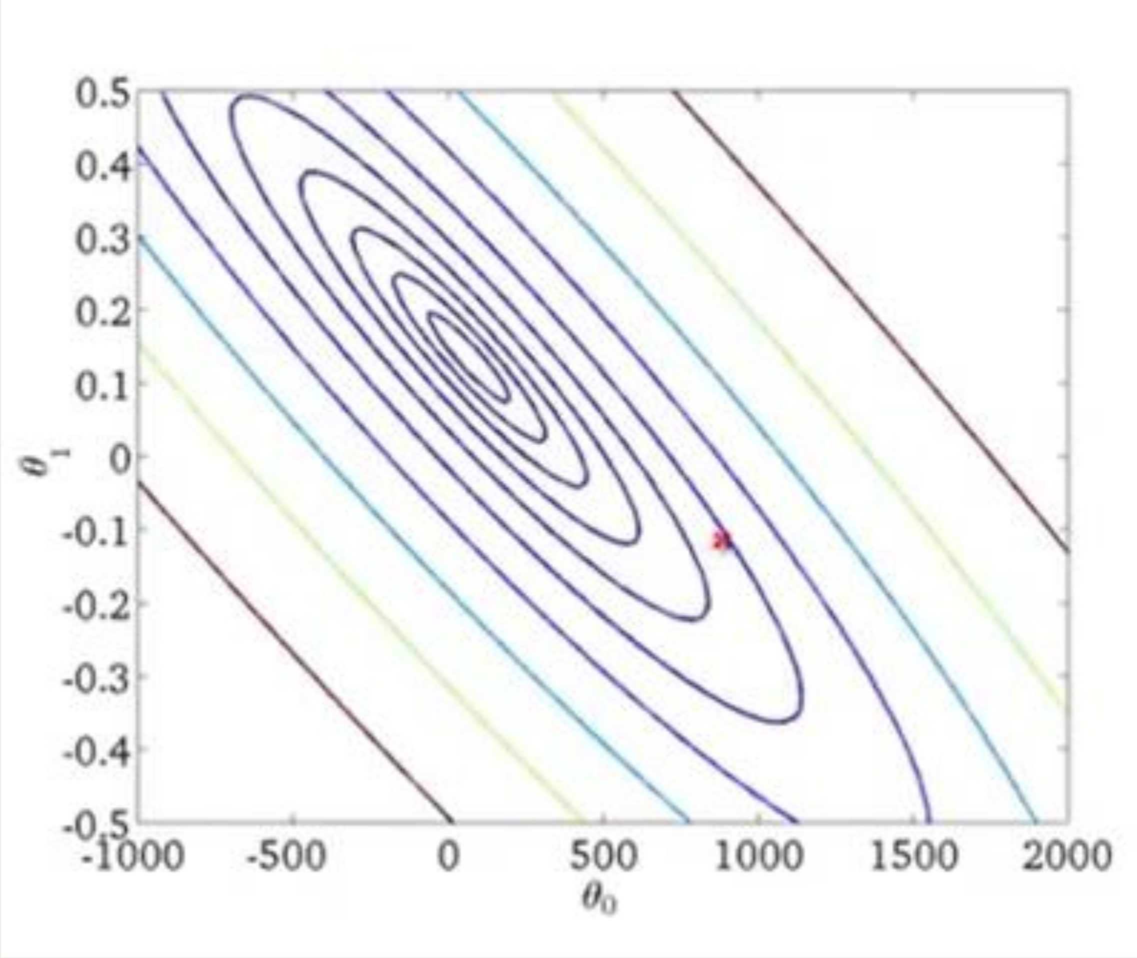
Some Algorithms

SAMPLE AVERAGE APPROXIMATION
(a.k.a Empirical Risk Minimization)

1. $\min_{g \in \mathcal{G}} E[l(Y, g(X))] \approx \min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} l(y_i, g(x_i))$

   (consistent estimator approximation)

2. Bounds based on concentration of mean
3. Indirect bounds (choice optimization alg.)



[Vapnik, 92]

https://www.coursera.org/course/ml

Some Algorithms

SAMPLE AVERAGE APPROXIMATION
(a.k.a Empirical Risk Minimization)

1. $\min_{g\in\mathcal{G}} E[l(Y, g(X))] \approx \min_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} l(y_i, g(x_i))$

   (consistent estimator approximation)

2. Bounds based on concentration of mean
3. Indirect bounds (choice optimization alg.)

[Vapnik, 92]

SAMPLE APPROXIMATION
(a.k.a Stochastic Gradient Descent)

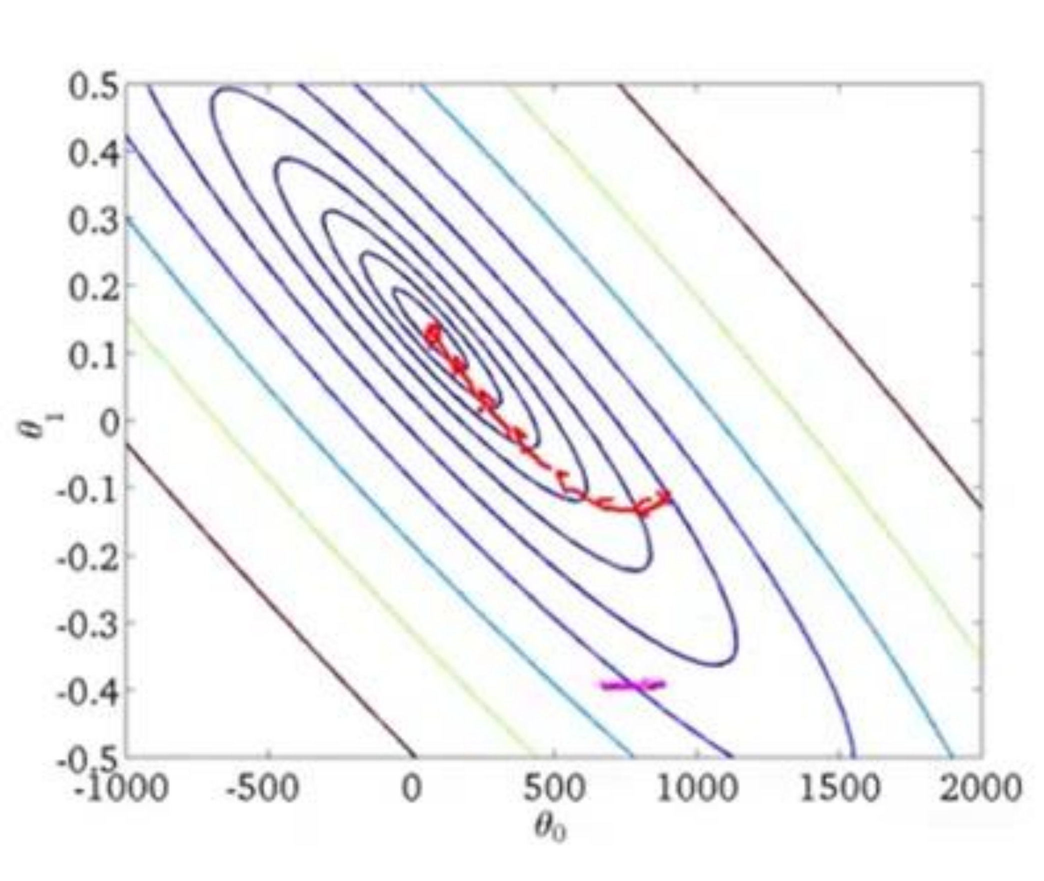
1. Update $g^{(k)}$ using $l(y_k, x_k)$ and $\hat{g} \equiv \frac{1}{m}\sum_{k=1}^{m} g^{(k)}$

   (weak estimator approximation)

2. Online learning literature
3. Direct bounds on risk

[Robbins & Monro, 51]

Some Algorithms



https://www.coursera.org/course/ml

SAMPLE APPROXIMATION
(a.k.a Stochastic Gradient Descent)

1. Update $g^{(k)}$ using $l(y_k, x_k)$ and $\hat{g} \equiv \frac{1}{m}\sum_{k=1}^{m} g^{(k)}$

   (weak estimator approximation)

2. Online learning literature
3. Direct bounds on risk

[Robbins & Monro, 51]

Some Algorithms

SAMPLE AVERAGE APPROXIMATION
(a.k.a Empirical Risk Minimization)

1. $\min_{g \in \mathcal{G}} E[l(Y, g(X))] \approx \min_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} l(y_i, g(x_i))$

   (consistent estimator approximation)
2. Bounds based on concentration of mean
3. Indirect bounds (choice optimization alg.)

Focus of this talk

[Vapnik, 92]

SAMPLE APPROXIMATION
(a.k.a Stochastic Gradient Descent)

1. Update $g^{(k)}$ using $l(y_k, x_k)$ and $\hat{g} \equiv \frac{1}{m} \sum_{k=1}^{m} g^{(k)}$

   (weak estimator approximation)
2. Online learning literature
3. Direct bounds on risk

Summary of results

[Robbins & Monro, 51]

ERM consistency: Sufficient conditions

- $0 \leq R[\hat{g}_m] - R[g^*] = R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] + \hat{R}_m[\hat{g}_m] - \hat{R}_m[g^*] + \hat{R}_m[g^*] - R[g^*]$

ERM consistency: Sufficient conditions

- $0 \le R[\hat{g}_m] - R[g^*] = R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] + \hat{R}_m[\hat{g}_m] - \hat{R}_m[g^*] + \hat{R}_m[g^*] - R[g^*]$
- $\le \left(\max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g]\right) + \underbrace{\hat{R}_m[g^*] - R[g^*]}_{\xrightarrow{p} 0 \quad \because \mathrm{LLN}}$

ERM consistency: Sufficient conditions

- $0 \le R[\hat{g}_m] - R[g^*] = R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] + \hat{R}_m[\hat{g}_m] - \hat{R}_m[g^*] + \hat{R}_m[g^*] - R[g^*]$
- $\le \left(\max_{g\in\mathcal{G}} R[g] - \hat{R}_m[g]\right) + \underbrace{\hat{R}_m[g^*] - R[g^*]}_{\xrightarrow{p} 0 \quad \because \text{LLN}}$

- Hence **one-sided uniform convergence** is a sufficient condition for ERM consistency
  - *i.e.,* $\left\{\max_{g\in\mathcal{G}} R[g] - \hat{R}_m[g]\right\}_{m=1}^{\infty} \xrightarrow{p} 0$ *as* $m \to \infty$

ERM consistency: Sufficient conditions

- $0 \leq R[\hat{g}_m] - R[g^*] = R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] + \hat{R}_m[\hat{g}_m] - \hat{R}_m[g^*] + \hat{R}_m[g^*] - R[g^*]$
- $\leq \left(\max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g]\right) + \underbrace{\hat{R}_m[g^*] - R[g^*]}_{\xrightarrow{p} 0 \quad \because \text{LLN}}$

- Hence one-sided uniform convergence is a sufficient condition for ERM consistency
  - *i.e.,* $\left\{\max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g]\right\}_{m=1}^{\infty} \xrightarrow{p} 0$ *as* $m \to \infty$
  - *Vapnik proved this is necessary for “non-trivial” consistency (of ERM)*

Story so far …

- **Two algorithms:** Sample Average Approx., Sample Approx.
- One-sided **uniform convergence** of mean is sufficient for SAA consistency.

Candidate for Problem Complexity

$$\max_{g \in \mathcal{G}} \quad R \quad [g] \; - \hat{R}_m[g]$$

Candidate for Problem Complexity

$$E\left[\max_{g\in\mathcal{G}} \quad R \quad [g] \; - \hat{R}_m[g]\right]$$

Candidate for Problem Complexity

$$E\left[\max_{g\in\mathcal{G}} \quad R \quad [g] \; - \hat{R}_m[g]\right]$$

1. Ensure (asymptotically) goes to zero.
2. Show concentration around mean for max. div.

Candidate for Problem Complexity

$$E\left[\max_{g\in\mathcal{G}} \quad R\ \ [g]\ - \hat{R}_m[g]\right]$$

Candidate for Problem Complexity

$$E\left[\max_{g \in \mathcal{G}} E\left[\hat{R}'_m[g]\right] - \hat{R}_m[g]\right]$$

Candidate for Problem Complexity

$$\leq E\left[\max_{g \in \mathcal{G}} \quad \hat{R}'_m[g] \; - \hat{R}_m[g]\right]$$

MAXIMUM DISCREPANCY

Towards Rademacher Complexity

$$E\left[\max_{g\in\mathcal{G}} \quad \hat{R}'_m[g] \; - \hat{R}_m[g]\right]$$

Towards Rademacher Complexity

$$E\left[\max_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(l\big(Y_i',g(X_i')\big)-l\big(Y_i,g(X_i)\big)\right)\right)\right]$$

Towards Rademacher Complexity

$$E_\sigma E\left[\max_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(l\big(Y_i', g(X_i')\big) - l\big(Y_i, g(X_i)\big)\right)\right)\right]$$

iid Rademacher random variables $P[\sigma_i = 1] = 0.5$, $P[\sigma_i = -1] = 0.5$.

Rademacher Complexity

$$\leq 2\,E\underbrace{\left[\underbrace{E_\sigma\left[\max_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i l\big(Y_i, g(X_i)\big)\right)\right]}_{\text{Empirical term}}\right]}_{\text{Distribution-dependent term}}$$

Rademacher Complexity

$$= 2\,\underbrace{E\left[\underbrace{E_\sigma\left[\max_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i l\big(Y_i, g(X_i)\big)\right)\right]}_{\text{Empirical term}}\right]}_{\text{Distribution}-\text{dependent term}}$$

$f(Z_i)$

$f \in \mathcal{F}$

Rademacher Complexity

$$= 2\,E\underbrace{\left[\underbrace{E_\sigma\left[\max_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(Z_i)\right)\right]}_{\hat{\mathcal{R}}_m(\mathcal{F})}\right]}_{\mathcal{R}_m(\mathcal{F})}$$

$\mathcal{R}_m(\mathcal{F})$ is Rademacher Complexity; $\hat{\mathcal{R}}_m(\mathcal{F})$ is empirical Rademacher Complexity

Story so far …

- **Two algorithms:** Sample Average Approx., Sample Approx.
- One-sided **uniform convergence** of mean is sufficient for SAA consistency.
- Defined **Rademacher Complexity.**
- Pending:
  - *Concentration around mean for the max. term.*
  - $\{\mathcal{R}_{\mathrm{m}}(\mathcal{G})\}_{m=1}^{\infty} \rightarrow \mathbf{0} \;\; \Rightarrow$ *a Learnable problem*.

Closer look at $\mathcal{R}_m(\mathcal{F}) = E\left[\max_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(Z_i)\right)\right]$

- High if $\mathcal{F}$ correlates with random noise
  - *Classification problems: $\mathcal{F}$ can assign arbitrary labels*
- Higher $\mathcal{R}_m(\mathcal{F})$, lower confidence on prediction

Closer look at $\mathcal{R}_m(\mathcal{F}) = E\left[\max_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(Z_i)\right)\right]$

- High if $\mathcal{F}$ correlates with random noise
  - *Classification problems: $\mathcal{F}$ can assign arbitrary labels*
- Higher $\mathcal{R}_m(\mathcal{F})$, lower confidence on prediction

- $\mathcal{F}_1 \subseteq \mathcal{F}_2 \Rightarrow \mathcal{R}_m(\mathcal{F}_1) \leq \mathcal{R}_m(\mathcal{F}_2)$
- Lower $\mathcal{R}_m(\mathcal{F})$, higher chance we miss Bayes optimal

Closer look at $\mathcal{R}_m(\mathcal{F}) = E\left[\max_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(Z_i)\right)\right]$

- High if $\mathcal{F}$ correlates with random noise
  - *Classification problems: $\mathcal{F}$ can assign arbitrary labels*
- Higher $\mathcal{R}_m(\mathcal{F})$, lower confidence on prediction

Choose model with right trade-off using Domain knowledge.

- $\mathcal{F}_1 \subseteq \mathcal{F}_2 \Rightarrow \mathcal{R}_m(\mathcal{F}_1) \leq \mathcal{R}_m(\mathcal{F}_2)$
- Lower $\mathcal{R}_m(\mathcal{F})$, higher chance we miss Bayes optimal

Relation with classical measures

- **Growth Function:** $\Pi_m(\mathcal{F}) \equiv \max_{\{x_1,\ldots,x_m\}\subset\mathcal{X}} |\{(f(x_1),\ldots,f(x_m)) \mid f \in \mathcal{F}\}|$
  - *Classification case:* $\Pi_m(\mathcal{F})$ *is max. no. of distinct classifiers induced*
  - ***Massart's Lemma:*** $\boldsymbol{\mathcal{R}_m(\mathcal{F}) \leq \sqrt{\frac{2\Pi_m(\mathcal{F})}{m}}}$

- **VC-Dimension:** $VCdim(\mathcal{F}) \equiv \max_{m:\Pi_m(\mathcal{F})=2^m} m$
  - ***Sauer's Lemma:*** $\boldsymbol{\mathcal{R}_m(\mathcal{F}) \leq \sqrt{\frac{2d \log\frac{em}{d}}{m}}}$

Mean concentration: Observation

- Define $h\big((X_1, Y_1), \dots, (X_m, Y_m)\big) \equiv \max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g]$
- h is function:
  - *of iid random variables*
  - *Satisfies bounded difference property*
    - Δh when one (X_i, Y_i) changes $\leq \frac{\Delta l}{m}$ (∵ bounded loss)
  - *Concentration around mean – McDiarmid's inequality*

McDiarmid's Inequality

Let $X_1, \ldots, X_m \in \mathcal{X}^m$ be iid rvs and $h: \mathcal{X}^m \mapsto \mathbb{R}$ satisfying:

$$|h(x_1, \ldots, x_i, \ldots, x_m) - h(x_1, \ldots, x_i', \ldots, x_m)| \le c_i$$

Then the following hold for any $\epsilon > 0$:

$$P[h - E[h] \ge \epsilon] \le e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}},$$

$$P[h - E[h] \le -\epsilon] \le e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

McDiarmid's Inequality

Let $X_1, \dots, X_m \in \mathcal{X}^m$ be iid rvs and $h: \mathcal{X}^m \mapsto \mathbb{R}$ satisfying:

$$|h(x_1, \dots, x_i, \dots, x_m) - h(x_1, \dots, x_i', \dots, x_m)| \le c_i$$

Then the following hold for any $\epsilon > 0$:

0 ⇒learnable

$$e^{\frac{-2m\epsilon^2}{\Delta l^2}} \to 0$$

$$P[h - E[h] \ge \epsilon] \le e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}},$$

$$P[h - E[h] \le -\epsilon] \le e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

Learning Bounds

- Let $\delta \equiv e^{\frac{-2m\epsilon^2}{\Delta l^2}}$, i.e., $\boldsymbol{\epsilon = \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$
- $P[h - E[h] \geq \epsilon] \leq \delta$ is same as:
  - *with probability atleast $1 - \delta$, we have:*

$$\boldsymbol{R[g] \leq \hat{R}_m[g] + 2\mathcal{R}_m(\mathcal{F}) + \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ \forall\, g \in \mathcal{G}}$$

Learning Bounds

- Let $\delta \equiv e^{\frac{-2m\epsilon^2}{\Delta l^2}}$, i.e., $\boldsymbol{\epsilon = \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$
- $P[h - E[h] \geq \epsilon] \leq \delta$ is same as:
  - *with probability atleast* $1 - \delta$*, we have:*

$$\boldsymbol{R[g] \leq \hat{R}_m[g] + 2\mathcal{R}_m(\mathcal{F}) + \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\ \forall\ g \in \mathcal{G}}$$

Computable
except this term!

Learning Bounds

- Let $\delta \equiv e^{\frac{-2m\epsilon^2}{\Delta l^2}}$, i.e., $\boldsymbol{\epsilon = \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$
- $P[h - E[h] \geq \epsilon] \leq \delta$ is same as:
  - *with probability atleast $1 - \delta$, we have:*

$$\boldsymbol{R[g] \leq \hat{R}_m[g] + 2\mathcal{R}_m(\mathcal{F}) + \Delta l \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \;\forall\, g \in \mathcal{G}}$$

Use McDiarmid on $\hat{\mathcal{R}}_m(\mathcal{F})$

Learning Bounds

- Let $\delta \equiv e^{\frac{-2m\epsilon^2}{\Delta l^2}}$, i.e., $\boldsymbol{\epsilon = \Delta l \sqrt{\frac{\log \frac{1}{\delta}}{2m}}}$
- $P[h - E[h] \geq \epsilon] \leq \delta$ is same as:
  - *with probability atleast $1 - \delta$, we have:*

$$\boldsymbol{R[g] \leq \hat{R}_m[g] + 2\mathcal{R}_m(\mathcal{F}) + \Delta l \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \;\forall\; g \in \mathcal{G}}$$

- With probability atleast $1 - \delta$, we have:

$$\boldsymbol{R[g] \leq \hat{R}_m[g] + 2\hat{\mathcal{R}}_m(\mathcal{F}) + 3\Delta l \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \;\forall\; g \in \mathcal{G}}$$

Story so far ...

- **Two algorithms:** Sample Average Approx., Sample Approx.
- One-sided **uniform convergence** of mean is sufficient for SAA consistency.
- Defined **Rademacher Complexity.**
- Concentration around mean for the max. term.
- $\{\mathcal{R}_{\mathbf{m}}(\mathcal{G})\}_{m=1}^{\infty} \rightarrow \mathbf{0} \;\; \Rightarrow$ **a Learnable problem.**
- Examples of ***usable*** Learnable problems
  - *Shows sufficiency condition not loose*

Linear model with Lipschitz loss

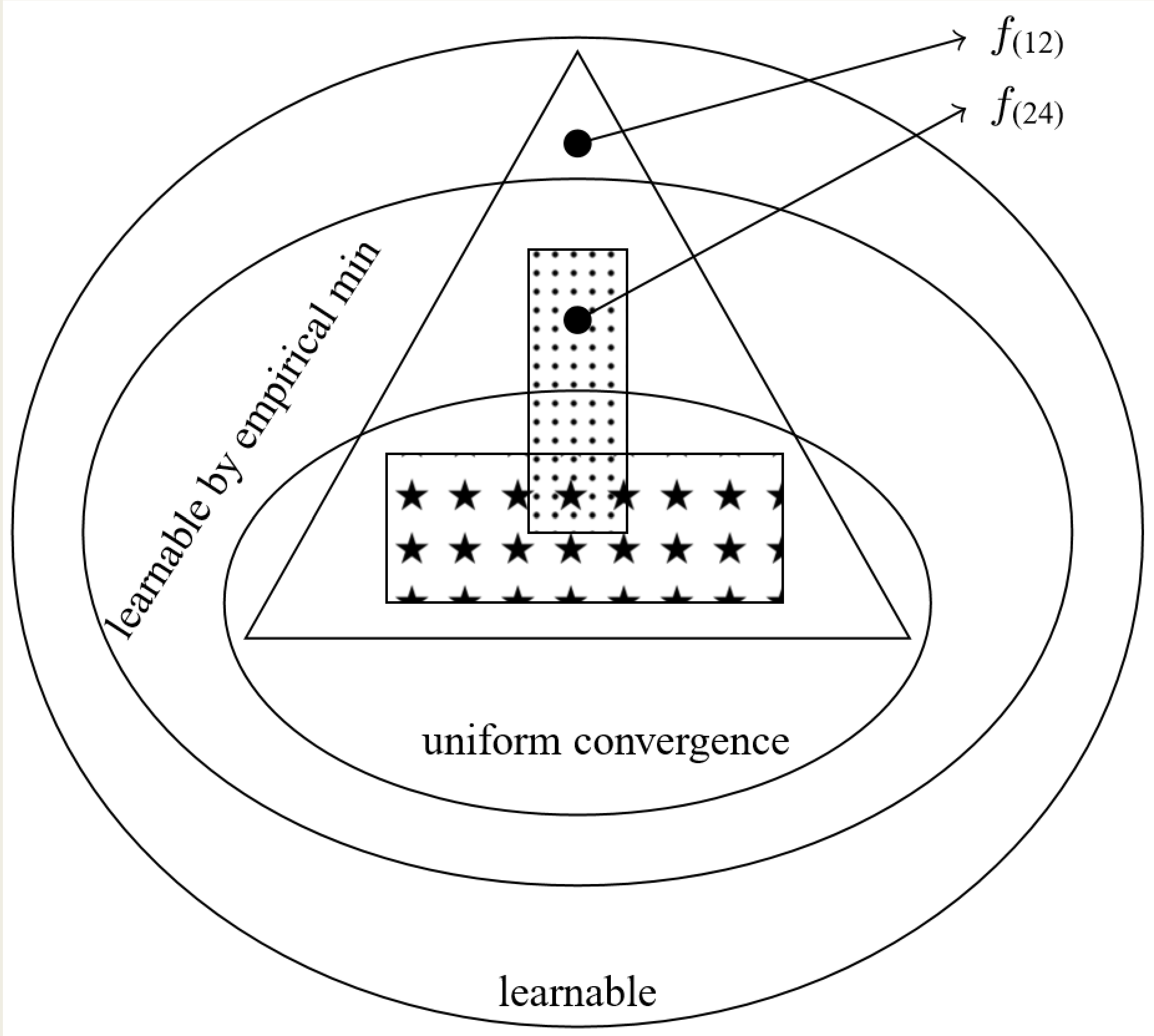
- Consider $\mathcal{G} \equiv \{g \mid \exists\, w \ni g(x) = \langle w, \phi(x) \rangle, \|w\| \leq W\},\ \phi: \mathcal{X} \mapsto \mathcal{H}$ (linear model)
- Contraction Lemma: $\hat{\mathcal{R}}_m(\mathcal{F}) \leq \hat{\mathcal{R}}_m(\mathcal{G})$

Linear model with Lipschitz loss

- Consider $\mathcal{G} \equiv \{g \mid \exists\, w \ni g(x) = \langle w, \phi(x) \rangle, \|w\| \le W\},\ \phi: \mathcal{X} \mapsto \mathcal{H}$ (linear model)
- Contraction Lemma: $\hat{\mathcal{R}}_m(\mathcal{F}) \le \hat{\mathcal{R}}_m(\mathcal{G})$
- $\hat{\mathcal{R}}_m(\mathcal{G}) = E_\sigma \left[\max_{\|w\| \le W} \frac{1}{m} \sum_{i=1}^m \sigma_i \langle w, \phi(x_i) \rangle \right]$
  - $= E_\sigma \left[\max_{\|w\| \le W} \left\langle w, \frac{1}{m} \sum_{i=1}^m \sigma_i\, \phi(x_i) \right\rangle \right]$
  - $= \frac{W}{m} E_\sigma \left[\left\| \frac{1}{m} \sum_{i=1}^m \sigma_i\, \phi(x_i) \right\| \right]$
  - $\le \frac{W}{m} \sqrt{E_\sigma \left[\left\| \frac{1}{m} \sum_{i=1}^m \sigma_i\, \phi(x_i) \right\|^2 \right]}$ (∵ Jensen's Inequality)
  - $= \frac{W}{m} \sqrt{\sum_{i=1}^m \|\phi(x_i)\|^2} \le \frac{WR}{\sqrt{m}} \to 0$ (if $\|\phi(x)\| \le R$)

Learnable Problems

Shai Shalev-Shwartz *et.al.*, 2009



THANK YOU