Credal Learning Theory

Michele Caprio
Joint work with Maryam Sultana, Eleni G. Elia, and Fabio Cuzzolin

Department of Computer Science, University of Manchester
Manchester Centre for AI Fundamentals

- Statistical Learning Theory (SLT) is the foundation of Machine Learning

- Statistical Learning Theory (SLT) is the foundation of Machine Learning
- Provides theoretical bounds for the risk of models learnt from a (single) training set

- Statistical Learning Theory (SLT) is the foundation of Machine Learning
- Provides theoretical bounds for the risk of models learnt from a (single) training set
    - Assumed to issue from a single unknown probability distribution

- Problem: predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, using mapping $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$

- Problem: predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, using mapping $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$
- Loss Function: $l : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$

- Problem: predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, using mapping $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$
- Loss Function: $l : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$
  - It measures the error committed by a model $h \in \mathcal{H}$

- Problem: predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, using mapping $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$
- Loss Function: $l : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$
  - It measures the error committed by a model $h \in \mathcal{H}$
  - Zero-one loss is defined as $l((x, y), h) \doteq \mathbb{I}[y \neq h(x)]$

- Problem: predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, using mapping $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$
- Loss Function: $l : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$
  - It measures the error committed by a model $h \in \mathcal{H}$
  - Zero-one loss is defined as $l((x, y), h) \doteq \mathbb{I}[y \neq h(x)]$
- Input-output pairs are usually assumed to be generated i.i.d. by a probability distribution $P^\star$, which is unknown

- Expected risk – or expected loss – of the model h,

$$\begin{aligned} L(h) \equiv L_{P^\star}(h) &\doteq \mathbb{E}_{P^\star}[l((x,y),h)] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} l((x,y),h)P^\star(\mathrm{d}(x,y)), \end{aligned}$$

measures the expected value w.r.t. $P^\star$ of loss l

- Expected risk – or expected loss – of the model h,

$$L(h) \equiv L_{P^\star}(h) \doteq \mathbb{E}_{P^\star}[l((x,y),h)] \\ = \int_{\mathcal{X}\times\mathcal{Y}} l((x,y),h) P^\star(\mathrm{d}(x,y)),$$

  measures the expected value w.r.t. $P^\star$ of loss l

- Expected risk minimizer

$$h^\star \in \arg\min_{h\in\mathcal{H}} L(h),$$

  is any hypothesis in $\mathcal{H}$ that minimizes the expected risk

- Consider a training dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - $(x_1, y_1), \ldots, (x_n, y_n) \sim P^\star$ i.i.d.

- Consider a training dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - $(x_1, y_1), \ldots, (x_n, y_n) \sim P^\star$ i.i.d.
- Empirical risk of hypothesis h

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), h)$$

- Consider a training dataset $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$
  - $(x_1, y_1), \dots, (x_n, y_n) \sim P^\star$ i.i.d.
- Empirical risk of hypothesis h

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), h)$$

- Empirical risk minimizer (ERM)

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$$

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
    - $\mathcal{H}$ is finite

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
    - $\mathcal{H}$ is finite
    - There exists a model $h^\star$ with zero expected risk (realizability)

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
    - $\mathcal{H}$ is finite
    - There exists a model $h^\star$ with zero expected risk (realizability)
- But what should we do when distribution shifts are allowed?

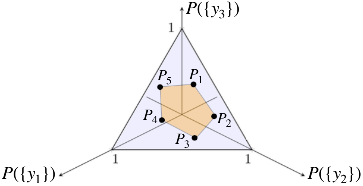
- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
    - $\mathcal{H}$ is finite
    - There exists a model $h^\star$ with zero expected risk (realizability)
- But what should we do when distribution shifts are allowed?
  - May cause issues of domain adaptation or domain generalization

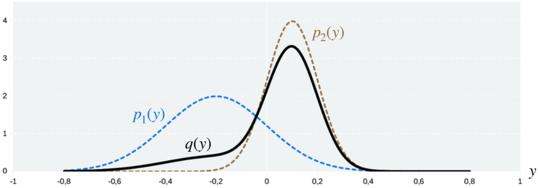
- SLT seeks upper bounds on the excess risk
    - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
    - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
        - $\mathcal{H}$ is finite
        - There exists a model $h^\star$ with zero expected risk (realizability)
- But what should we do when distribution shifts are allowed?
    - May cause issues of domain adaptation or domain generalization
    - Existing attempts: lack of generalizability, use of strong assumptions (Caprio et al., 2024, Section 2)

- SLT seeks upper bounds on the excess risk
  - Difference between the expected risk of the ERM $L(\hat{h})$, and the lowest expected risk $L(h^\star)$
  - Under increasingly more relaxed assumptions about the nature of the hypotheses space $\mathcal{H}$
    - $\mathcal{H}$ is finite
    - There exists a model $h^\star$ with zero expected risk (realizability)
- But what should we do when distribution shifts are allowed?
  - May cause issues of domain adaptation or domain generalization
  - Existing attempts: lack of generalizability, use of strong assumptions (Caprio et al., 2024, Section 2)
  - We use Credal Sets to address this issue

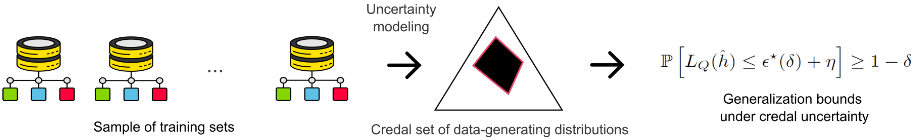
Credal Sets

- **Credal Set** Levi (1980): A set of probabilities $\mathcal{P}$ that is closed and convex
- **Finitely Generated Credal Set**: A credal set $\mathcal{P}$ with finitely many extreme elements $\text{ex}\mathcal{P}$



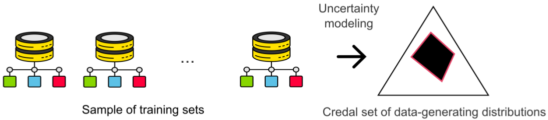


A Summary of our Learning Framework



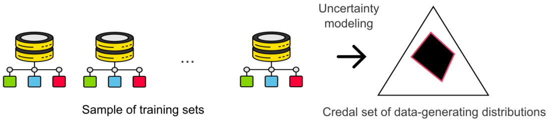
Deriving a Credal Sets

Available Evidence



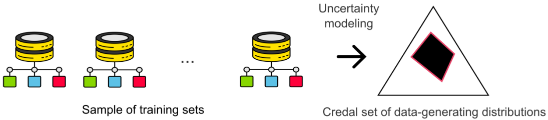
- Suppose that our evidence is a finite sample of training sets, $D_1, \ldots, D_N$

Deriving a Credal Sets
Available Evidence



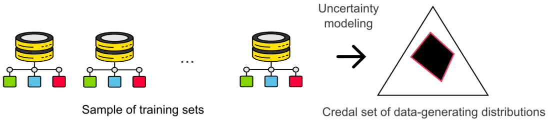
- Suppose that our evidence is a finite sample of training sets, $D_1, \ldots, D_N$
- $D_i = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,n_i}, y_{i,n_i})\}$, for all $i \in \{1, \ldots, N\}$

Deriving a Credal Sets
Available Evidence



- Suppose that our evidence is a finite sample of training sets, $D_1, \dots, D_N$
- $D_i = \{(x_{i,1}, y_{i,1}), \dots, (x_{i,n_i}, y_{i,n_i})\}$, for all $i \in \{1, \dots, N\}$
- $(x_{i,1}, y_{i,1}), \dots, (x_{i,n_i}, y_{i,n_i}) \sim P_i^\star$ i.i.d., for all $i \in \{1, \dots, N\}$

Deriving a Credal Sets
Available Evidence



- Suppose that our evidence is a finite sample of training sets, $D_1, \ldots, D_N$
- $D_i = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,n_i}, y_{i,n_i})\}$, for all $i \in \{1, \ldots, N\}$
- $(x_{i,1}, y_{i,1}), \ldots, (x_{i,n_i}, y_{i,n_i}) \sim P_i^\star$ i.i.d., for all $i \in \{1, \ldots, N\}$
- $P_i^\star$ need not be equal to $P_j^\star$, for all $i, j \in \{1, \ldots, N\}$, $i \neq j$

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Theorem 4.1)

Let $(x_{N+1,1}, y_{N+1,1}), \dots, (x_{N+1,n_{N+1}}, y_{N+1,n_{N+1}}) \equiv (x_1, y_1), \dots, (x_n, y_n)$ be sampled i.i.d. from $P^\star_{N+1} \equiv P \in \mathcal{P}$. Recall that the empirical risk minimizer is $\hat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l((x_i, y_i), h)$. Assume

- there exists a realizable hypothesis, i.e. $h^\star \in \mathcal{H}$ such that $L_P(h^\star) = 0$
- $\mathcal{H}$ is finite
- zero-one loss $l((x, y), h) = \mathbb{I}[y \neq h(x)]$

(C. et al, 2024, Theorem 4.1)

Let $(x_{N+1,1}, y_{N+1,1}), \dots, (x_{N+1,n_{N+1}}, y_{N+1,n_{N+1}}) \equiv (x_1, y_1), \dots, (x_n, y_n)$ be sampled i.i.d. from $P^\star_{N+1} \equiv P \in \mathcal{P}$. Recall that the empirical risk minimizer is $\hat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), h)$. Assume

- there exists a realizable hypothesis, i.e. $h^\star \in \mathcal{H}$ such that $L_P(h^\star) = 0$
- $\mathcal{H}$ is finite
- zero-one loss $l((x, y), h) = \mathbb{I}[y \neq h(x)]$

Fix any $\delta \in (0, 1)$. Then,

$$\mathbb{P}\left[L_P(\hat{h}) \leq \epsilon^\star(\delta)\right] \geq 1 - \delta,$$

where $\epsilon^\star(\delta)$ is a well-defined quantity that depends only on δ and on the elements of ex$\mathcal{P}$.

(C. et al, 2024, Corollary 4.3)

Retain the assumptions of Theorem 4.1. We have that

$$\epsilon^\star(\delta) \leq \epsilon_{\text{UB}}(\delta) \doteq \frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{n}.$$

In turn, the following holds for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) \leq \epsilon_{\text{UB}}(\delta)\right] \geq 1-\delta. \tag{1}$$

- $\epsilon_{\text{UB}}(\delta)$ is a uniform bound

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.3)

Retain the assumptions of Theorem 4.1. We have that

$$\epsilon^\star(\delta) \leq \epsilon_{\text{UB}}(\delta) \doteq \frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{n}.$$

In turn, the following holds for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) \leq \epsilon_{\text{UB}}(\delta)\right] \geq 1-\delta. \tag{1}$$

- $\epsilon_{\text{UB}}(\delta)$ is a uniform bound
- When only few samples are available, $\epsilon^\star(\delta)$ much smaller than $\epsilon_{\text{UB}}(\delta)$

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.3)

Retain the assumptions of Theorem 4.1. We have that

$$\epsilon^\star(\delta) \leq \epsilon_{\text{UB}}(\delta) \doteq \frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{n}.$$

In turn, the following holds for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) \leq \epsilon_{\text{UB}}(\delta)\right] \geq 1 - \delta. \tag{1}$$

- $\epsilon_{\text{UB}}(\delta)$ is a uniform bound
- When only few samples are available, $\epsilon^\star(\delta)$ much smaller than $\epsilon_{\text{UB}}(\delta)$
- $\mathcal{O}\left(\frac{\log|\cup_{P^{\text{ex}}\in\text{ex}\mathcal{P}} B_{P^{\text{ex}}}|}{n}\right) \leq \mathcal{O}\left(\frac{\log|\mathcal{H}|}{n}\right)$

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.3)

Retain the assumptions of Theorem 4.1. We have that

$$\epsilon^\star(\delta) \leq \epsilon_{\text{UB}}(\delta) \doteq \frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{n}.$$

In turn, the following holds for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) \leq \epsilon_{\text{UB}}(\delta)\right] \geq 1-\delta. \tag{1}$$

- $\epsilon_{\text{UB}}(\delta)$ is a uniform bound
- When only few samples are available, $\epsilon^\star(\delta)$ much smaller than $\epsilon_{\text{UB}}(\delta)$
- $\mathcal{O}\left(\frac{\log|\cup_{P^{\text{ex}}\in\text{ex}\mathcal{P}} B_{P^{\text{ex}}}|}{n}\right) \leq \mathcal{O}\left(\frac{\log|\mathcal{H}|}{n}\right)$
- Equation (1) corresponds to (Liang, 2016, Theorem 4)

- Allowing for distribution drift

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.4)

Consider a natural number $k < n$. Let $(x_1, y_1), \dots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \dots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.1.

Obtaining the Generalization Bounds

Realizability + Finite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.4)

Consider a natural number $k < n$. Let $(x_1, y_1), \dots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \dots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.1. Then,

$$\mathbb{P}\left[L_{P_1}(\hat{h}_1) + L_{P_2}(\hat{h}_2) \leq \epsilon^\star(\delta)\frac{n^2}{k(n-k)}\right] \geq 1-\delta,$$

where $\epsilon^\star(\delta)$ is the same quantity as in Theorem 4.1, and

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.4)

Consider a natural number $k < n$. Let $(x_1, y_1), \ldots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \ldots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.1. Then,

$$\mathbb{P}\left[L_{P_1}(\hat{h}_1) + L_{P_2}(\hat{h}_2) \leq \epsilon^\star(\delta)\frac{n^2}{k(n-k)}\right] \geq 1 - \delta,$$

where $\epsilon^\star(\delta)$ is the same quantity as in Theorem 4.1, and

$$\hat{h}_1 \in \arg\min_{h \in \mathcal{H}} \frac{1}{k}\sum_{i=1}^{k} l((x_i, y_i), h), \quad \hat{h}_2 \in \arg\min_{h \in \mathcal{H}} \frac{1}{n-k}\sum_{i=k+1}^{n} l((x_i, y_i), h).$$

Obtaining the Generalization Bounds

Further Results

- In (Caprio et al., 2024, Section 4.2): similar results when the realizability assumption is relaxed, but $\mathcal{H}$ is kept finite

Obtaining the Generalization Bounds

Further Results

- In (Caprio et al., 2024, Section 4.2): similar results when the realizability assumption is relaxed, but $\mathcal{H}$ is kept finite
- In (Caprio et al., 2024, Section 4.3): similar results when the realizability assumption is relaxed, and $\mathcal{H}$ is (potentially uncountably) infinite

Future Work

- In the future, we plan to

Future Work

- In the future, we plan to
  - Extend our results to different losses

Future Work

- In the future, we plan to
  - Extend our results to different losses
  - Derive PAC-like guarantees on the correct distribution P being an element of the credal set $\mathcal{P}$

Future Work

- In the future, we plan to
  - Extend our results to different losses
  - Derive PAC-like guarantees on the correct distribution P being an element of the credal set $\mathcal{P}$
  - Validate our findings on real datasets

Carthago Delenda Est

THANK YOU FOR YOUR ATTENTION!

References I

Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. To be submitted to NeurIPS 2024, 2024.

Isaac Levi. *The Enterprise of Knowledge*. London, UK : MIT Press, 1980.

Percy Liang. Statistical learning theory. Lecture notes for the course CS229T/STAT231 of Stanford University, 2016.

- Foregoing the Realizability assumption

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Theorem 4.5)

Let $(x_1, y_1), \dots, (x_n, y_n) \sim P$ i.i.d., where P is any element of credal set $\mathcal{P}$. Assume

- $\mathcal{H}$ is finite
- zero-one loss $l((x, y), h) = \mathbb{I}[y \neq h(x)]$

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Theorem 4.5)

Let $(x_1, y_1), \dots, (x_n, y_n) \sim P$ i.i.d., where P is any element of credal set $\mathcal{P}$. Assume

- $\mathcal{H}$ is finite
- zero-one loss $l((x, y), h) = \mathbb{I}[y \neq h(x)]$

Let $\hat{h}$ be the empirical risk minimizer, and $h^\star$ be the best theoretical model. Fix any $\delta \in (0, 1)$. Then,

$$\mathbb{P}\left[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star}(\delta)\right] \geq 1 - \delta,$$

where $\epsilon^{\star\star}(\delta)$ is a well-defined quantity that depends only on δ and on the elements of $\text{ex}\mathcal{P}$.

Go to Corollary 4.7

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

Proof. The proof builds on that of Liang (2016, Theorem 7). Fix any $\epsilon > 0$, and any $P \in \mathcal{P}$. Assume that the training dataset is given by n i.i.d. draws from P. By Liang (2016, Equations (158) and (186)), we have that

$$\begin{aligned} &\mathbb{P}[L_P(\hat{h}) - L_P(h^\star) > \epsilon] \\ &\qquad \leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} \left|\hat{L}_P(h) - L_P(h)\right| > \frac{\epsilon}{2}\right] \\ &\qquad < |\mathcal{H}| \cdot 2\exp\left(-2n\left(\frac{\epsilon}{2}\right)^2\right) \doteq \delta(\epsilon). \end{aligned} \tag{9}$$

Notice though, that we can improve on this bound, since we know that $P \in \mathcal{P}$, a finitely generated credal set. Let $B'_P \doteq \{h \in \mathcal{H} : |\hat{L}_P(h) - L_P(h)| > \epsilon/2\}$ be the set of "bad hypotheses" according to P. Then, it is immediate to see that

$$\sup_{h \in \mathcal{H}} \left|\hat{L}_P(h) - L_P(h)\right| = \sup_{h \in B'_P} \left|\hat{L}_P(h) - L_P(h)\right|.$$

Notice though that we do not know P; we only know it belongs to $\mathcal{P}$. Hence, we need to consider the set $B'_\mathcal{P}$ of bad hypotheses according to all the elements of $\mathcal{P}$, that is, $B'_\mathcal{P} \doteq \{h \in \mathcal{H} : \exists P \in \mathcal{P}, |\hat{L}_P(h) - L_P(h)| > \epsilon/2\} = \cup_{P \in \mathcal{P}} B'_P$. Since $\mathcal{P}$ is a finitely generated credal set, by the Bauer Maximum Principle and the linearity of the expectation operator we have that $B'_\mathcal{P} = B'_{\text{ex}\mathcal{P}} \doteq \{h \in \mathcal{H} : \exists P^{\text{ex}} \in \text{ex}\mathcal{P}, |\hat{L}_P(h) - L_P(h)| > \epsilon/2\} = \cup_{P^{\text{ex}} \in \text{ex}\mathcal{P}} B'_{P^{\text{ex}}}$. Hence, we obtain

$$\sup_{h \in \mathcal{H}} \left|\hat{L}_P(h) - L_P(h)\right| = \sup_{h \in B'_{\text{ex}\mathcal{P}}} \left|\hat{L}_P(h) - L_P(h)\right|.$$

In turn, (9) implies that

$$\begin{aligned} &\mathbb{P}[L_P(\hat{h}) - L_P(h^\star) > \epsilon] \\ &\qquad \leq \mathbb{P}\left[\sup_{h \in B'_{\text{ex}\mathcal{P}}} \left|\hat{L}_P(h) - L_P(h)\right| > \frac{\epsilon}{2}\right] \\ &\qquad < |B'_{\text{ex}\mathcal{P}}| \cdot 2\exp\left(-2n\left(\frac{\epsilon}{2}\right)^2\right) \doteq \delta_{\text{ex}\mathcal{P}}. \end{aligned}$$

Rearranging, we obtain

$$\epsilon = \sqrt{\frac{2\left(\log|B'_{\text{ex}\mathcal{P}}| + \log\left(\frac{2}{\delta_{\text{ex}\mathcal{P}}}\right)\right)}{n}}, \tag{10}$$

so if δ is fixed, we can write $\epsilon \equiv \epsilon^{\star\star}(\delta)$. In turn, this implies that $\mathbb{P}[L_P(\hat{h}) - L_P(h^\star) > \epsilon^{\star\star}(\delta)] < \delta$, or equivalently, $\mathbb{P}[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star}(\delta)] \geq 1 - \delta$. $\square$

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.6)

Retain the assumptions of Theorem 4.5. Denote by $Q \in \mathcal{P}$, $Q \neq P$, a generic distribution in $\mathcal{P}$ different from P. Pick any $\eta \in \mathbb{R}_{>0}$; if the TV-diameter $\text{diam}_{TV}(\mathcal{P}) = \eta$, we have that

$$\mathbb{P}\left[L_Q(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star}(\delta) + \eta\right] \geq 1 - \delta,$$

where $\epsilon^{\star\star}(\delta)$ is the same quantity as in Theorem 4.5.

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.6)

Retain the assumptions of Theorem 4.5. Denote by $Q \in \mathcal{P}$, $Q \neq P$, a generic distribution in $\mathcal{P}$ different from P. Pick any $\eta \in \mathbb{R}_{>0}$; if the TV-diameter $\text{diam}_{TV}(\mathcal{P}) = \eta$, we have that

$$\mathbb{P}\left[L_Q(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star}(\delta) + \eta\right] \geq 1 - \delta,$$

where $\epsilon^{\star\star}(\delta)$ is the same quantity as in Theorem 4.5.

- Probabilistic bound for the expected risk $L_Q(\hat{h})$ of the ERM $\hat{h}$, calculated w.r.t. the wrong distribution Q

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.6)

Retain the assumptions of Theorem 4.5. Denote by $Q \in \mathcal{P}$, $Q \neq P$, a generic distribution in $\mathcal{P}$ different from P. Pick any $\eta \in \mathbb{R}_{>0}$; if the TV-diameter $\text{diam}_{TV}(\mathcal{P}) = \eta$, we have that

$$\mathbb{P}\left[L_Q(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star}(\delta) + \eta\right] \geq 1 - \delta,$$

where $\epsilon^{\star\star}(\delta)$ is the same quantity as in Theorem 4.5.

- Probabilistic bound for the expected risk $L_Q(\hat{h})$ of the ERM $\hat{h}$, calculated w.r.t. the wrong distribution Q
  - Any distribution in $\mathcal{P}$ different from the one generating the training set

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.7)

Retain the assumptions of Theorem 4.5. Then,

$$\epsilon^{\star\star}(\delta) \leq \epsilon'_{\text{UB}}(\delta) \doteq \sqrt{\frac{2\left(\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)\right)}{n}}.$$

In turn, for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon'_{\text{UB}}(\delta)\right] \geq 1 - \delta, \tag{2}$$

- Main difference with Theorem 4.1: in Theorem 4.5, $L_P(\hat{h}) - L_P(h^\star)$ behaves as $\mathcal{O}\left(\sqrt{\frac{\log\left|\cup_{P^{\text{ex}}\in\text{ex}\mathcal{P}} B'_{P^{\text{ex}}}\right|}{n}}\right)$

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.7)

Retain the assumptions of Theorem 4.5. Then,

$$\epsilon^{\star\star}(\delta) \leq \epsilon'_{\text{UB}}(\delta) \doteq \sqrt{\frac{2\left(\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)\right)}{n}}.$$

In turn, for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon'_{\text{UB}}(\delta)\right] \geq 1 - \delta, \tag{2}$$

- Main difference with Theorem 4.1: in Theorem 4.5, $L_P(\hat{h}) - L_P(h^\star)$ behaves as $\mathcal{O}\left(\sqrt{\frac{\log|\cup_{P^{\text{ex}}\in \text{ex}\mathcal{P}} B'_{P^{\text{ex}}}|}{n}}\right)$
  - Slower than what we had in Theorem 4.1: relaxation of the realizability

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

(C. et al, 2024, Corollary 4.7)

Retain the assumptions of Theorem 4.5. Then,

$$\epsilon^{\star\star}(\delta) \leq \epsilon'_{\text{UB}}(\delta) \doteq \sqrt{\frac{2\left(\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)\right)}{n}}.$$

In turn, for all $P \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbb{P}\left[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon'_{\text{UB}}(\delta)\right] \geq 1 - \delta, \tag{2}$$

- Main difference with Theorem 4.1: in Theorem 4.5, $L_P(\hat{h}) - L_P(h^\star)$ behaves as $\mathcal{O}\left(\sqrt{\frac{\log|\cup_{P^{\text{ex}}\in\text{ex}\mathcal{P}} B'_{P^{\text{ex}}}|}{n}}\right)$
  - Slower than what we had in Theorem 4.1: relaxation of the realizability
- Equation (2) corresponds to (Liang, 2016, Theorem 7)

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

- Allowing for distribution drift

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.8)

Consider a natural number $k < n$. Let $(x_1, y_1), \ldots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \ldots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.5.

Obtaining the Generalization Bounds

No Realizability + Finite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.8)

Consider a natural number $k < n$. Let $(x_1, y_1), \ldots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \ldots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.5. Then,

$$\mathbb{P}\Bigg[\left(L_{P_1}(\hat{h}_1) - L_{P_1}\left(h^\star_{P_1}\right)\right) + \left(L_{P_2}(\hat{h}_2) - L_{P_2}\left(h^\star_{P_2}\right)\right) \leq \epsilon^{\star\star}(\delta)\sqrt{\frac{n}{k(n-k)}}(\sqrt{k} + \sqrt{n-k})\Bigg] \geq 1 - \delta,$$

where $\epsilon^{\star\star}(\delta)$ is the same quantity as in Theorem 4.5, and $\hat{h}_1$ and $\hat{h}_2$ are defined as in Corollary 4.4.

- Foregoing also finiteness of $\mathcal{H}$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Assume zero-one loss, $l((x, y), h) = \mathbb{I}[y \neq h(x)]$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Assume zero-one loss, $l((x,y),h) = \mathbb{I}[y \neq h(x)]$
- $\mathcal{A} \doteq \{(x,y) \mapsto l((x,y),h) : h \in \mathcal{H}\}$
- $\sigma_1, \ldots, \sigma_n \sim \text{Unif}(\{-1,1\})$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Assume zero-one loss, $l((x,y),h) = \mathbb{I}[y \neq h(x)]$
- $\mathcal{A} \doteq \{(x,y) \mapsto l((x,y),h) : h \in \mathcal{H}\}$
- $\sigma_1, \ldots, \sigma_n \sim \text{Unif}(\{-1,1\})$
- $R_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \mathbb{E}_{P^{\text{ex}}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i l((x_i,y_i),h)]$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Assume zero-one loss, $l((x,y),h) = \mathbb{I}[y \neq h(x)]$
- $\mathcal{A} \doteq \{(x,y) \mapsto l((x,y),h) : h \in \mathcal{H}\}$
- $\sigma_1, \ldots, \sigma_n \sim \text{Unif}(\{-1,1\})$
- $R_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \mathbb{E}_{P^{\text{ex}}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i l((x_i,y_i),h)]$

(C. et al, 2024, Theorem 4.9)

Let $(x_1,y_1),\ldots,(x_n,y_n) \sim P$ i.i.d., where P is any element of credal set $\mathcal{P}$. Let $\hat{h}$ be the ERM, and $h^\star$ be the best theoretical model. Fix any $\delta \in (0,1)$.

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Assume zero-one loss, $l((x,y),h) = \mathbb{I}[y \neq h(x)]$
- $\mathcal{A} \doteq \{(x,y) \mapsto l((x,y),h) : h \in \mathcal{H}\}$
- $\sigma_1, \ldots, \sigma_n \sim \text{Unif}(\{-1,1\})$
- $R_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \mathbb{E}_{P^{\text{ex}}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i l((x_i,y_i),h)]$

(C. et al, 2024, Theorem 4.9)

Let $(x_1,y_1),\ldots,(x_n,y_n) \sim P$ i.i.d., where P is any element of credal set $\mathcal{P}$. Let $\hat{h}$ be the ERM, and $h^\star$ be the best theoretical model. Fix any $\delta \in (0,1)$. Then, for all $P \in \mathcal{P}$,

$$\mathbb{P}\left[L_P(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star\star}(\delta)\right] \geq 1 - \delta,$$

where

$$\epsilon^{\star\star\star}(\delta) \doteq 4 \max_{P^{\text{ex}} \in \text{ex}\mathcal{P}} R_{n,P^{\text{ex}}}(\mathcal{A}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^{\star}$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification
- In real applications, we effectively cannot compute $R_{n,P^\star}(\mathcal{A})$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification
- In real applications, we effectively cannot compute $R_{n,P^\star}(\mathcal{A})$
- $R_{n,P^\star}(\mathcal{A})$ can be approximated via the *empirical Rademacher complexity* $\hat{R}_n(\mathcal{A})$ (Liang, 2016, Equation (219)), but

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification
- In real applications, we effectively cannot compute $R_{n,P^\star}(\mathcal{A})$
- $R_{n,P^\star}(\mathcal{A})$ can be approximated via the *empirical Rademacher complexity* $\hat{R}_n(\mathcal{A})$ (Liang, 2016, Equation (219)), but
  1. Especially in the case of low cardinality training set, i.e., if n is not "large enough": possible poor approximation of the classical bound

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification
- In real applications, we effectively cannot compute $R_{n,P^\star}(\mathcal{A})$
- $R_{n,P^\star}(\mathcal{A})$ can be approximated via the *empirical Rademacher complexity* $\hat{R}_n(\mathcal{A})$ (Liang, 2016, Equation (219)), but
  1. Especially in the case of low cardinality training set, i.e., if n is not "large enough": possible poor approximation of the classical bound
  2. The collected dataset $\{(x_i, y_i)\}_{i=1}^n$ may well be a realization of a stochastic process governed by a distribution different than $P^\star$

- Theorem 4.9 generalizes (Liang, 2016, Theorem 9), which focuses only on the "true" probability $P^\star$ on $\mathcal{X} \times \mathcal{Y}$
  - Our result holds for all the plausible distributions in credal set $\mathcal{P}$
  - Hedge against distribution misspecification
- In real applications, we effectively cannot compute $R_{n,P^\star}(\mathcal{A})$
- $R_{n,P^\star}(\mathcal{A})$ can be approximated via the *empirical Rademacher complexity* $\hat{R}_n(\mathcal{A})$ (Liang, 2016, Equation (219)), but
  1. Especially in the case of low cardinality training set, i.e., if n is not "large enough": possible poor approximation of the classical bound
  2. The collected dataset $\{(x_i, y_i)\}_{i=1}^n$ may well be a realization of a stochastic process governed by a distribution different than $P^\star$
     - Empirical Rademacher complexity $\hat{R}_n(\mathcal{A})$ is not able to distinguish between these two cases

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- $\overline{R}_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \max_{P^{\text{ex}} \in \text{ex}\mathcal{P}} R_{n,P^{\text{ex}}}(\mathcal{A})$ is more conservative (looser bound), but

- $\overline{R}_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \max_{P^{\text{ex}} \in \text{ex}\mathcal{P}} R_{n,P^{\text{ex}}}(\mathcal{A})$ is more conservative (looser bound), but
  - It can be computed explicitly – since we know credal set $\mathcal{P}$ and its extreme elements ex$\mathcal{P}$

- $\overline{R}_{n,P^{\text{ex}}}(\mathcal{A}) \doteq \max_{P^{\text{ex}} \in \text{ex}\mathcal{P}} R_{n,P^{\text{ex}}}(\mathcal{A})$ is more conservative (looser bound), but
  - It can be computed explicitly – since we know credal set $\mathcal{P}$ and its extreme elements ex$\mathcal{P}$
  - It holds for all $P \in \mathcal{P}$

Go to Corollary 4.12

(C. et al, 2024, Corollary 4.10)

Retain the assumptions of Theorem 4.9. If $\mathcal{P}$ is the singleton $\{P^\star\}$, we retrieve (Liang, 2016, Theorem 9).

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

(C. et al, 2024, Corollary 4.10)

Retain the assumptions of Theorem 4.9. If $\mathcal{P}$ is the singleton $\{P^\star\}$, we retrieve (Liang, 2016, Theorem 9).

(C. et al, 2024, Corollary 4.11)

Retain the assumptions of Theorem 4.9. Denote by $Q \in \mathcal{P}$, $Q \neq P$, a generic distribution in $\mathcal{P}$ different from P. Pick any $\eta \in \mathbb{R}_{>0}$; if $\text{diam}_{TV}(\mathcal{P}) = \eta$, we have that

$$\mathbb{P}\left[L_Q(\hat{h}) - L_P(h^\star) \leq \epsilon^{\star\star\star}(\delta) + \eta\right] \geq 1 - \delta,$$

where $\epsilon^{\star\star\star}(\delta)$ is the same quantity as in Theorem 4.9.

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Allowing for distribution drift

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.12)

Consider a natural number $k < n$. Let $(x_1, y_1), \ldots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \ldots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.9, and let

$$\epsilon^{\star\star\star}_{\text{shift}} \doteq 4\left[\overline{R}_{k,P^{\text{ex}}}(\mathcal{A}) + \overline{R}_{n-k,P^{\text{ex}}}(\mathcal{A})\right] + \sqrt{\frac{2\log(2/\delta)}{n(n-k)}}\left(\sqrt{n-k} + \sqrt{n}\right).$$

Obtaining the Generalization Bounds

No Realizability + (Possibly) Infinite $\mathcal{H}$

- Allowing for distribution drift

(C. et al, 2024, Corollary 4.12)

Consider a natural number $k < n$. Let $(x_1, y_1), \dots, (x_k, y_k) \sim P_1$ i.i.d., and $(x_{k+1}, y_{k+1}), \dots, (x_n, y_n) \sim P_2$ i.i.d., where P_1, P_2 are two generic elements of credal set $\mathcal{P}$. Retain the other assumptions of Theorem 4.9, and let

$$\epsilon^{\star\star\star}_{\text{shift}} \doteq 4 \left[\overline{R}_{k,P^{\text{ex}}}(\mathcal{A}) + \overline{R}_{n-k,P^{\text{ex}}}(\mathcal{A}) \right] + \sqrt{\frac{2\log(2/\delta)}{n(n-k)}} \left(\sqrt{n-k} + \sqrt{n} \right).$$

Then,

$$\mathbb{P}\left[\left(L_{P_1}(\hat{h}_1) - L_{P_1}\left(h^\star_{P_1}\right) \right) + \left(L_{P_2}(\hat{h}_2) - L_{P_2}\left(h^\star_{P_2}\right) \right) \leq \epsilon^{\star\star\star}_{\text{shift}} \right] \geq 1 - \delta.$$