EE Department - Sharif University of Technology

Information-Theoretic Analysis of Learning Algorithms

Behrad Moniri
bemoniri@ee.sharif.edu

February 8, 2021

Overview of Talk

Goal: Analysis of the performance of learning algorithms using tools from information theory.

Overview of Talk

Goal: Analysis of the performance of learning algorithms using tools from information theory.

1. Frequentist Setting.
2. Bayesian Setting.

Frequentist Setting

Problem Formulation

- **Instance Set**: $\mathcal{Z}$
  Hypothesis Set: $\mathcal{W}$
  Loss Function: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$

Problem Formulation

- **Instance Set**: $\mathcal{Z}$
  Hypothesis Set: $\mathcal{W}$
  Loss Function: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- An unknown distribution μ on $\mathcal{Z}$.

Problem Formulation

- **Instance Set**: $\mathcal{Z}$
  Hypothesis Set: $\mathcal{W}$
  Loss Function: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- An unknown distribution μ on $\mathcal{Z}$.
- Given $S = (Z_1, \ldots, Z_n) \sim \mu$.

Problem Formulation

- **Instance Set**: $\mathcal{Z}$
  Hypothesis Set: $\mathcal{W}$
  Loss Function: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- An unknown distribution μ on $\mathcal{Z}$.
- Given $S = (Z_1, \ldots, Z_n) \sim \mu$.
- For every $w \in \mathcal{W}$, define

$$\begin{cases} L_\mu(w) = \mathbb{E}_\mu[\ell(w, Z)] = \int \ell(w, z)\mu(dz) \\ \\ L_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i) \end{cases}$$

Problem Formulation

- **Instance Set**: $\mathcal{Z}$
  Hypothesis Set: $\mathcal{W}$
  Loss Function: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- An unknown distribution μ on $\mathcal{Z}$.
- Given $S = (Z_1, \ldots, Z_n) \sim \mu$.
- For every $w \in \mathcal{W}$, define

$$\begin{cases} L_\mu(w) = \mathbb{E}_\mu[\ell(w, Z)] = \int \ell(w, z)\mu(dz) \\ \\ L_S(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i) \end{cases}$$

- **Goal**: Algorithm picks $W \in \mathcal{W}$ according to some $P_{W|S}$. Control

$$\mathbb{E}[\text{gen}(W)] = \mathbb{E}[L_\mu(W) - L_S(W)],$$

where the expectation is over $P_{S,W} = \mu^{\otimes n} P_{W|S}$.

History

- Traditional ways of analyzing generalization:
  - VC Dimension
  - Radamacher Complexities

History

- Traditional ways of analyzing generalization:
  - VC Dimension
  - Radamacher Complexities
- **Uniform Bounds**: $\mathbb{E}[\text{gen}(W)] \leq \mathbb{E}[\sup_{w \in \mathcal{W}} \text{gen}(w)]$

History

- Traditional ways of analyzing generalization:
  - VC Dimension
  - Radamacher Complexities
- **Uniform Bounds**: $\mathbb{E}[\text{gen}(W)] \leq \mathbb{E}[\sup_{w \in \mathcal{W}} \text{gen}(w)]$
- These bounds depend only on the hypothesis set: **pessimistic**.

History

- Traditional ways of analyzing generalization:
  - VC Dimension
  - Radamacher Complexities
- **Uniform Bounds**: $\mathbb{E}[\text{gen}(W)] \leq \mathbb{E}[\sup_{w \in \mathcal{W}} \text{gen}(w)]$
- These bounds depend only on the hypothesis set: **pessimistic**.
- **Intuition**:

$$\text{less information usage from } S \implies \text{less overfitting}$$

Definition

The random variable X is called σ-subgaussian if

$$\forall \lambda \in \mathbb{R} : \quad \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\lambda^2\sigma^2/2}.$$

Definition

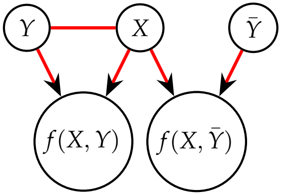
The random variable X is called σ-subgaussian if

$$\forall \lambda \in \mathbb{R}: \quad \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\lambda^2\sigma^2/2}.$$

Lemma [Xu and Raginsky, 2017] & [Russo and Zou, 2016]

Let $(X, Y) \sim P_{XY}$, and $\bar{Y} \sim P_Y$ be independent copy. For any real valued $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, if $f(X, \bar{Y})$ is σ-subgaussian, then

$$\left|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(X,\bar{Y})]\right| \leq \sqrt{2\sigma^2 I(X;Y)}$$



Decoupling Estimate

Proof: Donsker-Varadhan variational representation:

$$\mathrm{KL}(\pi||\rho) = \sup_F \left\{ \int_\Omega F d\pi - \log \int_\Omega e^F d\rho \right\}.$$

Proof: Donsker-Varadhan variational representation:

$$\mathrm{KL}(\pi||\rho) = \sup_{F} \left\{ \int_{\Omega} F d\pi - \log \int_{\Omega} e^{F} d\rho \right\}.$$

For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}\mathrm{KL}(P_{XY}||P_X \otimes P_Y) &\geq \mathbb{E}[\lambda f(X,Y)] - \log \mathbb{E}[e^{\lambda f(X,\bar{Y})}] \\ &\geq \lambda \mathbb{E}[f(X,Y)] - \lambda \mathbb{E}[f(X,\bar{Y})] - \frac{\lambda^2 \sigma^2}{2}\end{aligned}$$

Decoupling Estimate

Proof: Donsker-Varadhan variational representation:

$$\mathrm{KL}(\pi||\rho) = \sup_F \left\{ \int_\Omega F d\pi - \log \int_\Omega e^F d\rho \right\}.$$

For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} \mathrm{KL}(P_{XY}||P_X \otimes P_Y) &\geq \mathbb{E}[\lambda f(X,Y)] - \log \mathbb{E}[e^{\lambda f(X,\bar{Y})}] \\ &\geq \lambda \mathbb{E}[f(X,Y)] - \lambda \mathbb{E}[f(X,\bar{Y})] - \frac{\lambda^2 \sigma^2}{2} \end{aligned}$$

Discriminant must be non-positive, which concludes the proof. $\square$

Capacity Theorem

- Let $f(s,w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$.

Capacity Theorem

- Let $f(s,w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$. We have

$$\begin{aligned}\mathbb{E}[\text{gen}(W)] &= \mathbb{E}[L_\mu(W) - L_S(W)] \\ &= \mathbb{E}[f(\bar{S}, W)] - \mathbb{E}[f(S, W)].\end{aligned}$$

Capacity Theorem

- Let $f(s, w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$. We have

$$\begin{aligned} \mathbb{E}[\text{gen}(W)] &= \mathbb{E}[L_\mu(W) - L_S(W)] \\ &= \mathbb{E}[f(\bar{S}, W)] - \mathbb{E}[f(S, W)]. \end{aligned}$$

- If $l(w, Z)$ is σ-subgaussian $\implies f(S, w)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian.

Capacity Theorem

- Let $f(s,w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$. We have

$$\begin{aligned}\mathbb{E}[\mathrm{gen}(W)] &= \mathbb{E}[L_\mu(W) - L_S(W)] \\ &= \mathbb{E}[f(\bar{S}, W)] - \mathbb{E}[f(S, W)].\end{aligned}$$

- If $l(w, Z)$ is σ-subgaussian $\implies f(S, w)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian.

Theorem [Xu and Raginsky, 2017]

Suppose that $\ell(w, Z)$ is σ-subgaussian for μ, under all $w \in \mathcal{W}$. We have

$$|\mathbb{E}[\mathrm{gen}(W)]| \leq \sqrt{\frac{2\sigma^2}{n} I(S; W)}$$

Capacity Theorem

- Let $f(s, w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$. We have

$$\begin{aligned}\mathbb{E}[\text{gen}(W)] &= \mathbb{E}[L_\mu(W) - L_S(W)] \\ &= \mathbb{E}[f(\bar{S}, W)] - \mathbb{E}[f(S, W)].\end{aligned}$$

- If $l(w, Z)$ is σ-subgaussian $\implies f(S, w)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian.

Theorem [Xu and Raginsky, 2017]

Suppose that $\ell(w, Z)$ is σ-subgaussian for μ, under all $w \in \mathcal{W}$. We have

$$|\mathbb{E}[\text{gen}(W)]| \leq \sqrt{\frac{2\sigma^2}{n} I(S; W)}$$

Remark

- The learning algorithm $P_{W|S}$: **channel** from S to W.
- $\sup_\mu I(S; W)$: channel **capacity** of the channel, under the constraint that the input distribution is of a product form.

Chaining Mutual Information

- When W is deterministic given S: $I(S;W)$ is infinite.

Chaining Mutual Information

- When W is deterministic given S: $I(S; W)$ is infinite.
- **Can we tighten our bounds?**

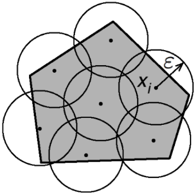
Chaining Mutual Information

- When W is deterministic given S: $I(S;W)$ is infinite.
- **Can we tighten our bounds?**
- Assume we intend to upper bound $\mathbb{E}\left[\sup_{t\in T} X_t\right]$.
- (T,d) is a metric space. $X_t - X_s \sim d^2(s,t)$-subgaussian.

Chaining Mutual Information

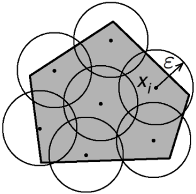
- When W is deterministic given S: $I(S;W)$ is infinite.
- **Can we tighten our bounds?**
- Assume we intend to upper bound $\mathbb{E}\left[\sup_{t\in T} X_t\right]$.
- (T,d) is a metric space. $X_t - X_s \sim d^2(s,t)$-subgaussian.

Let $\mathcal{N}_k$ be $\epsilon_k = 2^{-k}$-net of T



- When W is deterministic given S: $I(S;W)$ is infinite.
- **Can we tighten our bounds?**
- Assume we intend to upper bound $\mathbb{E}\left[\sup_{t\in T} X_t\right]$.
- (T,d) is a metric space. $X_t - X_s \sim d^2(s,t)$-subgaussian.

Let $\mathcal{N}_k$ be $\epsilon_k = 2^{-k}$-net of T



Dudley bound: multi-scale approximation of T

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 6 \sum_{k\in\mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T,d,2^{-k})}.$$

Chaining Mutual Information [Asadi, Abbe & Verdu. 2019]

If $\{\text{gen}(w)\}_{w\in\mathcal{W}}$ is a subgaussian process in $(\mathcal{W}, d)$:

$$\mathbb{E}[\text{gen}(W)] \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\pi_k(W); S)}.$$

Bayesian Setting

- **Generation Model**:

$$P_{W,S,Z} = P_W \otimes \prod_{i=1}^{n} P_{Z_i|W} \otimes P_{Z|W}$$
$$\forall\, i \in [n], \quad P_{Z_i|W} = P_{Z|W}$$

- *Predicting Modeling Framework*: $Z = (X, Y)$, and $Z_i = (X_i, Y_i)$.

$$P_{Z|W} = P_X \otimes P_{Y|X,W}$$

- **Generation Model**:

$$P_{W,S,Z} = P_W \otimes \prod_{i=1}^{n} P_{Z_i|W} \otimes P_{Z|W}$$

$$\forall\, i \in [n], \quad P_{Z_i|W} = P_{Z|W}$$

- *Predicting Modeling Framework*: $Z = (X, Y)$, and $Z_i = (X_i, Y_i)$.

$$P_{Z|W} = P_X \otimes P_{Y|X,W}$$

- **Goal**: predict Y based on X and observations $S = \{Z_1, \ldots, Z_n\}$.

- **Generalized Entropy**:

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y, \psi(X)]$$

Minimum Excess Risk

- **Generalized Entropy**:

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y, \psi(X)] \rightsquigarrow \psi^*_{Y|X}(x).$$

Minimum Excess Risk

- **Generalized Entropy**:

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y, \psi(X)] \rightsquigarrow \psi^*_{Y|X}(x).$$

- **Minimum Excess Risk (MER)**:

$$\mathrm{MER}^n_\ell = R_\ell(Y|S, X) - R_\ell(Y|W, X)$$

Minimum Excess Risk

- **Generalized Entropy**:

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y,\psi(X)] \rightsquigarrow \psi^*_{Y|X}(x).$$

- **Minimum Excess Risk (MER)**:

$$\mathrm{MER}_\ell^n = R_\ell(Y|S,X) - R_\ell(Y|W,X)$$

- MER is algorithm *independent*.

Minimum Excess Risk

- **Generalized Entropy**:

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y,\psi(X)] \rightsquigarrow \psi^*_{Y|X}(x).$$

- **Minimum Excess Risk (MER)**:

$$\mathrm{MER}_\ell^n = R_\ell(Y|S,X) - R_\ell(Y|W,X)$$

- MER is algorithm *independent*.

Theorem [Xu & Raginsky 2020], [Hafez & Moniri, 2021]

The following bound can be derived for MER:

$$\mathrm{MER}_\ell^n \leq \sqrt{\frac{b^2}{2} I(Y;W|S,X)}.$$

Minimum Excess Risk: Lower Bounds

- Lower bounds were left as an open problem in [Xu & Raginsky 2020].

- Lower bounds were left as an open problem in [Xu & Raginsky 2020].

Remark [Hafez & Moniri 2020]

It is *impossible* to find a matching lower bound such that

$$\mathrm{MER}_\ell^n \geq \alpha \sqrt{I(Y;W|S,X)}.$$

Minimum Excess Risk: Lower Bounds

- Define the following distortion function:

$$d(w, \hat{h}(.)) = \mathbb{E}_{XY|W=w}[\ell(Y, \hat{h}(X)) - \ell(Y, \psi^*_{Y|XW}(w, X))].$$

- Define the following distortion function:

$$d(w, \hat{h}(.)) = \mathbb{E}_{XY|W=w}[\ell(Y, \hat{h}(X)) - \ell(Y, \psi^*_{Y|XW}(w, X))].$$

- **Optimal algorithm**: outputs $\hat{h}(.) = \psi^*_{Y|SX}(s, .)$.

$$\begin{aligned}
&\mathbb{E}_{WS}[d(W, \psi^*_{Y|SX}(S, .))] \\
&\qquad = \mathbb{E}_{WSXY}[\ell(Y, \psi^*_{Y|SX}(S, X)) - \ell(Y, \psi^*_{Y|WX}(W, X))] \\
&\qquad = R_\ell(Y|S, X) - R_\ell(Y|W, X) = \mathrm{MER}^n_\ell.
\end{aligned}$$

Minimum Excess Risk: Lower Bounds

- Define the (constrained) rate-distortion optimization:

$$D_n(R) = \inf_{P_{\hat{h}|S}} \mathbb{E}[d(W, \hat{h})], \quad \text{s.t. } I(W; \hat{h}) \leq R.$$

- Define the (constrained) rate-distortion optimization:

$$D_n(R) = \inf_{P_{\hat{h}|S}} \mathbb{E}[d(W, \hat{h})], \quad \text{s.t.} \quad I(W; \hat{h}) \leq R.$$

- Note that $W \to S \to \hat{h}$ and $P_{S|W}$ is fixed.

Theorem

For a given training set size n, for all rates $R \geq I(W; S)$, we have

$$D_n(R) = \mathrm{MER}_\ell^n.$$

Explicit Bounds on MER

Assume that $\mathcal{W}$ is a d-dimensional subset of $\mathbb{R}^d$.

- **Upper Bounds** under some regularity conditions on $P_{Z|W}$:
  - $\text{MER}_l^n = O(\frac{1}{n})$ for bounded loss.
  - $\text{MER}_2^n = O(\frac{1}{n})$ for quadratic loss.

Explicit Bounds on MER

Assume that $\mathcal{W}$ is a d-dimensional subset of $\mathbb{R}^d$.

- **Upper Bounds** under some regularity conditions on $P_{Z|W}$:
  - $\mathrm{MER}_l^n = O(\frac{1}{n})$ for bounded loss.
  - $\mathrm{MER}_2^n = O(\frac{1}{n})$ for quadratic loss.
- **Lower Bounds** using the R/D view and the Shannon Lower Bound, in some cases, we prove $\Omega(\frac{1}{n})$ rates.

Explicit Bounds on MER

Assume that $\mathcal{W}$ is a d-dimensional subset of $\mathbb{R}^d$.

- **Upper Bounds** under some regularity conditions on $P_{Z|W}$:
  - $\text{MER}_l^n = O(\frac{1}{n})$ for bounded loss.
  - $\text{MER}_2^n = O(\frac{1}{n})$ for quadratic loss.
- **Lower Bounds** using the R/D view and the Shannon Lower Bound, in some cases, we prove $\Omega(\frac{1}{n})$ rates.
- For example, in $Y = W^\top X + \sigma\nu$ with

$$\begin{cases} W \sim \mathcal{N}(0, I_{p\times p}) \\ X \sim \mathcal{N}(0, \Sigma_X) \\ \nu \sim \mathcal{N}(0, I_{p\times p}) \end{cases}$$

  we have $\text{MER}_2^n = \Omega(\frac{p}{n})$.

Conclusion

Using information theoretic tools:

- We upper bounded generalization gap in a frequentist setting.
- We upper and lower bounded minimum excess risk in Bayesian statistics.

References

1. Aolin Xu and Maxim Raginsky.
   Minimum Excess Risk in Bayesian Learning
   Arxiv, 2020
2. Aolin Xu and Maxim Raginsky.
   Information-Theoretic Analysis of Generalization Capability of Learning Algorithms
   NeurIPS, 2017
3. Daniel Russo and James Zou.
   How much does you data exploration overfit?
   AISTAT 2016, **IEEE Trans. Info. Theory** 2020
4. Amir R. Asadi, Emmanuel Abbe, and Sergio Verdu.
   Chaining Mutual Information and Tightening Generalization Bounds
   NeurIPS 2019
5. Hafez-Kolahi, Golgooni, Kasaei, and Soleymani.
   Conditioning and Processing: Thechniques to improve information theoretic generalization bounds
   NeurIPS, 2020.
6. Bu, Zou, and Veeravalli.
   Tightening Mutual Information Based Bounds on Generalization Error
   IEEE Sel. Areas in Info. Theory, 2020.
7. Assadi & Abbe.
   Maximum Multiscale Entropy and Neural Network Regularization
   Arxiv, 2020.