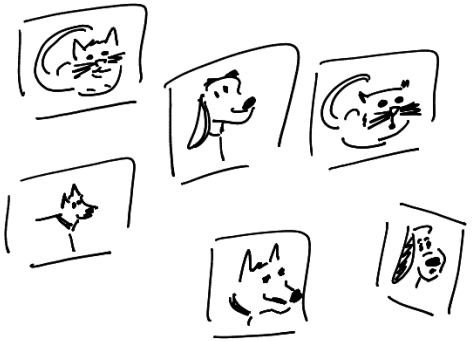
INFORMATION-THEORETIC GENERALIZATION BOUNDS

8 FEB 2021

EUGENIO CLERICO

1/ STATISTICAL LEARNING

$\mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ | $(\mathcal{Z}, P_Z)$



$(\mathcal{S} = \mathcal{Z}^n, P_Z^{\otimes n})$

$\mathcal{S}$ TRAINING SAMPLES

$x \mapsto$ [ω] $\to f_\omega(x)$ | [ω=?] $\xrightarrow[s \in \mathcal{S}]{}$ [ω_s]

TRAINING...

2/ JUDGING AN ALGORITHM

$\ell : W \times Z \to \mathbb{R}$

SQUARED LOSS: $\ell(w,(x,y)) = \frac{1}{2}\left(f_w(x) - y\right)^2$

0-1 LOSS: $\ell(w,(x,y)) = \begin{cases} 1 & \text{if } f_w(x) \neq y \\ 0 & \text{otherwise} \end{cases}$

POPULATION RISK : $L_P(w) = \mathbb{E}_z\left[\ell(w,z)\right]$

EMPIRICAL RISK: $L_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i)$

$\left(S = (z_1 \dots z_m)\right)$

GOAL: TO FIND A "GOOD" $P_{w|S}$

3/ ALGORITHM : $S \longmapsto W \sim P_{W|S=S}$

$S \quad W$

WE HAVE ACCESS TO THE EMPIRICAL RISK $L_S(w)$

GOAL: CONTROL ON THE POPULATION RISK $L_P(w)$

GENERALIZATION GAP : $L_P(w) - L_S(w)$

WE WILL FOCUS ON ITS EXPECTATION

$$\mathcal{G} = \mathbb{E}_{S,W}\left[L_P(W) - L_S(W) \right]$$

4/ Generalization Bounds

→ Various ideas:

- Complexity of $\mathcal{W}$
- Properties of the algorithm
- Dependency Output - Training Sample

Bounds based on Information Measures ⟷ Stability

"The less dependent the output hypothesis W is on the input dataset S, the better the learning algorithm generalizes" [Xu & Raginsky, 2017]

5/ Mutual Information Bound

Kullback - Leibler Divergence

$$KL(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{X\sim\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}(X)\right]$$

$KL \geq 0$

$KL = 0 \iff \mathbb{P} = \mathbb{Q}$

KL might be ∞

Variational Characterization

$$KL(\mathbb{P}\|\mathbb{Q}) = \sup_{F : e^F \in L^1(\mathbb{Q})} \left(\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \log \mathbb{E}_{X\sim\mathbb{Q}}[e^{f(X)}]\right)$$

$$\Rightarrow \mathbb{E}_{X\sim\mathbb{P}}[f(X)] \leq KL(\mathbb{P}\|\mathbb{Q}) + \log \mathbb{E}_{X\sim\mathbb{Q}}[e^{f(X)}]$$

[How much does your data exploration overfit? Controlling bias via information usage, Russo & Zou, 2016]
[Information-theoretic analysis of generalization capability of learning algorithms, Xu & Raginsky, 2017]

6/ Mutual Information

$(X, Y) \sim \mathbb{P}_{X,Y}$ $\quad I(X,Y)$ measures the "dependence" $X \longleftrightarrow Y$

$$I(X,Y) = KL(\mathbb{P}_{X,Y} \,\|\, \underbrace{\mathbb{P}_X \otimes \mathbb{P}_Y}_{\text{As if } X \perp\!\!\!\perp Y})$$

If $\bar{X} \sim X$, $\bar{Y} \sim Y$, $\bar{X} \perp\!\!\!\perp \bar{Y}$

$\Rightarrow I(X,Y) = KL(\mathbb{P}_{X,Y} \,\|\, \mathbb{P}_{\bar{X},\bar{Y}})$

Subgaussianity

X is σ^2-subgaussian

$$\mathbb{E}[e^{\lambda X}] \le e^{\lambda \mathbb{E}[X]} e^{\frac{\lambda^2 \sigma^2}{2}}$$

(i.e. $\log \mathbb{E}[e^{\lambda X}] \le \lambda \mathbb{E}[X] + \frac{\lambda^2 \sigma^2}{2}$)

Examples

- $X \sim \mathcal{N}(\mu, \sigma^2)$ is σ^2-subgauss
- $X \in [a,b]$ is $\frac{(b-a)^2}{4}$-subgauss

7/ **LEMMA 1**

(X, Y) , $(\bar{X}, \bar{Y})$: $P_{\bar{X},\bar{Y}} \sim P_X \otimes P_Y$

$\ell(\bar{X}, \bar{Y})$ σ^2-SUBGAUSS wrt $P_{\bar{X},\bar{Y}}$

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

Proof

$$\forall \lambda \quad I(X,Y) \geq \lambda \mathbb{E}[\ell(X,Y)] - \log \mathbb{E}[e^{\lambda \ell(\bar{X},\bar{Y})}]$$

↑ Variational def of KL

$$\geq \lambda \left(\mathbb{E}[\ell(X,Y)] - \mathbb{E}[\ell(\bar{X},\bar{Y})] \right) + \frac{\lambda^2 \sigma^2}{2}$$

↳ Subgauss

Taking the optimal λ $\Rightarrow$ BOUND. $\square$

8/ Back to Statistical Learning...

$$\mathcal{G} = \mathbb{E}_{S,W}\left[L_P(W) - L_S(W)\right]$$

$$L_P(\omega) = \mathbb{E}_Z\left[\ell(\omega, Z)\right] = \mathbb{E}_S\left[L_S(\omega)\right] \quad \forall \text{ fixed } \omega$$

$$\Rightarrow \mathbb{E}_{S,W}\left[L_P(W)\right] = \mathbb{E}_W\left[L_P(W)\right] = \mathbb{E}_W\left[\mathbb{E}_S\left[L_S(W)\right]\right]$$

$$\mathcal{G} = \mathbb{E}_{S\otimes W}\left[L_S(W)\right] - \mathbb{E}_{S,W}\left[L_S(W)\right]$$

Let $\bar{W} \sim W$, $\bar{S} \sim S$, $\bar{W} \perp\!\!\!\perp \bar{S}$

$$\mathcal{G} = \mathbb{E}_{\bar{S},\bar{W}}\left[L_{\bar{S}}(\bar{W})\right] - \mathbb{E}_{S,W}\left[L_S(W)\right]$$

9/ Simple Bound

Proposition 1

$\ell(w, Z)$ σ^2-subgauss $\forall w$

$\Rightarrow |\mathcal{G}| \leq \sqrt{\frac{2\sigma^2}{n} I(S;W)}$

Lemma 1

(X, Y), $(\bar{X}, \bar{Y})$: $P_{\bar{X},\bar{Y}} \sim P_X \otimes P_Y$

$\ell(\bar{X}, \bar{Y})$ σ^2-subgauss wrt $P_{\bar{X},\bar{Y}}$

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

Proof

Let $f(s, w) = L_s(w)$

$\ell(w, \cdot)$ σ^2-subgauss $\Rightarrow$ $f(S, w)$ $\frac{\sigma^2}{n}$-subgauss

$\Rightarrow$ $f(\bar{S}, \bar{W})$ $\frac{\sigma^2}{n}$-subgauss

$\Rightarrow$ Apply Lemma 1 and conclude. $\square$

10/ More General (and Tighter!) Bound

Proposition 2

$\Psi : \mathbb{R}^+ \to \mathbb{R}$, convex, smooth, $\Psi(0) = \Psi'(0) = 0$

$$\log \mathbb{E}_S\left[e^{\lambda (L_P(w) - L_S(w))} \right] \leq \Psi(\lambda) \quad \forall \lambda > 0, \ \forall w$$

$$\Rightarrow |\mathcal{G}| \leq \Psi^{*-1}\left(I(W; L_S) \right)$$

(with $\Psi^* : x \mapsto \sup_{\lambda \in \mathbb{R}^+} \{ x\lambda - \Psi(\lambda) \}$)

Remark: If $\Psi(\lambda) = \frac{\lambda^2 \sigma^2}{2n} \Rightarrow \frac{\sigma^2}{n}$-subgauss

$$|\mathcal{G}| \leq \sqrt{\frac{2\sigma^2}{n} I(W; L_S)} \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

11/ Some comments on the MI bound

- It might be loose or even ∞
  (if the algorithm is a deterministic function of the sample)
- It requires assumptions on P_Z, which is unknown
- Intuitive idea which allows many "variation"
- It takes into account the algorithm itself and not only the hypothesis space.
- It can lead to the design of new algorithms ($\neq$ERM)
- Does not take into account the dependencies between different hypotheses

12/ The Chaining Method

Subgaussian Process

(T, d) Metric space

$\{X_t\}_{t \in T}$ is subgaussian if $\forall \lambda \geq 0, \forall s, t \in T$

$$\mathbb{E}\left[e^{\lambda (X_s - X_t)}\right] \leq e^{\frac{\lambda^2}{2} d(s,t)^2} ; \quad \mathbb{E}[X_t] = 0$$

Separable Process

$\exists S \subset T$ countable s.t.

$$X_t \in \lim_{\substack{s \to t \\ s \in S}} X_s \quad \text{a.s.}$$

[Probability in High Dimensions, Van Handel, 2016]

13/ MAXIMAL INEQUALITY

$(X_1, \dots, X_n)$ FINITE COLLECTION OF RVs

X_i σ^2-SUBGAUSS $\forall i = 1 \dots n$

$$\Rightarrow \mathbb{E}\left[\max_{i=1\dots n} X_i\right] \leq \sqrt{2\sigma^2 \log n}$$

GOAL: FIND AN UPPER BOUND FOR $\mathbb{E}\left[\sup_{t \in T} X_t\right]$ FOR A GENERAL T

ASSUMPTIONS: $\{X_t\}_{t \in T}$ SUBGAUSSIAN AND SEPARABLE

14/ IDEAS: - DISCRETIZE THE SPACE

- EXPLOIT THE DEPENDENCIES $X_s \leftrightarrow X_t$

ε-NET ($\varepsilon > 0$): $(\mathcal{N}_\varepsilon, \pi)$, $\mathcal{N}_\varepsilon \subset T$, $\pi : T \to \mathcal{N}_\varepsilon$

SUCH THAT $d(t, \pi(t)) \leq \varepsilon \quad \forall t \in T$

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} X_{\pi(t)}\right] + \mathbb{E}\left[\sup_{t\in T} (X_t - X_{\pi(t)})\right]$$

MIGHT BE EASY TO CONTROL IF $\mathcal{N}_\varepsilon$ IS "SIMPLE"

SOMETHING SMALL

KEY IDEA: ITERATION ...

$$(\mathcal{N}_{\varepsilon/2}, \pi') \Rightarrow \begin{cases} \mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} X_{\pi(t)}\right] + \\ \quad + \mathbb{E}\left[\sup_{t\in T} \left(X_{\pi'(t)} - X_{\pi(t)}\right)\right] + \mathbb{E}\left[\sup_{t\in T} \left(X_t - X_{\pi'(t)}\right)\right] \end{cases}$$

15/ Case of Finite T $(|T| < \infty)$

$(\mathcal{N}_{2^{-k}}, \pi_k)$: Refining Sequence of Nets

- $\exists k_0 : \operatorname{diam}|T| < 2^{-k_0}$
- $\exists \bar{k} : d(s,t) > 2^{-\bar{k}}$ $\forall s,t \in T \quad (s \neq t)$

$\Rightarrow \mathcal{N}_{2^{-k_0}} = \{t_0\}$

$\pi_{\bar{k}}(t) = t \quad \forall t \in T$

Define $N_k = |\mathcal{N}_{2^{-k}}|$

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] = \cancel{\mathbb{E}[X_{t_0}]} + \sum_{k=k_0}^{\bar{k}-1} \mathbb{E}\Big[\underbrace{\sup_{t\in T}}_{\text{it's a max on } N_k \cdot N_{k+1} \le N_{k+1}^2 \text{ elements}} \underbrace{\left(X_{\pi_{k+1}(t)} - X_{\pi_k(t)}\right)}_{(3\cdot 2^{-k+1})^2 \text{- SUBGAUSSIAN}}\Big]$$

$$\overset{\text{MAXIMAL INEQUALITY}}{\lesssim} 6 \sum_{k=k_0}^{\bar{k}} 2^{-k}\sqrt{\log N_k}$$

16/ Case $|T| = \infty$

Theorem 1

$\{X_t\}_{t \in T}$ subgaussian and separable

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

Proof

By separability there is a sequence $\{t_j\}_{j \geq 1} \subset T$ s.t.

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] = \mathbb{E}\left[\sup_{j \geq 1} X_{t_j}\right] = \sup_{J \geq 1} \mathbb{E}\left[\sup_{j=1 \dots J} X_{t_j}\right]$$

Apply the previous bound to $\mathbb{E}\left[\sup_{j=1 \dots J} X_{t_j}\right]$ $\forall k$.

Conclude by noting that $N(\{t_j\}_{j=1 \dots J}, d, \varepsilon) \leq N(T, d, \varepsilon)$.

$\square$

17/ Remarks on the Chaining Bound

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

- RHS $= \infty$ if $\operatorname{diam} T = \infty$
- $\log N(T,d,2^{-k})$ is the Entropy of $\mathcal{N}_k$ under a uniform distribution
- In the Statistical Learning framework we can seek for an "Algorithm-Dependent" Bound, that is we are interested in something like $\mathbb{E}[X_{\tilde{c}}]$, rather than $\mathbb{E}[\sup_t X_t]$.
  ↳ random variable

18/ Chaining Mutual Information Bound

Goal: Find a generalization bound which

1 - Exploits dependencies between the hypotheses

(≃ chaining)

2 - Exploits dependence on the algorithm

(≃ mutual information)

[Chaining Mutual Information and Tightening Generalization Bounds, Asadi & Abbe & Verdù, 2018]

19/ THEOREM 2

(T, d) BOUNDED METRIC SPACE, $\operatorname{diam} T \leq 2^{-k_0}$

$\{X_t\}_{t \in T}$ SEPARABLE AND SUBGAUSSIAN PROCESS

$(\mathcal{N}_{2^{-k}}, \pi_k)$ SEQUENCE OF "REFINING" NETS

such that $\pi_{k-1} \circ \pi_k = \pi_k$

τ RANDOM VARIABLE WITH VALUES IN T

$$\Rightarrow \mathbb{E}[X_\tau] \leq 3 \sum_{k=k_0+1}^{\infty} 2^{-k} \sqrt{2\, I(\tau_k; X_T)}$$

with $\tau_k = \pi_k(\tau)$, $X_T = \{X_t\}_{t \in T}$

Remark: $I(\tau_k; X_T) \leq H(\tau_k) \leq \log\left(N(T, d, 2^{-k})\right)$

20/ Proof

$$\mathbb{E}[X_{\tilde{\tau}}] = \mathbb{E}[X_{t_0}] + \sum_{k=K_0+1}^{\infty} \mathbb{E}[X_{\tilde{\tau}_k} - X_{\tilde{\tau}_{k-1}}]$$

$(3\cdot 2^{-k})^2$-SUBGAUSS

LEMMA 1

(X,Y), $(\bar{X},\bar{Y})$: $P_{\bar{X},\bar{Y}} \sim P_{\bar{X}} \otimes P_{\bar{Y}}$

$\ell(\bar{X},\bar{Y})$ σ^2-SUBGAUSS wrt $P_{\bar{X},\bar{Y}}$

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

APPLY LEMMA 1 WITH

$X \leftrightarrow (\tilde{\tau}_{k-1}, \tilde{\tau}_k)$

$Y \leftrightarrow X_T$

$$\ell((\tilde{\tau}_{k-1},\tilde{\tau}_k), X_T) = X_{\tilde{\tau}_k} - X_{\tilde{\tau}_{k-1}}$$

$$\Rightarrow \mathbb{E}[X_{\tilde{\tau}_k} - X_{\tilde{\tau}_{k-1}}] \le 3\cdot 2^{-k}\sqrt{2\, I((\tilde{\tau}_{k-1},\tilde{\tau}_k), X_T)}$$

SINCE $\Pi_{k-1} \circ \Pi_k = \Pi_k \Rightarrow (\tilde{\tau}_{k-1}, \tilde{\tau}_k) \mapsto \tilde{\tau}_k$ IS A BIJECTION

$$\Rightarrow I((\tilde{\tau}_{k-1},\tilde{\tau}_k), X_T) = I(\tilde{\tau}_k, X_T). \qquad \square$$

21/ Back to Statistical Learning...

Proposition 3

$(\mathcal{W}, d)$ Bounded Metric Space $\left(\operatorname{diam}\mathcal{W} \leq 2^{-k_0}\right)$

$\{gen(w)\}_{w\in\mathcal{W}} = \{L_P(w) - L_S(w)\}_{w\in\mathcal{W}}$ | Subgaussian and Separable Process on $(\mathcal{W}, d)$

$(\mathcal{N}_{2^{-k}}, \Pi_k)$ Refining sequence of nets on $\mathcal{W}$

$$\Rightarrow \mathcal{G} = \mathbb{E}_{S,W}\left[gen(W)\right] \leq 3\sum_{k=k_0+1}^{\infty} 2^{-k}\sqrt{2\, I(\Pi_k(W); S)}$$

Remark: $I(\Pi_k(W); S) \leq I\left(\Pi_k(W); \{gen(w)\}_{w\in\mathcal{W}}\right)$ since

$\{gen(w)\}_{w\in\mathcal{W}} \leftrightarrow S \leftrightarrow W \leftrightarrow \Pi_k(W)$ is a Markov chain.

22/ Some comments on the Chaining MI Bound

- In general it's tighter than the basic MI and Chaining Bounds.
- It requires some additional structure on W
- It can be finite when the MI bound is ∞
- The idea of chaining can be applied to other information-theoretic generalization bound
- It can be hard to apply to "real-life" algorithms

23/ Variants of the Mutual Information Bound

Wasserstein-Based Bound

MI Bound	Wass Bound
Subgaussianity of the loss	Lipschitzianity of the loss
Mutual Information	Wasserstein Distance
$I(S;W) = KL(P_{S,W} \| P_S \otimes P_W)$	$W(P_{S/W}, P_S)$
∞ if $S \mapsto W_S$ deterministic	Always finite

[Generalization error bounds using Wasserstein distances, Tovar Lopez & Jog, 2018]

24 / p-Wasserstein Distance

$\mathbb{P}$ and $\mathbb{Q}$ probability measures on some space $\mathcal{Z}$

$$W_p(\mathbb{P},\mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \left(\int_{\mathcal{Z}\times\mathcal{Z}} \underbrace{\|z - z'\|_p^p}_{\text{(in general } \mathrm{dist}(z,z'))} \, d\pi(z,z') \right)^{1/p}$$

with $\Pi(\mathbb{P},\mathbb{Q}) = \left\{ \begin{array}{l} \text{probability measures on } \mathcal{Z}\times\mathcal{Z} \text{ with} \\ \text{marginals } \mathbb{P} \text{ and } \mathbb{Q} \end{array} \right\}$

Remarks:

- W_p is a metric on the space of probability measures
- $\exists\, \pi^* \in \Pi(\mathbb{P},\mathbb{Q})$ which attains the inf

25 / PROPOSITION 4

ASSUME $\exists p \geq 1,\ C > 0 : \forall h, h' \in \mathcal{H}$

$$\| \ell(h, \cdot) - \ell(h', \cdot) \|_\infty \leq C \, \| h - h' \|_p$$

$$\Rightarrow |\mathcal{G}| \leq \frac{C}{M^{1/p}} \, \mathbb{E}_W \left[W_p(\mathbb{P}_S, \mathbb{P}_{S|W})^p \right]^{1/p}$$

Remarks:

- PRIOR KNOWLEDGE ON $\mathbb{P}_Z$ IS NOT REQUIRED
- DOES NOT HOLD FOR THE 0-1 LOSS

26/ Proof

1) Reformulation of $\mathcal{G}$

$$\mathcal{G} = \mathbb{E}_w\left[\mathbb{E}_S[L_S(w)] - \mathbb{E}_{S|w}[L_S(w)]\right]$$

$\bar{S}, \bar{\omega}$ copies of S, ω

$$\mathcal{G} = \mathbb{E}_\pi\left[L_{\bar{S}}(\bar{w}) - L_S(w)\right]$$

$\forall \pi$ probability on $(\bar{\omega}, \bar{S}) \times (\omega, S)$ with marginals

$P_w \otimes P_S$ on $(\bar{\omega}, \bar{S})$

$P_{w,S}$ on (ω, S)

$$\Rightarrow |\mathcal{G}| \leq \mathbb{E}_\pi\left[|L_{\bar{S}}(\bar{w}) - L_S(w)|\right]$$

2) Construction of π_0

Looking for π_0 such that:

- $\mathbb{P}_W \otimes \mathbb{P}_S$ on $(\bar{\Omega}, \bar{\mathcal{S}})$
- $\mathbb{P}_{W,S}$ on $(\Omega, \mathcal{S})$
- optimal Wasserstein coupling on $\bar{\mathcal{S}} \times \mathcal{S}$
- Diagonal on $\bar{\Omega} \times \Omega$

$\forall \omega$, π_ω on $\bar{\mathcal{S}} \times \mathcal{S}$ "optimal", i.e.

$$W_p^p(\mathbb{P}_S, \mathbb{P}_{S|W=\omega}) = \mathbb{E}_{\pi_\omega}\left[\|\bar{S} - S\|_p^p\right]$$

Define $\bar{\pi}_0$ on $\bar{\mathcal{S}} \times \mathcal{S} \times \Omega$ as

$$d\bar{\pi}_0(\bar{s}, s, \omega) = d\pi_\omega(\bar{s}, s)\, d\mathbb{P}_W(\omega)$$

28/ "EXTEND" $\overline{\Pi}_0$ TO $(\bar{\Omega}, \bar{\mathcal{S}}) \times (\Omega, \mathcal{S})$ SO THAT IT IS DIAGONAL ON $\bar{\Omega} \times \Omega$

$$D = \{(\bar{\omega}, \omega) \in \bar{\Omega} \times \Omega : \bar{\omega} = \omega\}$$

$$\underline{P} : \bar{\Omega} \times \Omega \twoheadrightarrow \Omega$$
$$(\bar{\omega}, \omega) \longmapsto \omega$$

DEFINE $$\Pi_0 \left(\underset{\bar{\mathcal{S}} \times \mathcal{S}}{\cap} A \times \underset{\bar{\Omega} \times \Omega}{\cap} B \right) = \overline{\Pi}_0 \left(A \times \underline{P}(B \cap D) \right)$$

Π_0 IS WELL DEFINED AS A PROBABILITY ON $(\bar{\Omega}, \bar{\mathcal{S}}) \times (\Omega, \mathcal{S})$ BY CARATHÉODORY'S THM.

MARGINALS: $\begin{cases} \mathbb{P}_W \otimes \mathbb{P}_S & \text{ON } (\bar{\Omega}, \bar{\mathcal{S}}) \\ \mathbb{P}_{W,S} & \text{ON } (\Omega, \mathcal{S}) \end{cases}$

3) Conclusion of the Proof

Lemme 2 : $\left[\|\ell(t,\cdot)-\ell(t',\cdot)\|_\infty \le C\|t-t'\|_p\right] \Rightarrow \left[\|L_s(\cdot)-L_{s'}(\cdot)\|_\infty \le \frac{C}{n^{1/p}}\|s-s'\|_p\right]$

$$|\mathcal{E}| \le \mathbb{E}_{\pi_0}\left[|L_{\bar{s}}(\bar{w})-L_s(w)|\right] = \mathbb{E}_{\bar{\pi}_0}\left[|L_{\bar{s}}(w)-L_s(w)|\right]$$

(π_0 is diagonal on $\bar{\omega}\times\omega$)

$$= \mathbb{E}_w\left[\mathbb{E}_{\pi_w}\left[|L_{\bar{s}}(w)-L_s(w)|\right]\right] \quad \text{(def of } \bar{\pi}_0\text{)}$$

$$\le \frac{C}{n^{1/p}}\,\mathbb{E}_w\left[\mathbb{E}_{\pi_w}\left[\|\bar{s}-s\|_p\right]\right] \quad \text{(Lemme 2)}$$

$$\le \frac{C}{n^{1/p}}\,\mathbb{E}_w\left[\mathbb{E}_{\pi_w}\left[\|\bar{s}-s\|_p^p\right]\right]^{1/p} \quad (p\ge 1,\ \text{Jensen})$$

$$\le \frac{C}{n^{1/p}}\,\mathbb{E}_w\left[W_p^p\left(P_s, P_{s/w}\right)\right]^{1/p} \quad (\pi_w \text{ is "optimal"})$$

□

30/ ISMI Bound

Individual Sample Mutual Information:

Mutual Information + Pointwise Stability

Proposition 5

1) If $\ell(w,Z)$ is σ^2-subgaussian under P_Z, $\forall w \in \mathcal{W}$

$$\Rightarrow |\mathcal{G}| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma^2 I(W;Z_i)}$$

2) If $\ell(W,Z)$ is σ^2-subgaussian under $P_W \otimes P_Z$

$$\Rightarrow |\mathcal{G}| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma^2 I(W;Z_i)}$$

[Tightening Mutual Information Based Bounds on Generalization Error, Bu & Zou & Veeravalli, 2020]

31/ Conditional Mutual Information Bound

Conditional Mutual Information

$$I(X;Y|Z) = \mathbb{E}_Z[I(X|Z;Y|Z)]$$

$\tilde{S}$ a $2n$-"super sample", $\tilde{S} = (S^0, S^1)$

U uniformly distributed on $\{0,1\}^n$

$\tilde{S}^U = (Z_1^{U_1}, \ldots, Z_n^{U_n})$ n-sample chosen from $\tilde{S}$ "according to U"

$$CMI = I(W_{\tilde{S}^U}; U | \tilde{S})$$

Proposition 6

$$|\mathcal{E}_\mu| \leq \sqrt{\frac{2}{n} \mathbb{E}_{Z \sim \tilde{Z}} \left[\sup_{w \in W} \left(\ell(w,Z) - \ell(w,Z') \right)^2 \right] \cdot CMI}$$

[Reasoning About Generalization via Conditional Mutual Information, Steinke & Zakynthinou, 2020]

32/ Some Final Comments

- Wasserstein-based bound is hard to compare with MI bound, but it can be much tighter.
- Wasserstein distance might be very hard to estimate
- Both ISMI and conditional MI are tighter than the MI bound.
- Chaining techniques can be applied to ISMI and conditional MI
- Some recent papers proposed new algorithms inspired by information-theoretic generalization bounds
- Applications to Langevin dynamics algorithm (noisy iterative algorithm)
- It might be of interest to study more in depth the connections with PAC-Bayesian bounds

THANK YOU FOR YOUR ATTENTION

Bibliography

[How much does your data exploration overfit? Controlling bias via information usage, Russo & Zou, 2016]
[Information-theoretic analysis of generalization capability of learning algorithms, Xu & Raginsky, 2017]
[Probability in High Dimensions, Van Handel, 2016]
[Chaining Mutual Information and Tightening Generalization Bounds, Asadi & Abbe & Verdù, 2018]
[Generalization error bounds using Wasserstein distances, Tovar Lopez & Jog, 2018]
[Tightening Mutual Information Based Bounds on Generalization Error, Bu & Zou & Veeravalli, 2020]
[Reasoning About Generalization via Conditional Mutual Information, Steinke & Zakynthinou, 2020]

[Conditioning and Processing: Techniques to Improve Information-Theoretic Generalization Bounds, Hafez-Kolahi et al., 2020]
[Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms, Haghifam et al., 2020]
[An Information-Theoretic View of Generalization via Wasserstein Distance, Wang et al., 2019]
[Chaining Meets Chain Rule: Multilevel Entropic Regularization and Training of Neural Networks, Asadi & Abbe, 2020]