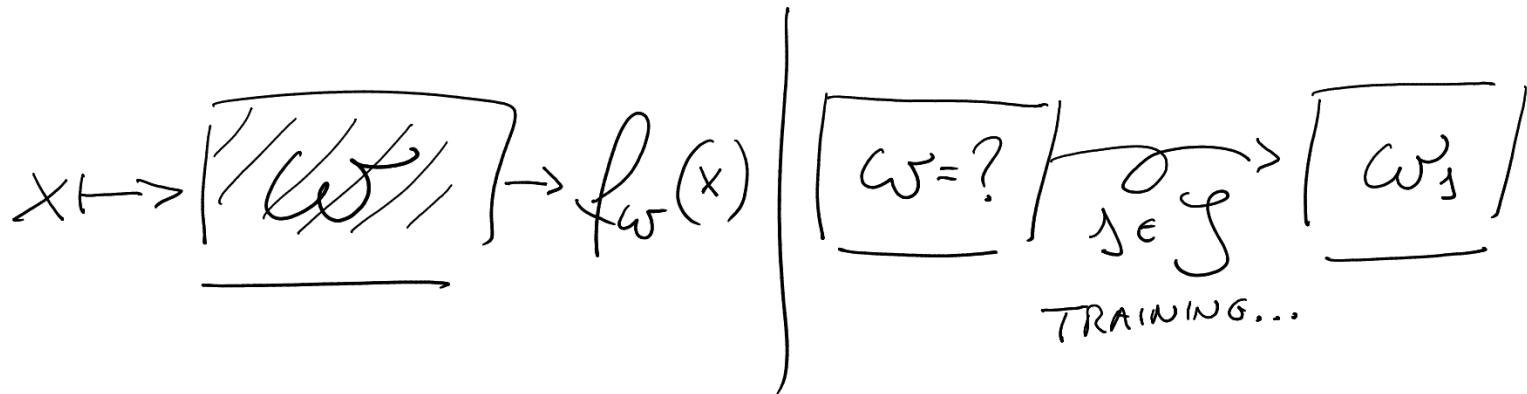
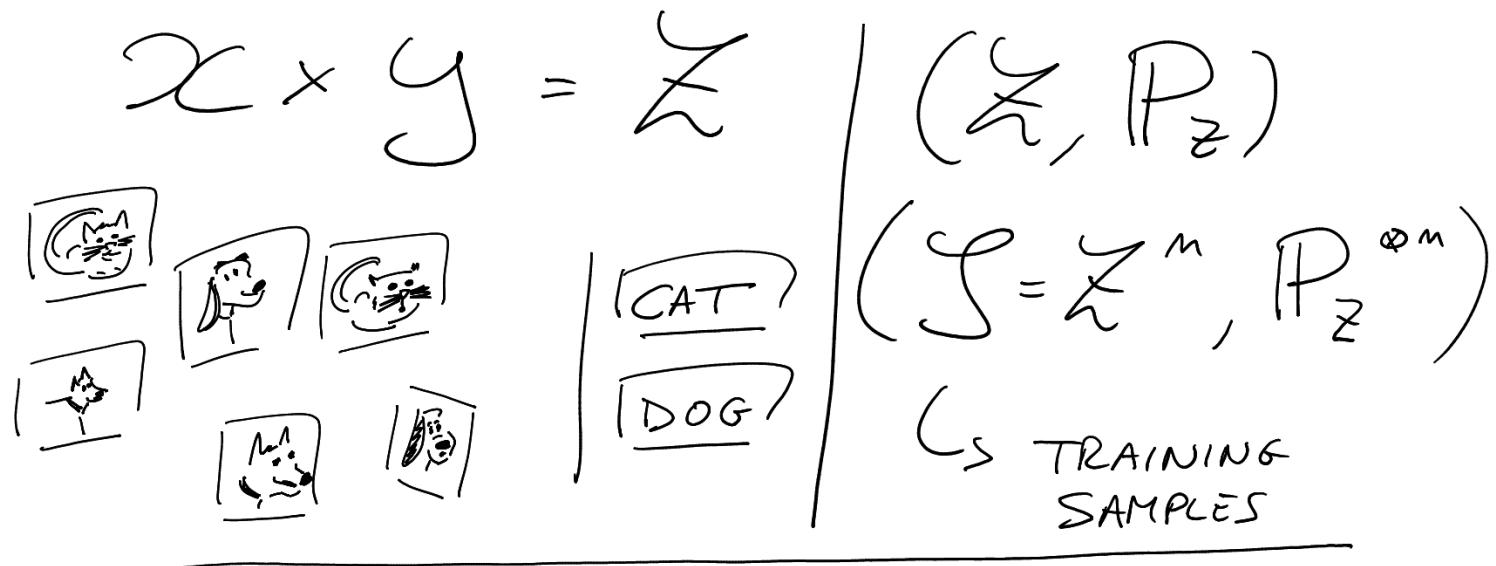


# INFORMATION-THEORETIC GENERALIZATION BOUNDS

8 FEB 2021

EUGENIO CLERICO

# 1/ STATISTICAL LEARNING



## 2/ JUDGING AN ALGORITHM

$$\ell: \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$$

SQUARED LOSS:  $\ell(\omega, (x, y)) = \frac{1}{2} (f_\omega(x) - y)^2$

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

---

POPULATION RISK:  $L_p(\omega) = E_Z [\ell(\omega, Z)]$

EMPIRICAL RISK:  $L_s(\omega) = \frac{1}{N} \sum_{i=1}^m \ell(\omega, Z_i)$   
( $s = (Z_1 \dots Z_m)$ )

GOAL: TO FIND A "GOOD"  $\hat{\mathcal{P}}_{\mathcal{W} | s}$

3/ ALGORITHM :  $\mathcal{S} \xrightarrow{\text{ } \eta \text{ }} W \sim P_{W|S=\mathcal{S}}$

---

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

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

## 4/ GENERALIZATION Bounds

→ VARIOUS IDEAS:

- COMPLEXITY OF  $W$
- PROPERTIES OF THE ALGORITHM
- DEPENDENCY OUTPUT - TRAINING SAMPLE

---

Bounds BASED ON INFORMATION MEASURES  $\longleftrightarrow$  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 BOUNDS

### KULLBACK - LEIBLER DIVERGENCE

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

$KL \geq 0$   
 $KL = 0 \Leftrightarrow P = Q$   
 $KL$  might be  $\infty$

### VARIATIONAL CHARACTERIZATION

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

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

[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

$$(\underline{X}, Y) \sim P_{\underline{X}, Y}$$

$$I(\underline{X}; Y)$$

MEASURES THE "DEPENDENCE"

$$\underline{X} \longleftrightarrow Y$$

$$I(\underline{X}; Y) = KL(P_{\underline{X}, Y} \parallel \underbrace{P_{\underline{X}} \otimes P_Y}_{\text{As if } \underline{X} \perp\!\!\!\perp Y})$$

$$\begin{array}{c} \text{if } \bar{\underline{X}} \sim \underline{X}, \bar{Y} \sim Y \\ \bar{\underline{X}} \perp\!\!\!\perp \bar{Y} \end{array}$$

$$\Rightarrow I(\underline{X}; Y) = KL(P_{\underline{X}, Y} \parallel P_{\bar{\underline{X}}, \bar{Y}})$$

## SUBGAUSSIANITY

$\underline{X}$  IS  $\sigma^2$ -SUBGAUSSIAN

$$E[e^{\lambda \underline{X}}] \leq e^{\lambda E[\underline{X}]} e^{\frac{\lambda^2 \sigma^2}{2}}$$

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

EXAMPLES

$\underline{X} \sim N(\mu, \sigma^2)$  IS  $\sigma^2$ -SUBGAUSS

$\underline{X} \in [\alpha, \beta]$  IS  $\frac{(\beta - \alpha)^2}{4}$ -SUBGAUSS

~~X~~ / LEMMA 1

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

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

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

Proof

$$\begin{aligned} \forall \lambda \quad I(\bar{X}, Y) &\stackrel{\text{Variational def of KL}}{\geq} \lambda \mathbb{E}[f(\bar{X}, Y)] - \log \mathbb{E}[e^{\lambda f(\bar{X}, Y)}] \\ &\stackrel{\text{Subgauss}}{\geq} \lambda (\mathbb{E}[f(\bar{X}, Y)] - \mathbb{E}[f(\bar{\bar{X}}, \bar{Y})]) + \frac{\lambda^2 \sigma^2}{2} \end{aligned}$$

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

# 8/ BACK TO STATISTICAL LEARNING...

$$G = \overline{E}_{S, w} [L_P(w) - L_S(w)]$$

$$L_P(w) = E_z [\ell(w, z)] = E_S [L_S(w)] \quad \forall \text{fixed } w$$

$$\Rightarrow \overline{E}_{S, w} [L_P(w)] = \overline{E}_w [L_P(w)] = \overline{E}_w [\overline{E}_S [L_S(w)]]$$

$$G = \overline{E}_{S \otimes w} [L_S(w)] - \overline{E}_{S, w} [L_S(w)]$$

Let  $\bar{w} \sim w$ ,  $\bar{s} \sim S$ ,  $\bar{w} \perp\!\!\!\perp \bar{s}$

$$G = \overline{E}_{\bar{s}, \bar{w}} [L_{\bar{s}}(\bar{w})] - \overline{E}_{S, w} [L_S(w)]$$

# 9/ SIMPLE BOUND

## PROPOSITION 1

$\ell(\omega, Z)$   $\sigma^2$ -SUBGAUSS  $\forall \omega$

$$\Rightarrow |\ell| \leq \sqrt{\frac{2\sigma^2}{m} I(S; W)}$$

## Proof

$$\text{LET } \ell(S, \omega) = L_S(\omega)$$

$\ell(\omega, \cdot)$   $\sigma^2$ -SUBGAUSS  $\Rightarrow \ell(S, \omega)$   $\frac{\sigma^2}{m}$ -SUBGAUSS

$\Rightarrow \ell(\bar{S}, \bar{\omega})$   $\frac{\sigma^2}{m}$ -SUBGAUSS

$\Rightarrow$  APPLY LEMMA 1 AND CONCLUDE.  $\square$

## LEMMA 1

$(X, Y), (\bar{X}, \bar{Y}) : P_{\bar{X}, \bar{Y}} \sim P_X \otimes P_Y$   
 $\ell(\bar{X}, \bar{Y})$   $\sigma^2$ -SUBGAUSS w.r.t  $P_{\bar{X}, \bar{Y}}$

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

# 10/ More General (and Tighter!) Bound

## Proposition 2

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

$$\log E_S [e^{\lambda (L_P(\omega) - L_S(\omega))}] \leq \Psi(\lambda) \quad \forall \lambda > 0 \quad \forall \omega$$

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

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

REMARK: If  $\Psi(\lambda) = \frac{\lambda^2 \sigma^2}{2m} \Rightarrow \frac{\sigma^2}{m}$ -SUBGAUSS

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

# 11/ SOME COMMENTS ON THE MI BOUND

- IT MIGHT BE LOOSE OR EVEN  $\infty$   
(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

$\{\bar{X}_t\}_{t \in T}$  IS SUBGAUSSIAN IF  $\forall \lambda \geq 0, \forall s, t \in T$

$$E[e^{\lambda(\bar{X}_s - \bar{X}_t)}] \leq e^{\frac{\lambda^2}{2}d(s, t)^2}; E[\bar{X}_t] = 0$$

---

### SEPARABLE PROCESS

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

$\bar{X}_t \in \lim_{\substack{s \rightarrow t \\ s \in S}} \bar{X}_s \quad a.s.$

# 13/ MAXIMAL INEQUALITY

$(\bar{X}_1, \dots, \bar{X}_m)$  FINITE COLLECTION OF RVs

$\bar{X}_i$ :  $\sigma^2$ -SUBGAUSS  $\forall i = 1 \dots m$

$$\Rightarrow \mathbb{E}[\sup_{i=1 \dots m} \bar{X}_i] \leq \sqrt{2\sigma^2 \log m}$$

---

GOAL: FIND AN UPPER BOUND FOR  $\mathbb{E}[\sup_{t \in T} \bar{X}_t]$   
FOR A GENERAL  $T$

ASSUMPTIONS:  $\{\bar{X}_t\}_{t \in T}$  SUBGAUSSIAN AND SEPARABLE

14 / IDEAS:

- DISCRETIZE THE SPACE

- EXPLOIT THE DEPENDENCIES  $\bar{X}_t \leftrightarrow \bar{X}_{\pi(t)}$

---

$\epsilon$ -NET :  $(N_\epsilon, \pi)$ ,  $N_\epsilon \subset T$ ,  $\pi: T \rightarrow N_\epsilon$   
 $(\epsilon > 0)$

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

---

$$\mathbb{E}[\sup_{t \in T} \bar{X}_t] \leq \underbrace{\mathbb{E}[\sup_{t \in T} \bar{X}_{\pi(t)}]}_{\text{MIGHT BE EASY TO CONTROL IF } N_\epsilon \text{ IS "SIMPLE"}} + \underbrace{\mathbb{E}[\sup_{t \in T} (\bar{X}_t - \bar{X}_{\pi(t)})]}_{\text{SOMETHING SMALL}}$$

MIGHT BE EASY TO  
CONTROL IF  $N_\epsilon$  IS  
"SIMPLE"

SOMETHING  
SMALL

KEY IDEA: ITERATION ...

$$(N_{\epsilon_2}, \pi') \Rightarrow \left\{ \begin{array}{l} \mathbb{E}[\sup_{t \in T} \bar{X}_t] \leq \mathbb{E}[\sup_{t \in T} \bar{X}_{\pi'(t)}] + \\ + \mathbb{E}[\sup_{t \in T} (\bar{X}_{\pi'(t)} - \bar{X}_{\pi(t)})] + \mathbb{E}[\sup_{t \in T} (\bar{X}_t - \bar{X}_{\pi'(t)})] \end{array} \right.$$

# 15/ CASE OF FINITE T ( $|T| < \infty$ )

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

- $\exists_{K_0} : \text{diam } |T| < 2^{-K_0}$
  - $\exists \bar{k} : d(s, +) > 2^{-\bar{k}}$
- $\forall s, t \in T \quad (s \neq t)$
- $\mathcal{N}_{2^{-K_0}} = \{t_0\}$
- $\Rightarrow \pi_{\bar{k}}(+) = + \quad \forall t \in T$

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

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

MAXIMAL INEQUALITY

$$\leq 6 \sum_{k=k_0}^{\bar{k}} 2^{-k} \sqrt{\log N_k}$$

$N_k \cdot N_{k+1} \leq N_{k+1}^2$  elements,

$(3 \cdot 2^{-k+1})^2$  - SUBGAUSSIAN

# 16/ CASE $|T| = \infty$

## THEOREM 1

$\{\bar{X}_t\}_{t \in T}$  SUBGAUSSIAN AND SEPARABLE

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

## Proof

By SEPARABILITY THERE IS A SEQUENCE  $\{t_j\}_{j \geq 1} \subset T$  S.T.

$$\mathbb{E}[\sup_{t \in T} \bar{X}_t] = \mathbb{E}[\sup_{j \geq 1} \bar{X}_{t_j}] = \sup_{J \geq 1} \mathbb{E}[\sup_{j=1 \dots J} \bar{X}_{t_j}]$$

APPLY THE PREVIOUS BOUND TO  $\mathbb{E}[\sup_{j=1 \dots J} \bar{X}_{t_j}] \forall k$ .

CONCLUDE BY NOTING THAT  $N(\{t_j\}_{j=1 \dots J}, d, \varepsilon) \leq N(T, d, \varepsilon)$ .

□

# 17/ REMARKS ON THE CHAINING BOUND

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

- RHS =  $\infty$  if  $\text{diam } T = \infty$
- $\log N(T, d, 2^{-k})$  IS THE ENTROPY OF  $N_k$   
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  
 $\mathbb{E}[X_{\tilde{z}}]$ , 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

( $\approx$  CHAINING)

2 - EXPLOITS DEPENDENCE ON THE ALGORITHM

( $\approx$  MUTUAL INFORMATION)

# 19/ THEOREM 2

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

$\{\bar{X}_t\}_{t \in T}$  SEPARABLE AND SUBGAUSSIAN PROCESS

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

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

$\tilde{Z}$  RANDOM VARIABLE WITH VALUES IN  $\overline{T}$

$$\Rightarrow E[\bar{X}_{\tilde{Z}}] \leq 3 \sum_{k=k_0+1}^{\infty} 2^{-k} \sqrt{2 I(\tilde{Z}_k; \bar{X}_T)}$$

with  $\tilde{Z}_k = \pi_k(\tilde{Z})$ ,  $\bar{X}_T = \{\bar{X}_t\}_{t \in T}$

---

Remark:  $I(\tilde{Z}_k; \bar{X}_T) \leq H(\tilde{Z}_k) \leq \log(N(T, d, 2^{-k}))$

# 20 / Proof

$$E[\bar{X}_z] = E[\cancel{\bar{X}_{t_0}}] + \sum_{k=K_0+1}^{\infty} E[\tilde{\bar{X}}_{\tilde{z}_k} - \bar{X}_{\tilde{z}_{k-1}}]$$

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

LEMMA 1

$$\begin{aligned} & (\bar{X}, Y), (\tilde{X}, \tilde{Y}) : P_{\bar{X}, Y} \sim P_{\tilde{X}} \circ P_Y \\ & f(\bar{X}, Y) \text{ } \sigma^2\text{-SUBGAUSS w.r.t. } P_{\bar{X}, Y} \\ \Rightarrow & |E[f(\bar{X}, Y)] - E[f(\tilde{X}, \tilde{Y})]| \leq \sqrt{2\sigma^2 I(\bar{X}; Y)} \end{aligned}$$

APPLY LEMMA 1 WITH

$$\begin{cases} \bar{X} \leftrightarrow (\tilde{z}_{k-1}, \tilde{z}_k) \\ Y \leftrightarrow \bar{X}_T \end{cases} \quad \left\{ f((\tilde{z}_{k-1}, \tilde{z}_k), \bar{X}_T) = \bar{X}_{\tilde{z}_k} - \bar{X}_{\tilde{z}_{k-1}} \right.$$

$$\Rightarrow E[\bar{X}_{\tilde{z}_k} - \bar{X}_{\tilde{z}_{k-1}}] \leq 3 \cdot 2^{-k} \sqrt{2 I((\tilde{z}_{k-1}, \tilde{z}_k), \bar{X}_T)}$$

SINCE  $\overline{\Pi}_{k-1} \circ \overline{\Pi}_k = \overline{\Pi}_k \Rightarrow (\tilde{z}_{k-1}, \tilde{z}_k) \mapsto \tilde{z}_k$  IS A BIJECTION

$$\Rightarrow I((\tilde{z}_{k-1}, \tilde{z}_k), \bar{X}_T) = I(\tilde{z}_k, \bar{X}_T)$$

□

# 21/ BACK TO STATISTICAL LEARNING...

## PROPOSITION 3

$(\mathcal{W}, d)$  BOUNDED METRIC SPACE ( $\text{diam } \mathcal{W} \leq 2^{-k_0}$ )

$$\left\{ \text{gen}(w) \right\}_{w \in \mathcal{W}} = \left\{ L_p(w) - L_s(w) \right\}_{w \in \mathcal{W}} \quad \begin{cases} \text{SUBGAUSSIAN AND SEPARABLE} \\ \text{PROCESS ON } (\mathcal{W}, d) \end{cases}$$

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

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

Remark:  $I(\pi_k(W); S) \leq I(\pi_k(W); \{\text{gen}(w)\}_{w \in \mathcal{W}})$  since

$\{\text{gen}(w)\}_{w \in \mathcal{W}} \leftrightarrow S \leftrightarrow W \leftrightarrow \pi_k(W)$  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  $\infty$
- 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 BOUNDS

### WASSERSTEIN-BASED BOUND

MI Bound

SUBGAUSSIANITY  
OF THE LOSS

MUTUAL INFORMATION

$$I(s; w) = KL(P_{s,w} \parallel P_s \otimes P_w)$$

as if  $s \mapsto w_s$  DETERMINISTIC

WASS Bound

LIPSCHITZIANITY  
OF THE LOSS

WASSERSTEIN DISTANCE

$$W(P_{s|w}, P_s)$$

ALWAYS FINITE

# 24 / P-WASSERSTETN DISTANCE

$P$  AND  $Q$  PROBABILITY MEASURES ON SOME SPACE  $\mathbb{Z}$

$$W_p(P, Q) = \inf_{\pi \in \overline{\Pi}(P, Q)} \left( \sum_{\mathbb{Z} \times \mathbb{Z}} \|z - z'\|_p^p d\pi(z, z') \right)^{1/p}$$

(in general  $\text{dist}(z, z')$ )

with  $\overline{\Pi}(P, Q) = \underbrace{\{ \text{PROBABILITY MEASURES ON } \mathbb{Z} \times \mathbb{Z} \text{ WITH} \}}_{\text{MARGINALS } P \text{ AND } Q}$

- Remarks:
- $W_p$  IS A METRIC ON THE SPACE OF PROBABILITY MEASURES
  - $\exists \pi^* \in \overline{\Pi}(P, Q)$  WHICH ATTAINS THE  $\inf$

## 25 / Proposition 4

Assume  $\exists p \geq 1, C > 0 : \forall z, z' \in \mathcal{Z}$

$$\|\ell(z, \cdot) - \ell(z', \cdot)\|_{\infty} \leq C \|z - z'\|_p$$

$$\Rightarrow |G| \leq \frac{C}{M^{1/p}} \left( \mathbb{E}_w [\mathcal{W}_p(P_S, P_{S|w})^p] \right)^{1/p}$$

---

Remarks:

- PRIOR KNOWLEDGE ON  $P_Z$  IS NOT REQUIRED
- DOES NOT HOLD FOR THE 0-1 LOSS

# 26 / Proof

## 1) REFORMULATION OF $\mathcal{G}$

$$\mathcal{G} = \mathbb{E}_w [ \mathbb{E}_s [ L_s(w) ] - \mathbb{E}_{s|w} [ L_s(w) ] ]$$

$\bar{s}, \bar{w}$  COPIES OF  $s, w$

$$\mathcal{G} = \mathbb{E}_{\pi} [ L_{\bar{s}}(\bar{w}) - L_s(w) ]$$

$\forall \pi$  PROBABILITY ON  $(\bar{w}, \bar{s}) \times (w, s)$  WITH MARGINALS

$P_w \otimes P_s$  ON  $(\bar{w}, \bar{s})$

$P_{w,s}$  ON  $(w, s)$

$$\Rightarrow |\mathcal{G}| \leq \mathbb{E}_{\pi} [ |L_{\bar{s}}(\bar{w}) - L_s(w)| ]$$

# 27/ 2) CONSTRUCTION OF $\Pi_0$

LOOKING FOR  
 $\Pi_0$  SUCH THAT:

- $P_w \otimes P_s$  on  $(\bar{W}, \bar{S})$
  - $P_{w,s}$  on  $(W, S)$
  - OPTIMAL WASSERSTEIN COUPLING on  $\bar{S} \times S$
  - DIAGONAL on  $\bar{W} \times W$
- 

$\forall w, \Pi_w$  on  $\bar{S} \times S$  "OPTIMAL", i.e.

$$W_p^p(P_s, P_{S|W=w}) = \mathbb{E}_{\Pi_w} [\|\bar{S} - S\|_p^p]$$

DEFINE  $\bar{\Pi}_0$  on  $\bar{S} \times S \times W$  AS

$$d\bar{\Pi}_0(\bar{s}, s, w) = d\Pi_w(\bar{s}, s) dP_w(w)$$

28/ "Extend"  $\bar{\pi}_0$  to  $(\bar{\omega}, \bar{S}) \times (\omega, S)$  so that

IT IS DIAGONAL ON  $\bar{\omega} \times \omega$

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

$$\begin{aligned} P : \bar{\omega} \times \omega &\rightarrow \omega \\ (\bar{\omega}, \omega) &\mapsto \omega \end{aligned}$$

DEFINE  $\pi_0(A \times B) = \bar{\pi}_0(A \times P(B \cap D))$

$$\begin{matrix} \cap & \cap \\ \bar{\omega} \times \omega & \bar{\omega} \times \omega \end{matrix}$$

$\pi_0$  IS WELL DEFINED AS A PROBABILITY ON  $(\bar{\omega}, \bar{S}) \times (\omega, S)$   
BY CARATHÉODORY'S THM.

MARGINALS:  $\left\{ \begin{array}{ll} P_w \otimes P_S & \text{on } (\bar{\omega}, \bar{S}) \\ P_{w,S} & \text{on } (\omega, S) \end{array} \right.$

# 29 / 3) Conclusion of the Proof

$$\text{Lemme 2 : } \left[ \|l(\tau, \cdot) - l(\tau', \cdot)\|_{\infty} \leq C \|\tau - \tau'\|_p \right] \Rightarrow \left[ \|L_s(\cdot) - L_{s'}(\cdot)\|_{\infty} \leq \frac{C}{n^{1/p}} \|s - s'\|_p \right]$$

$$|\mathcal{G}| \leq \mathbb{E}_{\pi_0} [ |L_{\bar{s}}(\bar{w}) - L_s(w)| ] \stackrel{\pi_0 \text{ is diagonal on } \bar{W} \times W}{=} \mathbb{E}_{\bar{\pi}_0} [ |L_{\bar{s}}(w) - L_s(w)| ]$$

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

$$\stackrel{\text{Lemme 2}}{\leq} \frac{C}{n^{1/p}} \mathbb{E}_w [ \mathbb{E}_{\pi_{\bar{w}}} [ \| \bar{s} - s \|_p ] ]$$

$$\stackrel{P \geq 1}{\stackrel{\text{Jensen}}{\leq}} \frac{C}{n^{1/p}} \mathbb{E}_w [ \mathbb{E}_{\pi_{\bar{w}}} [ \| \bar{s} - s \|_p^p ] ]^{1/p}$$

$$\stackrel{\pi_w \text{ is "optimal"}}{\leq} \frac{C}{n^{1/p}} \mathbb{E}_w [ W_p^p ( P_s, P_{s|w} ) ]^{1/p} . \quad \square$$

# 30 / ISMI Bound

INDIVIDUAL SAMPLE MUTUAL INFORMATION :

MUTUAL INFORMATION + POINTWISE STABILITY

## PROPOSITION 5

1) If  $\ell(w, z)$  is  $\sigma^2$ -SUBGAUSSIAN under  $P_z$ ,  $\forall w \in W$

$$\Rightarrow |\mathcal{G}| \leq \frac{1}{m} \sum_{i=1}^m \sqrt{2\sigma^2 I(w; z_i)}$$

2) If  $\ell(w, z)$  is  $\sigma^2$ -SUBGAUSSIAN under  $P_w \otimes P_z$

$$\Rightarrow |\mathcal{G}| \leq \frac{1}{m} \sum_{i=1}^m \sqrt{2\sigma^2 I(w; z_i)}$$

# 31 / CONDITIONAL MUTUAL INFORMATION BOUND

## CONDITIONAL MUTUAL INFORMATION

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

$\tilde{S}$  A  $2m$ -“SUPER SAMPLE”,  $\tilde{S} = (S^0, S^1)$

$\cup$  UNIFORMLY DISTRIBUTED ON  $\{0,1\}^m$

$\tilde{S}^0 = (Z_1^{01}, \dots, Z_m^{0m})$   $m$ -SAMPLE CHOSEN FROM  $\tilde{S}$   
“ACCORDING TO  $\cup$ ”

$$CMI = I(W_{\tilde{S}^0}; \cup | \tilde{S})$$

### PROPOSITION 6

$$|\mathcal{G}| \leq \sqrt{\frac{2}{m} E_{Z \in \tilde{S}^0} \left[ \sup_{w \in W} (\ell(w, Z) - \ell(w, Z'))^2 \right]} \cdot CMI$$

## 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

33/

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]

