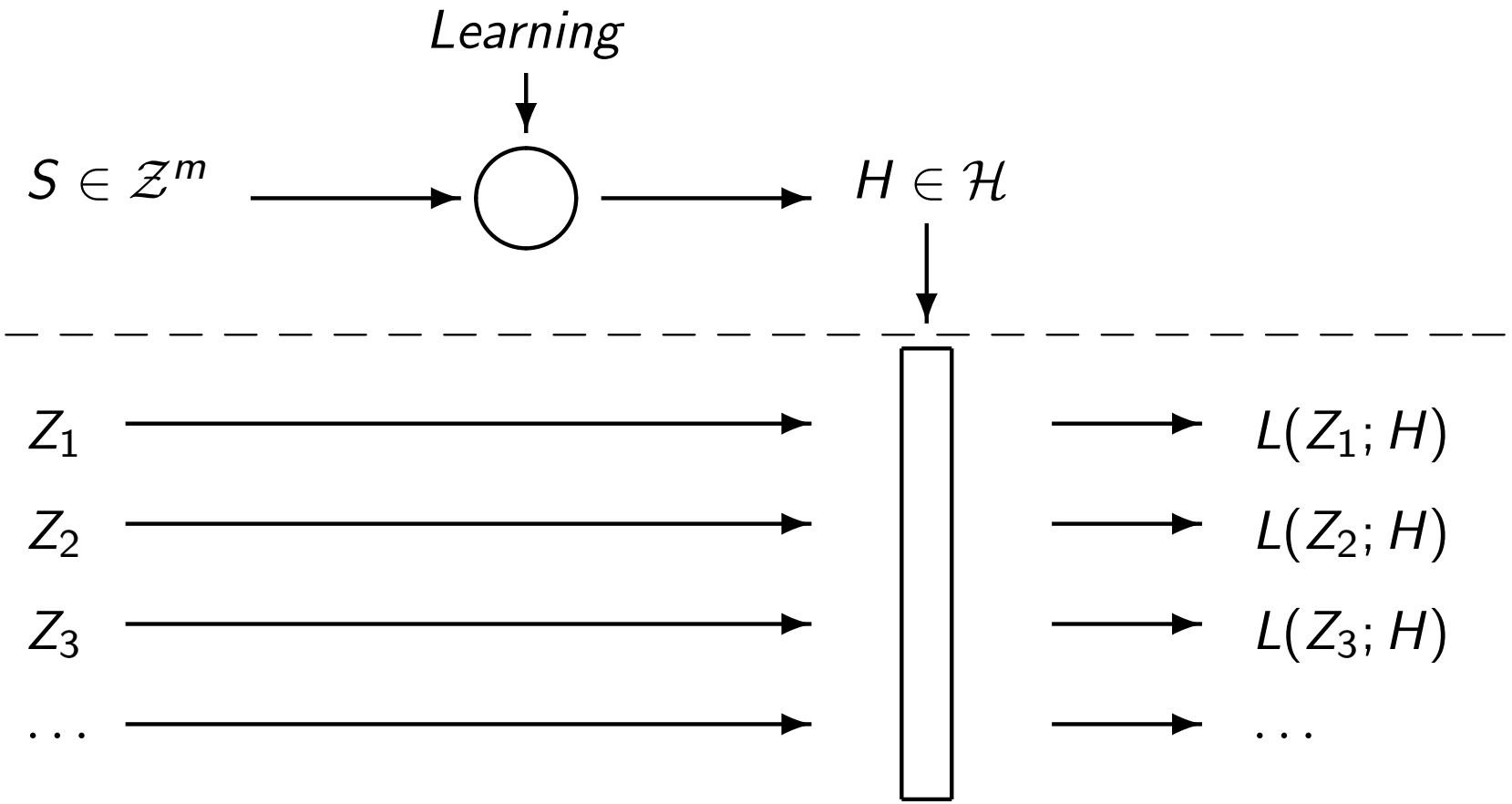
On the Interplay between Information, Stability, and Generalization

[Workshop on Adaptive Data Analysis]

Ibrahim Alabdulmohsin

December 9, 2016

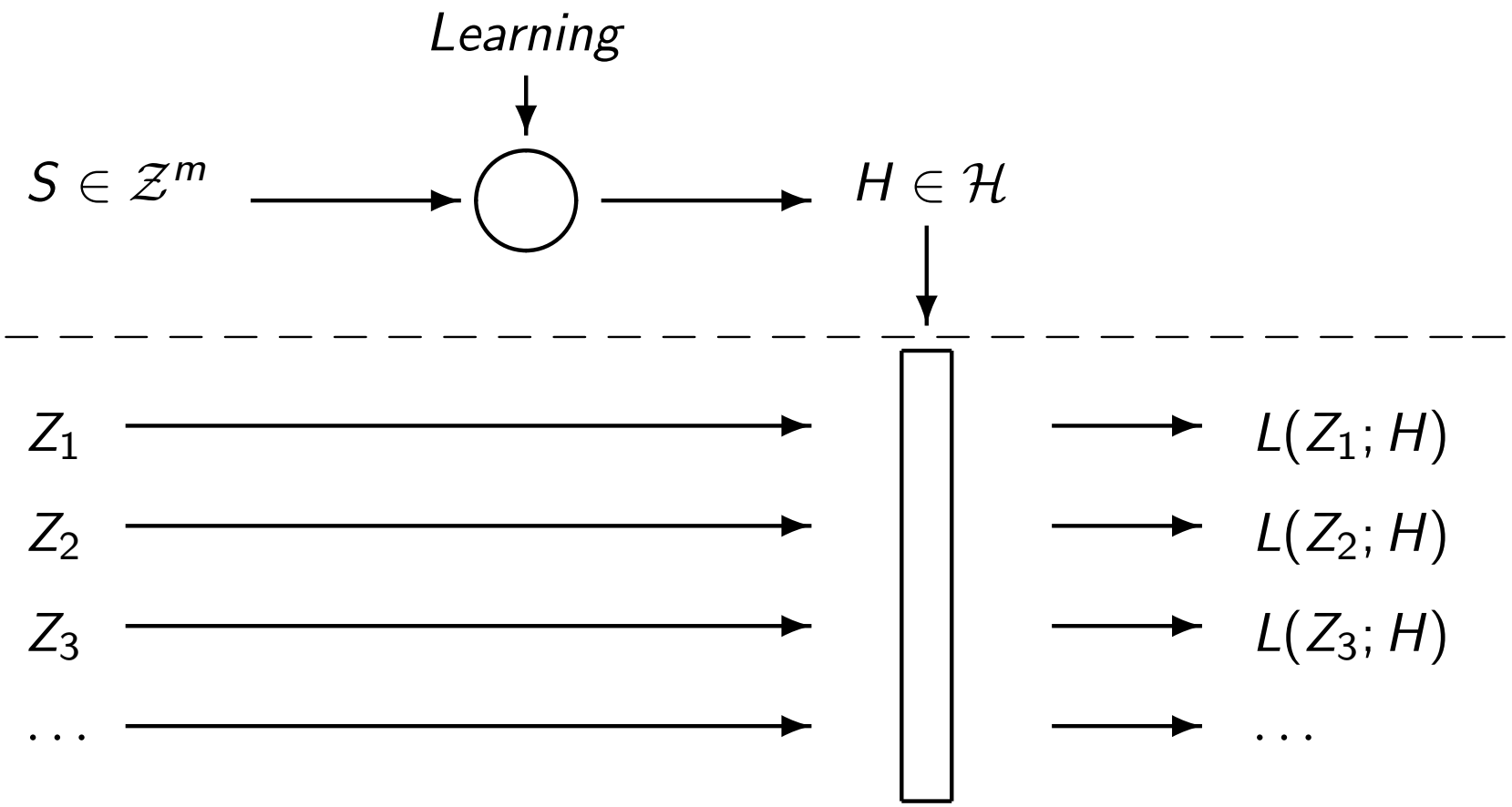
General Setting of Learning



We look into three aspects of the learning setting:

1. **Generalization**: Does the empirical performance faithfully represent the true performance of the algorithm?
2. **Information**: Does the hypothesis H reveal "lots" of information about the sample?
3. **Stability**: Will H be heavily "impacted" by a "small" perturbation in the training sample?

In what sense, if any, are they equivalent?


Learning
$S \in \mathcal{Z}^m$
$H \in \mathcal{H}$
Z_1
Z_2
Z_3
...
$L(Z_1; H)$
$L(Z_2; H)$
$L(Z_3; H)$
...

Uniform Generalization

Definition (Uniform Generalization)

A learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ generalizes uniformly with rate $\epsilon > 0$, if for all parametric losses and all distributions $p(z)$ on $\mathcal{Z}$, we have $|R_{emp}(\mathcal{L}) - R_{true}(\mathcal{L})| \leq \epsilon$.

All risks are defined in expectation over the randomness of the sample and the internal randomness of the algorithm.

Equivalence Relationship

Theorem (NIPS, 2015)

The uniform generalization rate is **equal** *to*
$\mathcal{J}(Z_{trn};\, H) = ||p(Z_{trn})p(H)\,,\;\; p(Z_{trn}, H)||_{\mathcal{T}}$

This is the mutual information, measured in the total variation distnace.

The risk of overfitting depends on many factors:

1. The hypothesis space $\mathcal{H}$.
2. The domain $\mathcal{Z}$.
3. The learning algorithm $\mathcal{L}$.
4. $\ldots$

Does $\mathcal{J}(Z_{trn}; H)$ capture the phenomenon of overfitting in its full generality?

Stability

We write:

$$\mathcal{J}(Z_{trn};\, H) = \mathbb{E}_{Z_{trn}} || p(H)\,,\;\; p(H|Z_{trn}) ||_{\mathcal{T}}$$

This is a stability constraint. Hence, *uniform generalization* is **equivalent** to *algorithmic stability*.

Information

We write:

$$\mathcal{J}(Z_{trn};\, H) = \mathbb{E}_H ||p(Z_{trn})\,,\;\; p(Z_{trn}|H)||_{\mathcal{T}}$$

This is an information leakage constraint. In fact, we also have:

$$\mathcal{J}(Z_{trn};\, H) \leq \sqrt{\frac{I(S_m; H)}{2m}}$$

This is the setting recently considered by Russo and Zou (2016) for controlling the bias of estimators in the adaptive setting.

Domain

If $\mathcal{Z}$ is a countable space, then:

$$\mathcal{I}(Z_{trn};\, H) \leq \sqrt{\frac{\mathbf{Ess}[\mathcal{Z}] - 1}{2\pi m}},$$

where:

$$\mathbf{Ess}\left[\mathcal{Z};\, p(z)\right] \doteq 1 + \Big(\sum_{z \in \mathcal{Z}} \sqrt{p(z)\left(1 - p(z)\right)}\Big)^2 \leq |\mathcal{Z}|$$

is a measure of the *effective* size of $\mathcal{Z}$.

Size of the Hypothesis Space

It can be shown that:

$$\mathcal{J}(Z_{trn};\, H) \leq \sqrt{\frac{\mathbf{H}(H)}{2m}} \leq \sqrt{\frac{\log|\mathcal{H}|}{2m}},$$

where $\mathbf{H}(H)$ is the Shannon entropy.

VC Dimension

A finite VC dimension does not imply uniform generalization; one can encode the sample in H.

Hence, we need to define a VC dimension in an information-theoretic manner.

Definition (Induced Concept Class)

The concept class $\mathcal{C}$ induced by a learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ is defined to be the set of total Boolean functions $c(z) = \mathbb{I}\{p(Z_{trn} = z \,|\, H) \geq p(Z_{trn} = z)\}$ for all $H \in \mathcal{H}$.

We have:

$$\mathcal{J}(Z_{trn};\, H) \leq \frac{4 + \sqrt{d_{VC}(\mathcal{C})\,(1 + \log(2m))}}{\sqrt{2m}}$$

Post Processing

Post-processing, e.g. sparsification and pruning, improves the uniform generalization rate.

That is:

$$S \to H_1 \to H_2 \qquad \Rightarrow \qquad \mathcal{J}(Z_{trn};\, H_1) \geq \mathcal{J}(Z_{trn};\, H_2)$$

Composition

Learning more infomration cannot improve the uniform generalization risk.

That is, for **composition** (adaptive or non-adaptive), we have:

$$\mathcal{I}(Z_{trn}; (H_1, H_2)) \geq \mathcal{I}(Z_{trn}; H_1)$$

Randomization and Privacy

Randomization improves the uniform generalization risk.

In particular, an (ϵ, δ) differentially private learning algorithm satisfies:

$$\mathcal{I}(Z_{trn}; H) \leq \frac{e^{\epsilon} - 1 + \delta}{2}$$

Sample Compression

A sample compression scheme of size k satisfies:

$$\mathcal{J}(Z_{trn};\, H) \leq O\Big(\frac{k}{m} + \sqrt{\frac{k}{m}}\Big)$$

Regularized ERM

If the hypothesis is learned using:

$$H = \arg\min_{h \in \mathcal{H}} \left\{ \frac{\lambda}{2} ||h||_2^2 + \frac{1}{m} \sum_{Z_i \in S_m} L(Z_i; h) \right\}$$

for some convex, twice differentiable loss $L(\cdot; h)$, then:

$$\mathcal{J}(Z_{trn};\, H) \leq \sqrt{\frac{d}{2m}} + o\big(m^{-\frac{1}{2}}\big),$$

which is valid when $m \gg \max\left\{d, \frac{1}{\gamma^2}\right\}$.

Composition: The General Rule

Writing:

$$\mathcal{J}(A;\, B \,|\, C) \doteq \mathbb{E}_C \, ||p(A, B \,|\, C)\,,\;\; p(A|C) \cdot p(B|C)||_{\mathcal{T}}$$

for the **conditional** variational information (analogous to conditional mutual information).

Theorem

We have:

$$\mathcal{J}(Z;\, (H_1, ..., H_k)) \leq \sum_{t=1}^{k} \mathcal{J}(Z;\, H_t \,|\, (H_1, ..., H_{t-1}))$$

Robustness

A finite amount of information (in bits) cannot alter the uniform generalization property significantly:

$$\mathcal{J}(Z_{trn};\,(H,K)) \leq (2+\frac{|\mathcal{K}|}{2})\cdot\mathcal{J}(Z_{trn};\,H)+\sqrt{\frac{\log|\mathcal{K}|}{2m}}$$

Concentration

Theorem

We have the concentration bound:

$$p\Big\{\big|R_{emp}(H;S_m) - R_{true}(H)\big| \geq t\Big\} \leq \frac{7}{2t}\Big[\mathcal{J}(Z_{trn};\, H) + \sqrt{\frac{\log 3}{49m}}\Big],$$

The concentration bound is tight because there exists learning algorithms that satisfy:

$$p\Big\{\big|R_{emp}(H;S_m) - R_{true}(H)\big| = t\Big\} = \frac{\mathcal{J}(Z_{trn};\, H)}{t}$$

Information-Theoretic Route

Theorem

1. *A generalization in expectation does not imply concentration.*
2. *A uniform generalization in expectation implies concentration.*

This gives an information-theoretic route from generalization in expectation to generalization in probability.

Q/A