Reasoning about Generalization via Conditional Mutual Information

Google Brain

Thomas Steinke*
Google Brain

Lydia Zakynthinou*
Northeastern University

Generalization in Machine Learning

- We sample an i.i.d dataset Z of size n from an unknown distribution D over $\mathcal{Z}$: $Z \leftarrow D^n$.
- Loss function ℓ indicates the quality of a model w.

Empirical loss	True loss
$\ell(w,Z)$	$\ell(w,D)$
$=\frac{1}{n}\sum_{i=1}^{n}\ell(w,Z_i)$	$=\mathbb{E}_{z\leftarrow D}[\ell(w,z)]$

- How can we ensure: $\ell(A(Z),Z) \approx \ell(A(Z),D)$

Methods used to prove generalization

- Uniform Convergence: VC dimension [VC71]
- Distributional Stability: Differential Privacy [DMNS06]
- Uniform Stability [BE02]
- Local Statistical Stability [LS19]
- **Mutual Information Methods**
  - ✓Low MI ⇒ Generalization [RZ16/RX17]
  - ✓ ε-DP ⇒ $I(A(Z);Z) \le \frac{1}{2}\varepsilon^2 n$
  - ✓ $I(A(Z);Z) \le \log|\mathcal{W}|$
  - ! (ε,δ)-DP $\not\Rightarrow$ bound on MI [D12/MMPRTV10]
  - ! For 1-D thresholds, *any consistent learner* has $I(A(Z);Z)=\infty$ for some D. [BMNSY18]

Conditional Mutual Information (CMI)

- Draw $2n$ samples $\tilde{Z} \leftarrow D^{2n}$.
- Selector function $S \in \{0,1\}^n$ uniformly random defines partition of real samples $\tilde{Z}_S$ and "ghost" samples $\tilde{Z}_{\bar{S}}$.

$$\tilde{Z}_S = (\tilde{Z}_{1,S_1},\dots,\tilde{Z}_{n,S_n})$$

$\tilde{Z}_{1,0}$	$\tilde{Z}_{1,1}$
$\tilde{Z}_{2,0}$	$\tilde{Z}_{2,1}$
$\vdots$	$\vdots$
$\tilde{Z}_{n,0}$	$\tilde{Z}_{n,1}$

- Run $A(\tilde{Z}_S)$.
- CMI of A with respect to D:

$$I(A(\tilde{Z}_S);S|\,\tilde{Z})$$

✓Post-processing and (non-adaptive) composition
✓CMI is finite ("normalized" MI)

Bounding CMI

- $\sqrt{\epsilon}$-Differential Privacy
- ϵ-Mutual Information Stability
- ϵ-KL Stability
- ϵ-Average-Leave-One-Out KL Stability

(all of the above) $CMI_D(A) \le \epsilon n$

- δ-TV Stability (i.e., $(0,\delta)$-DP)⇒ $CMI_D(A) \le \delta n$
- Compression schemes of size k ⇒ $CMI_D(A) \le k\log(n)$
- Hypothesis class with VC dimension d ⇒ ∃ ERM such that $CMI_D(A) \le d\log(n)$

Low CMI implies Generalization

For loss in [0,1].

- $|\mathbb{E}[\ell(A(Z),Z)-\ell(A(Z),D)]| \le \sqrt{\frac{2\cdot CMI_D(A)}{n}}$
- $\mathbb{E}\left[\left(\ell(A(Z),Z)-\ell(A(Z),D)\right)^2\right] \le \frac{3\cdot CMI_D(A)+2}{n}$
- $\mathbb{E}[\ell(A(Z),D)] \le 2\cdot\mathbb{E}[\ell(A(Z),Z)] + \frac{3\cdot CMI_D(A)}{n}$

Extension to "unbounded" loss (e.g. hinge loss) and to non-linear losses (e.g. Area Under the ROC Curve).

Conclusion

CMI is a new framework for reasoning about generalization, which:

- Unifies existing frameworks
- Provides a variety of forms of generalization guarantees.

Uniform Convergence
Compression Schemes
Distributional Stability
Differential privacy
KL/ALKL/TV-stability
⇒ Low $CMI(A)$ ⇒ Generalization bounds

* Work done while at IBM Research Almaden