23. Information and Generalization

Recap

- **Last class.** Generalization bounds via algorithmic stability
    - Generalize, if the algorithmic output (predictor) is stable w.r.t. dataset change

$$Z^n = (Z_1, Z_2, \ldots, Z_n) \longrightarrow \boxed{A} \longrightarrow \hat{f}$$

- Similar to continuity arguments

$$\mathrm{dist}(Z^n, \tilde{Z}^n) \leq \delta \quad \rightarrow \quad \mathrm{dist}(A(Z^n), A(\tilde{Z}^n)) \leq \epsilon$$

- Input. Hamming distance of the dataset
- Output. Uniform norm

- **This class.** Generalization bounds via information-theoretic arguments
    - If very little "information" about $(Z_1, Z_2, \ldots, Z_n)$ has been used for the determination of $\hat{f}$, then the model cannot overfit to the training data!

Information measures

- First, let us briefly review the measures of information
  - Consider a (discrete) random variable $X \sim P_X$.

Definition (Entropy).

The entropy of the random variable is

$$H(X) = H(P_X) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{P_X(x)} = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right]$$

- The amount of "uncertainty" in a random variable
  - Agnostic to re-labeling — nothing about x there!
- **Operational.** Expected length of the optimal binary code to store the outcomes of $X_1, X_2, \ldots,$
  - e.g., Huffman code

Information measures

- Analogous quantity can be defined for continuous random variables

Definition (Differential Entropy).

The differential entropy of the continuous random variable $X \sim p$ is

$$h(X) = h(p_X) = \mathbb{E}\left[\log \frac{1}{p_X(x)}\right] = \int_x \log \frac{1}{p_X(x)} \mathrm{d}x$$

- **Note.** Not generally invariant under relabeling
  - Consider $Y = 2X$

Information measures

Definition (Relative entropy; Kullback-Leibler divergence).

The KL divergence of the distribution P from Q is

$$D(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P(X)}{Q(X)}\right] = \mathbb{E}_{X\sim P}\left[\log \frac{1}{Q(X)}\right] - H(P)$$

- The penalty of using the codebook for $Q(\,\cdot\,)$ for the samples distributed as $P(\,\cdot\,)$
- **Properties.**
    - Nonnegative, with zero when $P = Q$
    - Asymmetric
    - Requires $P \ll Q$
        - i.e., $P(x) = 0$ for all x such that $Q(x) = 0$

Information measures

Definition (Mutual information).

The mutual information is

$$I(X;Y) = D(P_{XY} \| P_X P_Y)$$

- The "distance" of dependent signal from the independent ones
- For discrete variables, we have:

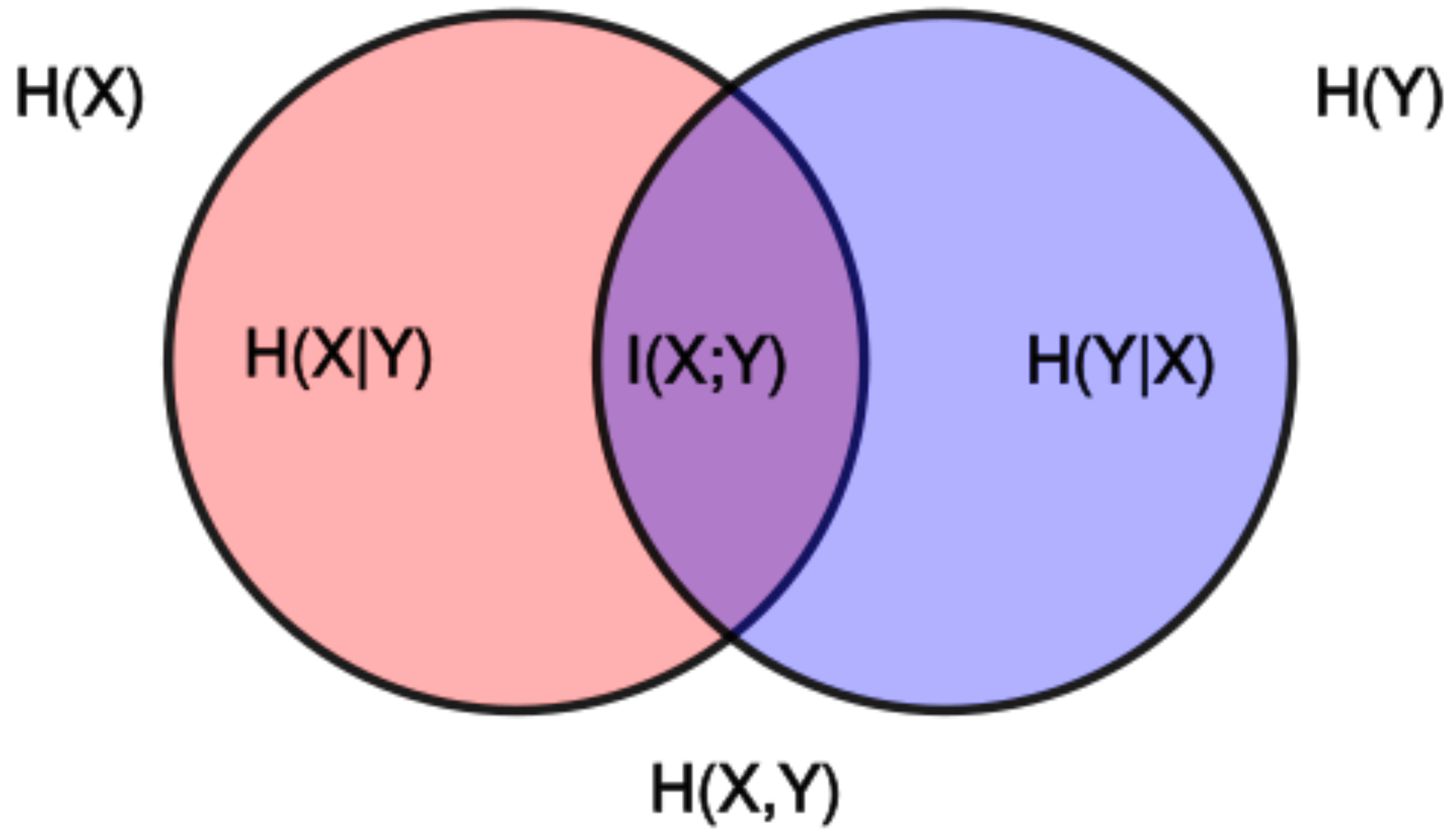
$$I(X;Y) = H(X) - H(X \mid Y)$$

where $H(X \mid Y) = \mathbb{E}_{y \sim P_Y}[H(P_{X|Y=y})]$

  - The uncertainty of X that can be reduced by knowing Y

- We also have:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$



MI-based Generalization bound

- Now, we can prove the following result.

Theorem 1 (Xu-Raginsky).

Suppose that the loss is bounded, i.e., $\ell(f, z) \in [0,1]$. Then, we have

$$|\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})]| \leq \sqrt{\frac{I(\hat{f}; Z^n)}{2n}}$$

- Exactly what we desired
  - If we use little information in Z^n to determine $\hat{f}$, never overfits
  - If $\hat{f}$ belongs to a small set, we'll have a small $H(\hat{f})$
  - Similar n-dependence, unless $I(\hat{f}; Z^n)$ grows too large

Tool: Donsker-Varadhan

- For the proof, we'll need a tool

Theorem (Donsker-Varadhan).

Let P and Q be two probability distributions on a common measurable space $\mathscr{X}$, such that $P \ll Q$. Then, for every $\varphi : \mathscr{X} \to \mathbb{R}$ such that $\mathbb{E}_Q[\exp(\varphi(X))] < \infty$, we have

$$D(P\|Q) \geq \mathbb{E}_P[\varphi(X)] - \log \mathbb{E}_Q[\exp(\varphi(X))]$$

- **Note.** Equality holds when we take supremum on the RHS w.r.t. φ
- **Proof idea.** Consider an exponentially-tilted version of Q, i.e.,

$$\tilde{Q} = \frac{\exp(\varphi(X))}{\mathbb{E}_Q[\exp(\varphi(X))]} Q$$

  - Then, we have

$$D(P\|Q) = D(P\|\tilde{Q}) + \mathbb{E}_P[\log(\tilde{Q}/Q)] \geq \mathbb{E}_P[\log(\tilde{Q}/Q)]$$

  - Evaluate the RHS and we get what we want

Proof sketch

$$D(P\|Q) \geq \mathbb{E}_P[\varphi(X)] - \log \mathbb{E}_Q[\exp(\varphi(X))]$$

- Consider the choice
  - $P = P_{\hat{f}Z^n}$
  - $Q = P_{\hat{f}} P_{Z^n}$
  - $\varphi(\hat{f}, Z^n) = \lambda(R(\hat{f}) - \hat{R}(\hat{f}))$,
- Then, by applying Donsker-Varadhan, we get:

$$I(\hat{f}; Z^n) \geq \lambda \mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] - \log \mathbb{E}_{P_{\hat{f}}} \mathbb{E}_{P_{Z^n}}[\exp(\lambda(R(\hat{f}) - \hat{R}(\hat{f}))]$$

- Now, as the $\ell(\cdot)$ is bounded in $[0,1]$, the Hoeffding's lemma implies that

$$\mathbb{E}_{P_{Z^n}}[\exp(\lambda(R(\hat{f}) - \hat{R}(\hat{f}))] \leq \exp(\lambda^2/8n)$$

- Thus, we have:

$$I(\hat{f}; Z^n) \geq \sup_{\lambda} \left(\lambda \mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] - \lambda^2/8n \right)$$

Bounding MI: Finite

- Now the question is: How do we bound the value $I(\hat{f}; Z^n)$?

- Suppose that we have a finite hypothesis space $\mathcal{F} = \{f_1, \ldots, f_k\}$
  - Then, we know that:

$$I(\hat{f}; Z^n) \leq H(\hat{f}) \leq \log |\mathcal{F}| = \log k$$

  - Thus, we get the bound:

$$\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] \leq \sqrt{\frac{\log k}{2n}}$$

- This result is similar to what we get via union bounds or finite class lemma

Bounding MI: Uncountable

- When $\mathscr{F}$ is uncountable, we can we do?
- **Naïve.** Do the same: $I(X;Y) = h(X) - h(X \mid Y)$, and upper bound $h(X)$?
    - Sadly, for continuous r.v., $h(X \mid Y)$ can be negative…
- We can think about the finite ϵ-covering $\mathscr{G}$

$$Z^n \to \hat{f} \to \hat{g} = \arg\min_{g \in \mathscr{G}} \|g - \hat{f}\|$$

- Thus, we get the bound:

$$\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] \leq \mathbb{E}[R(\hat{g}) - \hat{R}(\hat{g})] + \mathbb{E}[R(\hat{f}) - R(\hat{g})] + \mathbb{E}[\hat{R}(\hat{g}) - \hat{R}(\hat{f})]$$

$$\leq \sqrt{\frac{\log |\mathscr{N}(\mathscr{F}, \|\cdot\|, \epsilon)|}{2n}} + 2\epsilon$$

- Optimizing over ϵ, we typically get a bound dependent on $\sqrt{(d \log dn)/n}$

Bounding MI: Binary Classifiers

- Consider the case of binary classifiers
    - Given the samples Z^n, divide them into two subsets: D_1, D_2, with N_1 and N_2 samples.
    - Select a finite subset of $\mathscr{U} \subseteq \mathscr{F}$ such that

$$\{(f(Z_1), \ldots, f(Z_{N_1})) \mid f \in \mathscr{U}\} = \{(f(Z_1), \ldots, f(Z_{N_1})) \mid f \in \mathscr{F}\}$$

    - Then, select the risk minimizer on D_2

$$\hat{f} = \arg\min_{f \in \mathscr{U}} \hat{R}_{D_2}(f)$$

    - Then, we know that

$$\mathbb{E}[R(\hat{f}) - \hat{R}_{D_2}(\hat{f})] \leq \sqrt{\frac{\log |\mathscr{U}|}{N_2}}$$

- An upper bound on $\log |\mathscr{U}|$ is what we call the VC-dimension

Bounding MI: Randomized Algorithm

- MI-based generalization bounds can handle several cases which Rademacher analysis cannot

- Consider a randomized learning algorithm, where:
  - Hypothesis space is all binary classifiers on [0,1]:

$$\mathcal{F} = \{f : [0,1] \to \{-1, +1\}\}$$

  - Given N samples, we randomly select one data (x_i, y_i) and select

$$\hat{f}(x) = y_i$$

- Rademacher complexity: infinity
- mutual information:

$$I(\hat{f}; Z^n) \leq 1$$

Remarks

- There are many more cases where MI-based bounds are strictly better
  - Noise-adding algorithms (e.g., SGLD)
  - Adaptive data analysis (e.g., early stopping)
  - Gibbs posterior

- We did not cover conditional mutual information bounds
  - but you should check out ;)

Further avenues

- CMI bounds: https://proceedings.mlr.press/v125/steinke20a/steinke20a.pdf
- A survey: https://arxiv.org/abs/2309.04381