# Minimum Description Length and Generalization Guarantees for Representation Learning

Milad Sefidgaran, Abdellatif Zaidi, Piotr Krasnowski

Neurips 2023

# General problem setup

- Data $Z = (X, Y) \in \mathcal{Z}$ distributed according to $\mu$, where $Y \in \{1, \ldots, K\}$ is the label
- Training dataset $S = \{Z_1, \ldots, Z_n\} \sim \mu^{\otimes n}$
- Randomized algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{W}$
- Model $w$ for every $x$ makes the prediction $\hat{Y} \sim P_{\hat{Y}|X=x, W=w}$
- Loss function $\ell(z, w) = \mathbb{E}_{\hat{Y} \sim P_{\hat{Y}|X,W}(\hat{Y}|x,w)}\left[\mathbb{1}_{\{y \neq \hat{Y}\}}\right]$
- Empirical risk: $\hat{\mathcal{L}}(s, w) := \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, w)$ and Population risk: $\mathcal{L}(w) := \mathbb{E}_{Z\sim\mu}[\ell(Z, w)]$

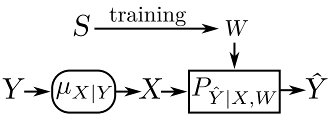
The goal is to study **generalization error**:

$$\mathrm{gen}(S, W) := \mathcal{L}(W) - \hat{\mathcal{L}}(S, W)$$

# Overview of results

- One-step prediction model:
  - a new notion of minimum description length (MDL) of predicted labels
  - Generalization bound:

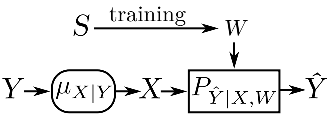
$$\sqrt{\frac{2 \times \mathrm{MDL}(\text{Predicted Labels})}{n}}$$



# Overview of results

- One-step prediction model:
  - a new notion of minimum description length (MDL) of predicted labels
  - Generalization bound:

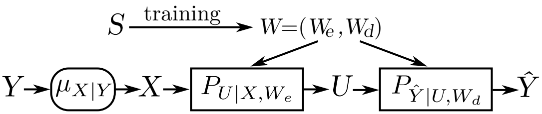
$$\sqrt{\frac{2 \times \mathrm{MDL}(\text{Predicted Labels})}{n}}$$




- Two-step prediction model:
  - a new notion of MDL of latent variables
  - Generalization bound:

$$2\sqrt{\frac{2 \times \mathrm{MDL}(\text{Latent Variables}) + K + 2}{n}}$$

  - Practical implications: suggests new symmetric data-dependent priors



# One-step prediction

- Approach. Extension of compressibility framework of Blum & Langford (2003) by considering:
  - **block-coding** or **information-theoretic** compression
  - **lossy** compression or **rate-distortion** analysis

# One-step prediction

- Approach. Extension of compressibility framework of Blum & Langford (2003) by considering:
  - **block-coding** or **information-theoretic** compression
  - **lossy** compression or **rate-distortion** analysis
- General idea. Consider a given training dataset $S$ and ghost dataset $S'$, that are **rearranged** in an **indistinguishable** manner as $\mathfrak{Z}^{2n}$.
  - If the set of rearranged predictions of $S$ and $S'$ can be "described" using few bits, then the algorithm generalizes well.
  - To "describe" the predictions, we use **source coding literature** in information theory and in particular the **information theoretic covering lemma**.

## One-step prediction

- Approach. Extension of compressibility framework of Blum & Langford (2003) by considering:
  - **block-coding** or **information-theoretic** compression
  - **lossy** compression or **rate-distortion** analysis
- General idea. Consider a given training dataset $S$ and ghost dataset $S'$, that are **rearranged** in an **indistinguishable** manner as $\mathfrak{Z}^{2n}$.
  - If the set of rearranged predictions of $S$ and $S'$ can be "described" using few bits, then the algorithm generalizes well.
  - To "describe" the predictions, we use **source coding literature** in information theory and in particular the **information theoretic covering lemma**.
  - This introduces a new notion of **MDL**:

$$D_{KL}\left(P^{\otimes 2n}_{\hat{Y}|X,W}(\hat{\mathbf{Y}},\hat{\mathbf{Y}}'|\mathbf{X},\mathbf{X}',W)\middle\|\mathbf{Q}\right),$$

for some appropriately "symmetric" prior $\mathbf{Q}$ over $\hat{Y}^{2n}$.

# Rearrangement strategies for one-step prediction model

- Type I symmetry. $(\mathfrak{Z}_i, \mathfrak{Z}_{i+n})$ is distributed uniformly over $\{(Z_i, Z_i'), (Z_i', Z_i)\}$.
  - We derive results similar to CMI (Steinke & Zakynthinou, 2020) and f-CMI (Harutyunyan et al., 2021) literature
  - Makes a connection between frameworks of Blum-Langford and CMI

# Rearrangement strategies for one-step prediction model

- Type I symmetry. $(\mathfrak{Z}_i, \mathfrak{Z}_{i+n})$ is distributed uniformly over $\{(Z_i, Z_i'), (Z_i', Z_i)\}$.
  - We derive results similar to CMI (Steinke & Zakynthinou, 2020) and f-CMI (Harutyunyan et al., 2021) literature
  - Makes a connection between frameworks of Blum-Langford and CMI
- Type II symmetry. $\mathfrak{Z}^{2n}$ is a a random permutation (reshuffle) of $(S, S')$.
  - new results in terms of the function

$$h_D(x, x') := 2h_b\left(\frac{x + x'}{2}\right) - h_b(x) - h_b(x'),$$

    which is two times Jensen-Shannon divergence between two binary Bernoulli distributions with parameters $x$ and $x'$.
  - The bounds are $\mathcal{O}(1/n)$ for the realizable setup.

# Rearrangement strategies for one-step prediction model

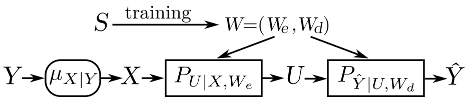
- Type I symmetry. $(\mathfrak{Z}_i, \mathfrak{Z}_{i+n})$ is distributed uniformly over $\{(Z_i, Z_i'), (Z_i', Z_i)\}$.
  - We derive results similar to CMI (Steinke & Zakynthinou, 2020) and f-CMI (Harutyunyan et al., 2021) literature
  - Makes a connection between frameworks of Blum-Langford and CMI
- Type II symmetry. $\mathfrak{Z}^{2n}$ is a a random permutation (reshuffle) of $(S, S')$.
  - new results in terms of the function

$$h_D(x, x') := 2h_b\left(\frac{x + x'}{2}\right) - h_b(x) - h_b(x'),$$

    which is two times Jensen-Shannon divergence between two binary Bernoulli distributions with parameters $x$ and $x'$.
  - The bounds are $\mathcal{O}(1/n)$ for the realizable setup.
- Lossy compressibility

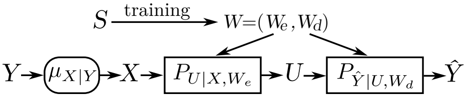
# Two-step prediction model

- Suitable for optimization:
  - **Encoder:** guarantees the good generalizability by extracting "good" representations,
  - **Decoder:** minimizes the empirical risk.



# Two-step prediction model

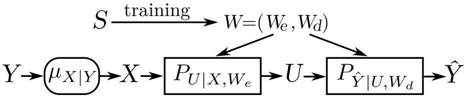
- Suitable for optimization:
  - **Encoder:** guarantees the good generalizability by extracting "good" representations,
  - **Decoder:** minimizes the empirical risk.




- Information bottleneck principle: $I(U;Y) - \beta I(U;X)$
  - $I(U;X)$ is perceived to capture MDL and hence the generalization performance,
  - $I(U;Y)$ captures the "relevance" for prediction and hence the empirical risk performance.

# Two-step prediction model

- Suitable for optimization:
  - **Encoder:** guarantees the good generalizability by extracting "good" representations,
  - **Decoder:** minimizes the empirical risk.




- Information bottleneck principle: $I(U;Y) - \beta I(U;X)$
  - $I(U;X)$ is perceived to capture MDL and hence the generalization performance,
  - $I(U;Y)$ captures the "relevance" for prediction and hence the empirical risk performance.
- Information bottleneck critics:
  - no non-vacuous theoretical guarantees,
  - Experimental evidence shows dependence of the generalization error on the so-called geometrical compression rather than $I(U;X)$,
  - Mutual information is invariant to bijection and does not reflect the "structure" or "simplicity" of the encoder/decoder.

# Main result

**Generalization Bound for Representation Learning Algorithms**

$$\mathbb{E}_{S,W}[\mathrm{gen}(S,W)] \leq 2\sqrt{\frac{2\mathbb{E}_{S,S',W_e}\left[D_{KL}\left(P^{\otimes 2n}_{U|X,W_e}(\mathbf{U},\mathbf{U}'|\mathbf{X},\mathbf{X}',W_e)\middle\|\mathbf{Q}\right)\right]+K+2}{n}},$$

where $\mathbf{Q}$ is a type-III symmetric prior.

- The bound only depends on the encoder and complexity of the latent variables.
- While the mutual information captures the information leakage, the above KL-divergence captures the encoder structure.
- The lossy version explains the geometrical compression.

# Experimental implications

- In Variational IB, the prior is fixed, e.g. $\mathcal{N}(0_m, I_m)$.
- In contrast, inspired by our results, we introduce new symmetric priors. These priors
    - are data-dependent,
    - are "learned" along the iterations,
    - can be applied in "lossless" and "lossy" manner.