Definition 1. *A foundation model is trained beyond task-relevant data, and these data are the maximum that research organizations can collect.*

Definition 2. *It is defined as a general model based on a large-scale predictive structure that can directly handle diverse tasks through an in-context learning mechanism without retraining for each task.*

Definition 3. *[2] (The original version of this definition uses category theory, and this version is the natural language version),* ***Ideal Foundation Model (IFM)****. For a specific task-defined data category, if a foundation model can map all data to a unified feature space, and there exists a matching rule independent of the training data that makes the model effective for the task, such that any relationship in the feature space yields results completely equivalent to the correct inherent relationship between samples in the original task, then the model is called an Ideal Foundation Model.*

Example: *In an ideal face recognition model, if the model extracts the vectors of two faces and computes their similarity through a fixed formula, and this similarity is completely equivalent to the real judgment of whether these two faces belong to the same person, then the model meets the requirements of an Ideal Foundation Model.*

Definition 4. *[1]* ***Bayes-optimal Foundation Model****. Assume the context conforms $P(H_t|\theta)$ where $\theta = \{\psi, \theta_{1:M}\}$, given multiple documents $D_{1:M} = \{D_m \mid D_m \sim P(D_m|\theta_m)\}$ and their contexts $H_t^{(m)} = X_{1:t}^{(m)}$ for $m \in [M]$, the model's prediction $\hat{P}(X_{t+1}^{(m)}) = P(X_{t+1}^{(m)} \mid H_t^{(m)}, D_{1:m-1}, \pi)$ has Bayes error:*

$$L_{T,M,\pi} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbb{E}_{X_{t+1}^{(m)} \sim P(\cdot|H_t^{(m)}, D_{1:m-1}, \theta)} \left[-\ln \hat{P}(X_{t+1}^{(m)}) \right] \tag{1}$$

The optimal Bayes error is:

$$L_{T,M} = \min_{\pi} L_{T,M,\pi} \tag{2}$$

A model achieving this error is called a Bayes-optimal Foundation Model.

Lemma 1. *[1] The optimal Bayes error upper bound decomposes into three terms: 1. Irreducible term $\mathbb{H}(D_{M+1} \mid \theta_{M+1})$ (conditional entropy of current document) 2. History term $\mathbb{I}(H_M; \psi)$ (mutual information between historical context and meta-parameters) 3. Context*

estimation term $\mathbb{I}(D_{M+1}; \theta_{M+1} \mid \psi)$ *(mutual information between current document and task parameters given meta-parameters)*

Where $H_m, t = \{D_{1:m-1}, X_{1:t}^{(m)}\}$*. For a new document and context, the Bayes error satisfies:*

$$L_{M+1} \leq \frac{\mathbb{H}(D_{M+1} \mid \theta_{M+1})}{r} + \frac{\mathbb{I}(H_M; \psi)}{Mr} + \frac{\mathbb{I}(D_{M+1}; \theta_{M+1} \mid \psi)}{r} \tag{3}$$

for some $r \leq T$*.*

This method can proves that linear representation learning with uniformly distributed parameters achieves the optimal performance upper bound for classical tasks [1], deriving the sample complexity when the learning algorithm is perfect (yielding the Bayes-optimal model).

References

[1] Hong Jun Jeon et al. "An Information-Theoretic Analysis of In-Context Learning". In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=NQn2tYLv5I.

[2] Yang Yuan. "On the Power of Foundation Models". In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 40519–40530. URL: https://proceedings.mlr.press/v202/yuan23b.html.