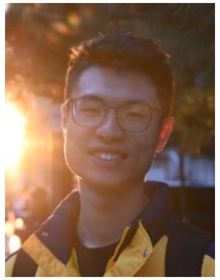
A Theory for Conditional Generative Modeling on Multiple Data Sources

Rongzhen Wang[1], Yan Zhang[2], Chenyu Zheng[1], Chongxuan Li[1,*], Guoqiang Wu[2,*]

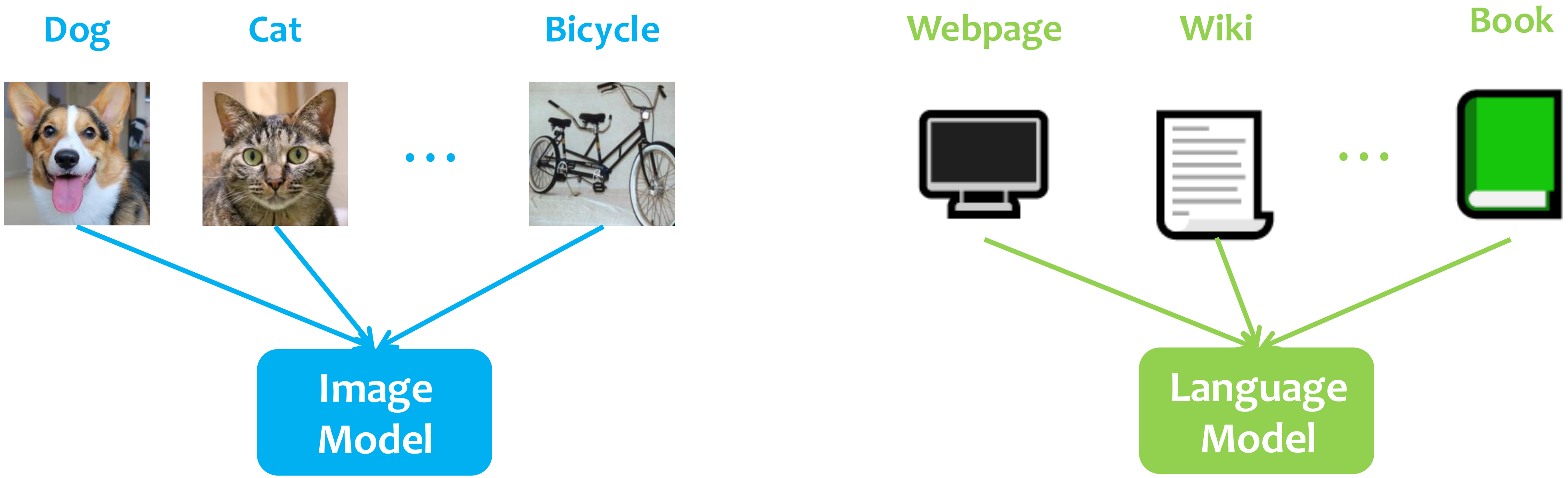




[1]Renmin University of China, [2]Shandong University, [*]Corresponding author

Motivation

- Large foundation models are trained on various data sources



*Is it more effective to train **separate models** on **individual data sources**, or to train a **single model** using data from **multiple sources**?*

Problem Formulation

- Data: X, source label: Y, number of sources: K
- Conditional distributions: $X|k \sim p^*_{X|k} = p_{\phi^*_k,\psi^*}$ for $k = 1, 2, \ldots, K$

$$p^*_{X|Y}(\boldsymbol{x}|y) = \prod_{k=1}^{K}\Big(p_{\phi^*_k,\psi^*}(\boldsymbol{x}|k)\Big)^{\mathbb{I}(y=k)}, p^*_{X,Y}(\boldsymbol{x},y) = p^*_{X|Y}(\boldsymbol{x}|y)p^*_Y(y)$$

- Dataset: $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ i.i.d. sampled from $p^*_{X,Y}$
- Average TV error: $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) := \mathbb{E}_Y\left[\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y})\right]$, where $\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y}) = \frac{1}{2}\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}$

Problem Formulation

- Maximum likelihood estimation (MLE)
  - Multi-source training:

$$\{\hat{\phi}_k^{\text{multi}}\}, \hat{\psi}^{\text{multi}} = \underset{\phi_k \in \Phi, \psi \in \Psi}{\arg\max} \prod_{i=1}^{n} \prod_{k=1}^{K} \left(p_{\phi_k, \psi}(\boldsymbol{x}_i | k)\right)^{\mathbb{I}(y_i = k)} \quad \Rightarrow \hat{p}_{X|Y}^{\text{multi}}$$

  - Single-source training: let $S_k = \{\boldsymbol{x}_j^k, k\}_{j=1}^{n_k} = \{(\boldsymbol{x}_i, y_i) \in S | y_i = k\}$,

$$\{\hat{\phi}_k^{\text{single}}\}, \{\hat{\psi}_k^{\text{single}}\} = \underset{\phi_k \in \Phi, \psi_k \in \Psi}{\arg\max} \prod_{j=1}^{n_k} p_{\phi_k, \psi_k}(\boldsymbol{x}_j^k | k) \quad \Rightarrow \hat{p}_{X|Y}^{\text{single}}$$

  - Realizable assumption: $\phi_k^* \in \Phi$ and $\psi^* \in \Psi$

Theoretical Results: General Error Bounds

- Multi-source training achieves **a lower error upper bound** than single-source training

Under such formulation, we have *(by **Theorem 3.2**)*:

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y}) = \mathcal{O}\left(\sqrt{\frac{1}{n}\log \underbrace{\mathcal{N}_{[]}\left(\frac{1}{n};\mathcal{P}^{\mathrm{multi}}_{X|Y}, L^1(\mathcal{X})\right)}_{}}\right),$$

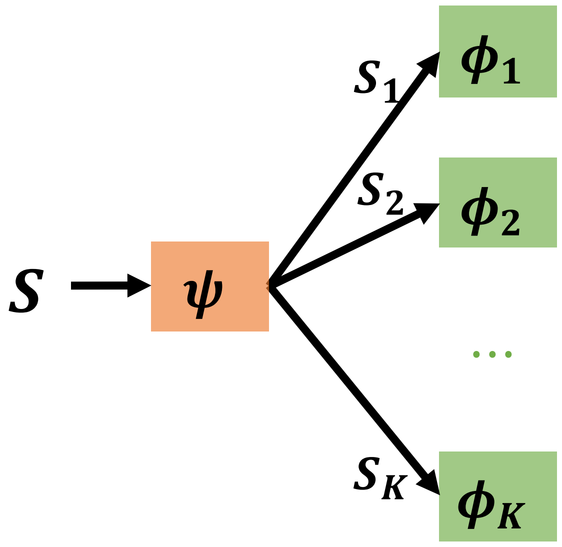
a complexity measure for conditional distribution space

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y}) = \mathcal{O}\left(\sqrt{\frac{1}{n}\log \underbrace{\mathcal{N}_{[]}\left(\frac{1}{n};\mathcal{P}^{\mathrm{single}}_{X|Y}, L^1(\mathcal{X})\right)}_{}}\right),$$

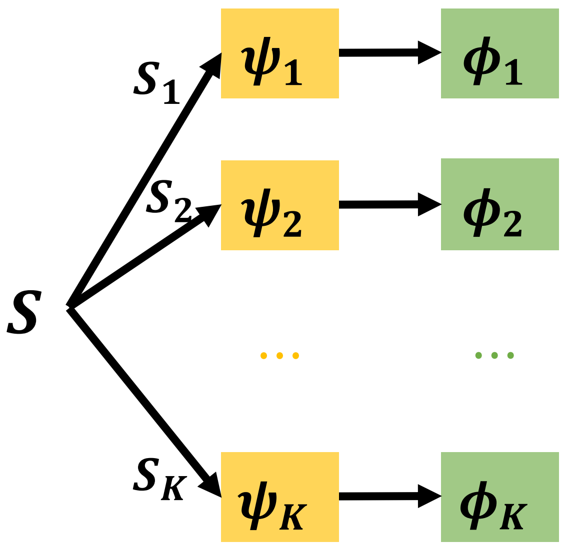
where *(by **Proposition 3.3**)*: $\mathcal{N}_{[]}\left(\epsilon;\mathcal{P}^{\mathrm{multi}}_{X|Y}, L^1(\mathcal{X})\right) \leq \mathcal{N}_{[]}\left(\epsilon;\mathcal{P}^{\mathrm{single}}_{X|Y}, L^1(\mathcal{X})\right)$.

Intuitive Illustration

- Theorem 3.2 resembles a generalization bound based on the model complexity
- Multi-source training reduces model complexity (or increases the sample size) by utilizing a shared parameter space Ψ to learn the common parameter φ



(a) Multi-source training.



(b) Single-source training.

Instantiations

- Parametric estimation
  - Conditional Gaussian distributions
- Deep generative models
  - Autoregressive model
  - Energy-based models

Theoretical Results: Instantiation of Gaussian Estimation

- Setup
  - $X|k \sim \mathcal{N}(\boldsymbol{\mu}_k^*, \boldsymbol{I}_d)$
  - The first d_1 dimensions of $\boldsymbol{\mu}_k^*$ are source-specific, while the remaining are shared:

$$\phi_k : \boldsymbol{\mu}_k^*[1:d_1], \quad \psi : \boldsymbol{\mu}_1^*[d_1+1:d] = \cdots = \boldsymbol{\mu}_K^*[d_1+1:d]$$

- Results

$$\mathcal{R}_{\mathrm{TV}}(\hat{p}_{X|Y}^{\mathrm{multi}}) = \tilde{\mathcal{O}}\left(\sqrt{\frac{(K-1)d_1+d}{n}}\right), \quad \mathcal{R}_{\mathrm{TV}}(\hat{p}_{X|Y}^{\mathrm{single}}) = \tilde{\mathcal{O}}\left(\sqrt{\frac{Kd}{n}}\right)$$

  - When $d_1 = 0$ (source distributions are exactly identical), $\mathcal{R}_{\mathrm{TV}}(\hat{p}_{X|Y}^{\mathrm{multi}}) = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$
  - When $d_1 = d$ (source distributions are entirely distinct), $\mathcal{R}_{\mathrm{TV}}(\hat{p}_{X|Y}^{\mathrm{multi}}) = \tilde{\mathcal{O}}\left(\sqrt{\frac{Kd}{n}}\right)$

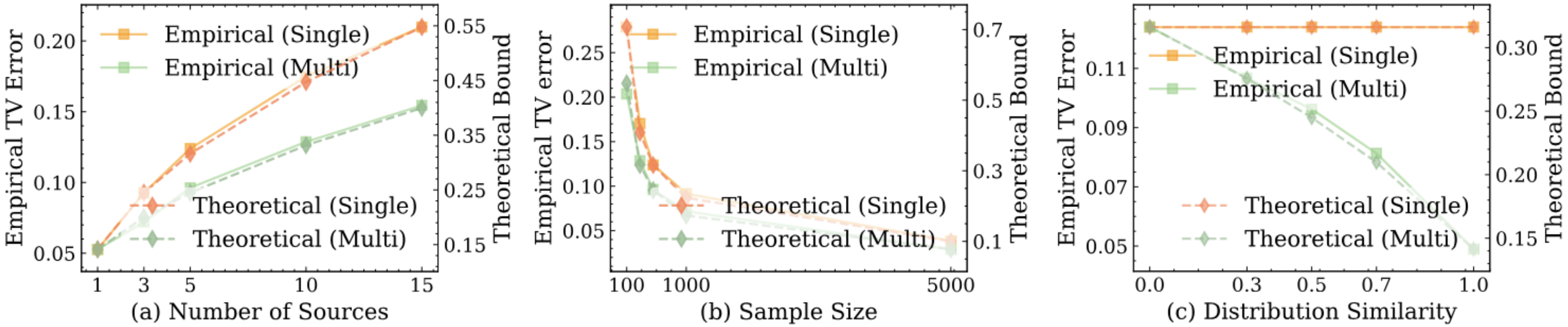
Experiments: Gaussian Estimation

- Analytical MLE solutions:

$$\hat{\phi}_k^{\text{multi}} = \frac{\sum_{y=1}^{n_k} \boldsymbol{x}_i^k[1:d_1]}{n_k}, \quad \hat{\psi}^{\text{multi}} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i[d_1+1:d]}{n}$$

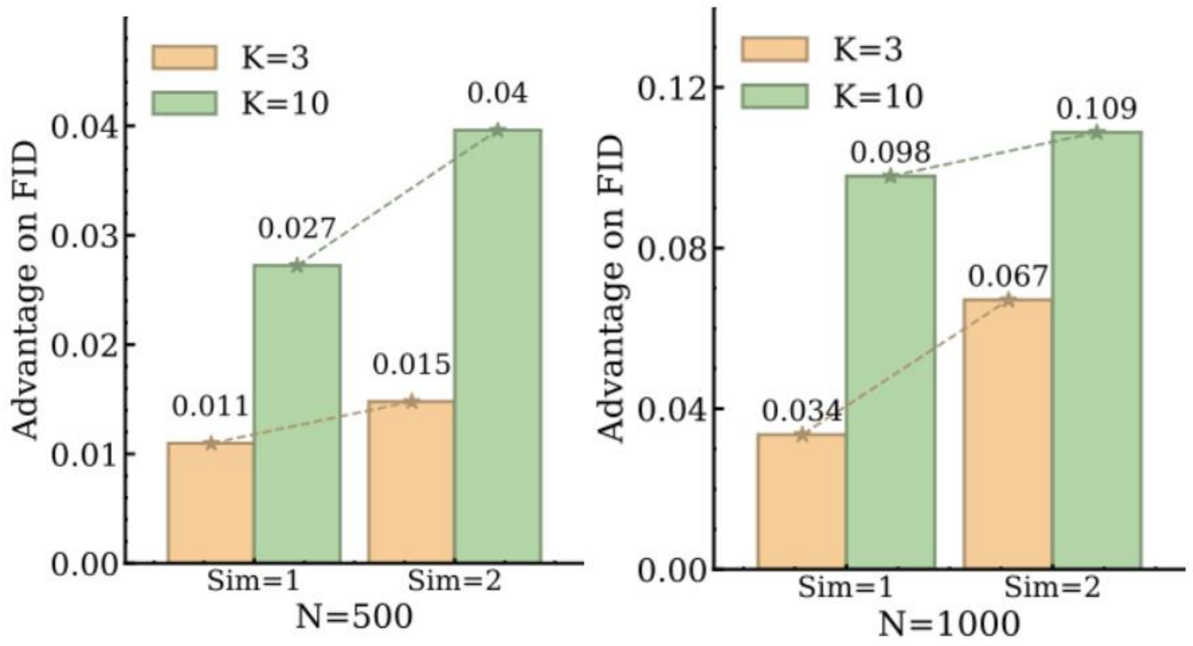
$$\hat{\phi}_k^{\text{single}} = \frac{\sum_{y=1}^{n_k} \boldsymbol{x}_i^k[1:d_1]}{n_k}, \quad \hat{\psi}_k^{\text{single}} = \frac{\sum_{y=1}^{n_k} \boldsymbol{x}_i^k[d_1+1:d]}{n}$$

- Results:



Experiments: Real-World Diffusion Model

- Dataset: ILSVRC2012 training set (a subset of ImageNet)
- Model: EDM2 (Karras et al., 2024)
- Result：



Conclusion

*Is it more effective to train **separate models** on **individual data sources**,*
*or to train a **single model** using data from **multiple sources**?*

- We find that under certain conditions, multi-source training provides stronger error guarantees than single-source training.

conditional modeling, realizable assumption, MLE

classical tool for analyzing MLE based on distribution space complexity [1,2]

- Together with the simulation experiments, this helps us understand the advantage of multi-source training quite clearly in some cases (i.e., the Gaussian estimation).

[1] Ge, J., Tang, S., Fan, J., and Jin, C. On the provable advantage of unsupervised pretraining, 2024.
[2] Geer, S. A. Empirical Processes in M-estimation, 2000.

Thank you for listening!