Building Better Black-Box Model Classification

Merrick Ohata[1], Carey E. Priebe[1], Hayden Helm[2]

[1] Johns Hopkins University Department of Applied Mathematics and Statistics [2]Helivan

The *discriminative factorization* bounds risk and boosts signal

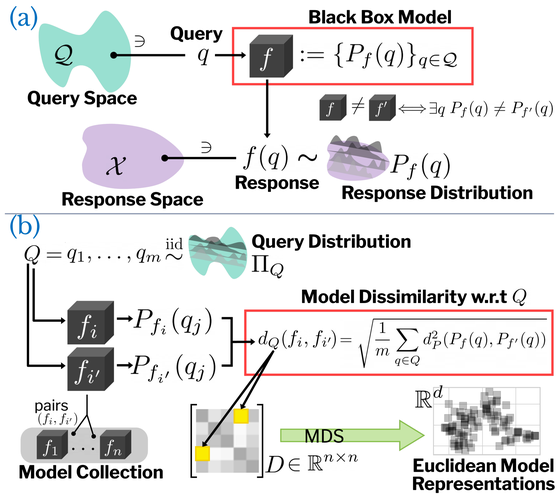

Figure 1: (a) the black box setting and (b) obtaining Euclidean model representations – each dot is a model

Introduction

- Most user-model interaction is black-box
- Can represent groups of models in $\mathbb{R}^d$ via *embedded* responses to a query set [2]
- Representations are consistent and efficient for model-level inference [1][3][4]
- Representation quality depends on task and query set choice
- "*signal*" queries probe relevant behavior (better), "*orthogonal*" queries do not (worse). "*uniform*" query set is all available queries.

The *discriminative factorization* framework

i. Formalizes signal vs orthogonal queries
ii. Yields bounds on classification risk
iii. Is estimable from the spectral structure of training data
iv. Improves classification efficiency without requiring task-specific knowledge

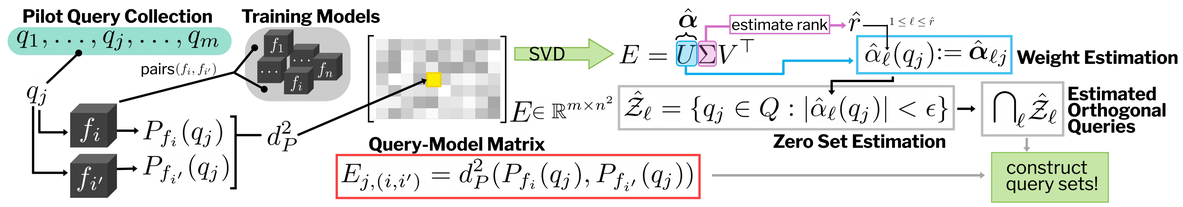

Figure 2: Obtaining the query-model matrix and estimating discriminative factorization values to select queries

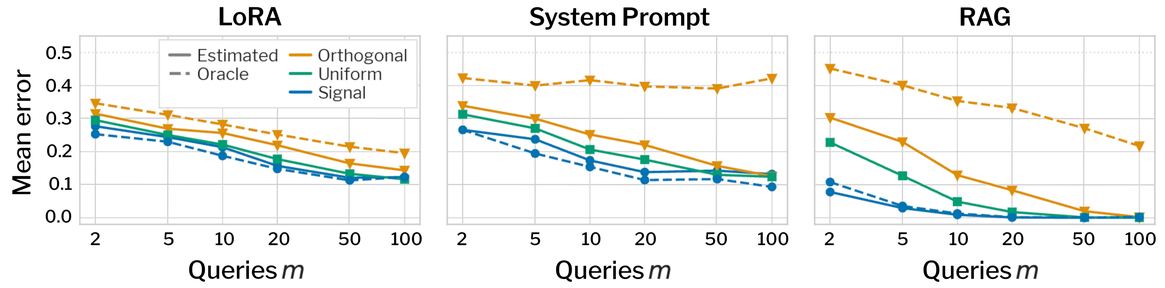

Figure 3: Mean classification error over 500 (LoRA) or 200 (System Prompt, RAG) trials with 80/20 training/test split on representations w.r.t. queries sampled uniformly.

Framework & Estimation

Rank r discriminative factorization:

$$d_P^2\left(P_f(q), P_{f'}(q)\right) = \sum_{\ell=1}^{r} \alpha_\ell(q)\ \phi_\ell(f, f')$$

for all $f, f' \in \mathcal{F}$ and $q \in \mathcal{Q}$ with maps
$\alpha: \mathcal{Q} \to [0,1]^r$ and $\phi: \mathcal{F}\times\mathcal{F} \to [0,\infty)^r$

$\alpha_\ell(q)$	*direction ℓ weight* for q
$\mathcal{Z}_\ell := \{q \in \mathcal{Q}: \alpha_\ell(q) = 0\}$	*zero set* of direction ℓ
$\rho_\ell = \Pi_Q(\mathcal{Z}_\ell)$	*zero set probability*
$\bigcap_\ell \mathcal{Z}_\ell$	*orthogonal queries*

- In practice, estimation as a one-time cost
- Use SVD of the *Query-Model Matrix* (Figure 2)
- Estimation results designate useful queries

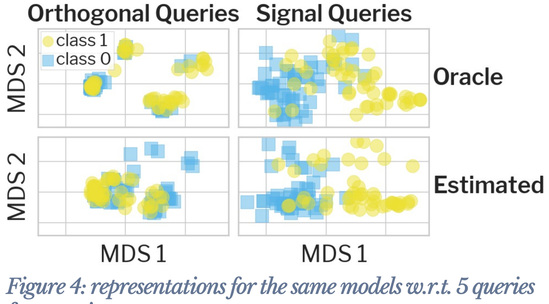

Figure 4: representations for the same models w.r.t. 5 queries from various sets

Classification Risk

- Information loss from using of embedded vs raw model responses
- With injective embedding, exponential bound on classification error: $\sum_{\ell=1}^{r} \rho_\ell^m + \gamma(n)$
- Non-injective case: worse-than-chance classification controlled by same bound
- $\gamma(n)$ from class-conditional distributions

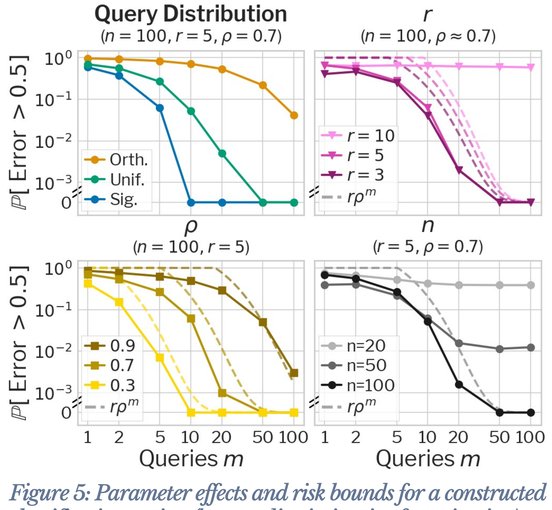

Figure 5: Parameter effects and risk bounds for a constructed classification setting (known discriminative factorization)

Experiments

Synthetic setting:
- Known discriminative factorization
- $\alpha_\ell(q) \sim$ Uniform[0,1] with parameter $\rho = \rho_\ell$
- $\theta_f \sim$ Uniform$\{0,1\}^r$: Class label = Parity(θ_f) and $\phi_\ell(f, f') = \mathbb{1}[\theta_{f\ell} \neq \theta_{f'\ell}]$

Real models:
- Binary classification on models augmented in 3 common ways with known signal and orthogonal queries
- Estimation of signal and orthogonal query sets via discriminative factorization

Discussion

- Estimated query selection mimics oracle selection in separation and mean error, improves classification efficiency
- Mean error always below 0.5; truly orthogonal queries may not exist
- Error bound holds for known factorization
- Theoretical results suggest black box model inference is always possible
- Further work needed to refine bounds, find "best" queries, extend to other tasks

Learn More

Scan Here

References

[1] Acharyya et al. Consistent estimation of generative model representations in the data kernel perspective space. arXiv:2409.17308, 2024.
[2] Helm et al. Tracking the perspectives of interacting language models. EMNLP 2024.
[3] Helm et al. Statistical inference on black-box generative models in the data kernel perspective space. ACL 2025.
[4] Helm et al. Query-efficient model evaluation using cached responses. arXiv preprint arXiv:2605.07096, 2026.

Acknowledgements

Aranyak Acharyya
Avanti Athreya
Brandon Duderstadt

Johns Hopkins Whiting School of Engineering

DARPA