Model Shapley: Equitable Model Valuation with Black-box Access

Xinyi Xu[1,2], Thanh Lam[1], Chuan Sheng Foo[2], Bryan Kian Hsiang Low[1]

[1]Department of Computer Science, National University of Singapore (NUS), Singapore

[2]Institute of Infocomm Research, A*STAR, Singapore

(Why not) Data Valuation

Practical challenges

- Operational difficulties:
  - **massively distributed** storage (e.g., over 1 million nodes),
  - **enormous data size** (e.g., over 45 TB of training data GPT-3),
  - **transient** nature (i.e., not persistently stored).

What then?

- **Data privacy regulations** (e.g., GRPR, CPA) prohibit direct access to data.
  Existing data valuations usually require an access to data.

Model Valuation

as a post-training valuation alternative

- The entire model is usually **stored in one piece** (i.e., not distributed storage).
- **Smaller** compared to the training data (e.g., size of GPT-3 $<$ 1% of its training data).
- Not transient (i.e., **persistently** stored).

- Models trained with **DP-ML methods can be available** for valuation.

- Existing AI marketplaces selling trained models require suitable pricing mechanisms, e.g., AWS Marketplace, Modzy.

Model Valuation

as a post-training valuation alternative

- The entire model is usually **stored in one piece** (i.e., not distributed storage).
- **Smaller** compared to the training data (e.g., size of GPT-3 $< 1\%$ of its training data).
- Not transient (i.e., **persistently** stored).

- Models trained with **DP-ML methods can be available** for valuation.

Data vs. Model Valuation

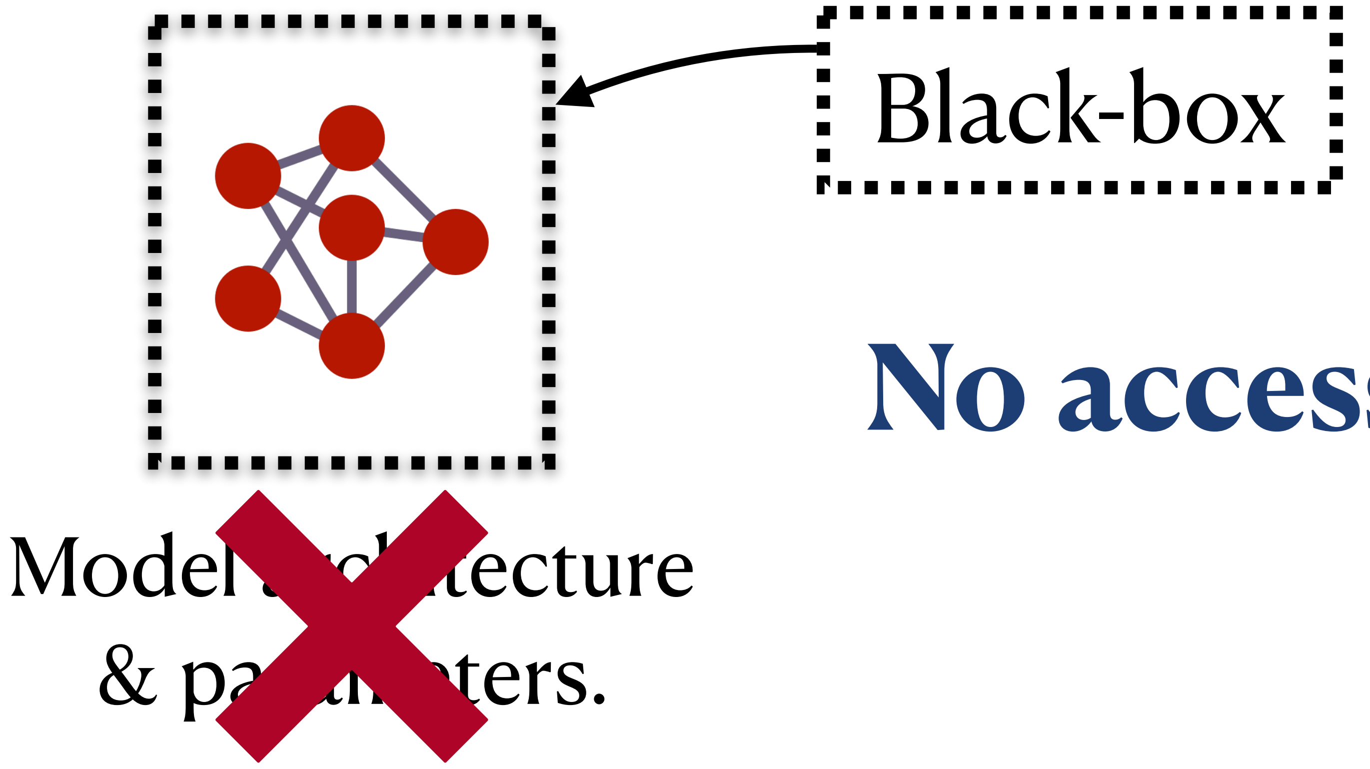
Model valuation is a more appealing choice

	Data valuation	Model valuation
Storage	Massively distributed	Stored in one piece
Size	Enormous	More manageable
Persistent storage	May not be	Yes
Privacy regulations	Cannot bypass	Less difficult

Model Valuation

Challenges

- Under the black-box access, what is a formal representation of the model?



No access to architecture, or parameters.

- What is a suitable valuation criterion?

Accuracy, predictive certainty, F1-score?

- How to ensure equitability (i.e., "fairness")?

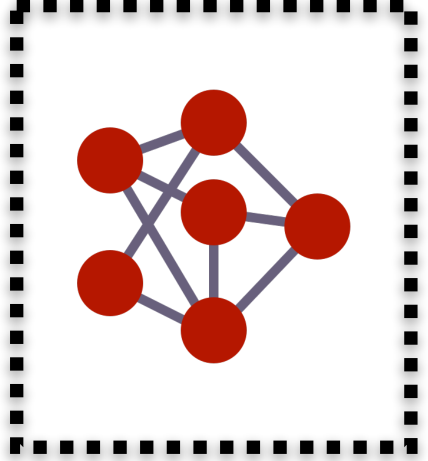
If two *different* models always make identical predictions. Are they equally valuable?

Dirichlet Abstraction

A homogeneous formalization of heterogeneous models

- What: "extract" the predictive distributions of different C-way classification models into *Dirichlet abstractions*.
- Why: To be able to compare different models, through their Dirichlet abstractions.

Black-box

$x_j \in \mathcal{X} \rightarrow$  $\rightarrow \mathbf{M}(x_j) \in \triangle(C)$

$\mathbf{M} : \mathcal{X} \mapsto \triangle(C)$

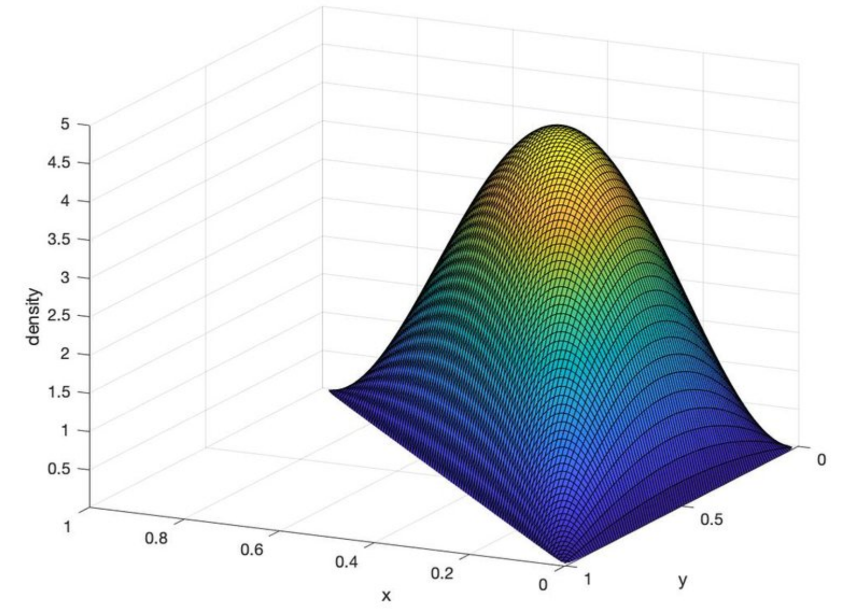
$\{\mathbf{M}(x_j)\}_{(x_j, y_j) \in D}$

$P_{\mathbf{M}(X)}$ where $X \sim P_X$

$\mathrm{supp}(X) = \mathcal{X}$

$\xrightarrow{\text{MLE}}$

$\mathbf{M}(X) \sim \mathrm{Dir}(\boldsymbol{\alpha})$



$\mathbb{Q} := \mathrm{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} \in (0,\infty)^C$

Model Shapley

Valuing a model via its (negated) distance to oracle

Hellinger distance: $d_{\mathrm{H}}(P, Q) := \left[1 - \int \sqrt{p(x)q(x)}\,\mathrm{d}x\right]^{1/2}.$

Value of $\mathbf{M}$ whose Dirichlet abstraction is $\mathbb{Q}$:

Closed-form expression available.

$$-d_{\mathrm{H}}(\mathbb{Q}^*, \mathbb{Q})$$

where $\mathbb{Q}^*$ is the Dirichlet abstraction of an oracle (i.e., optimal classifier).

Interpretation:

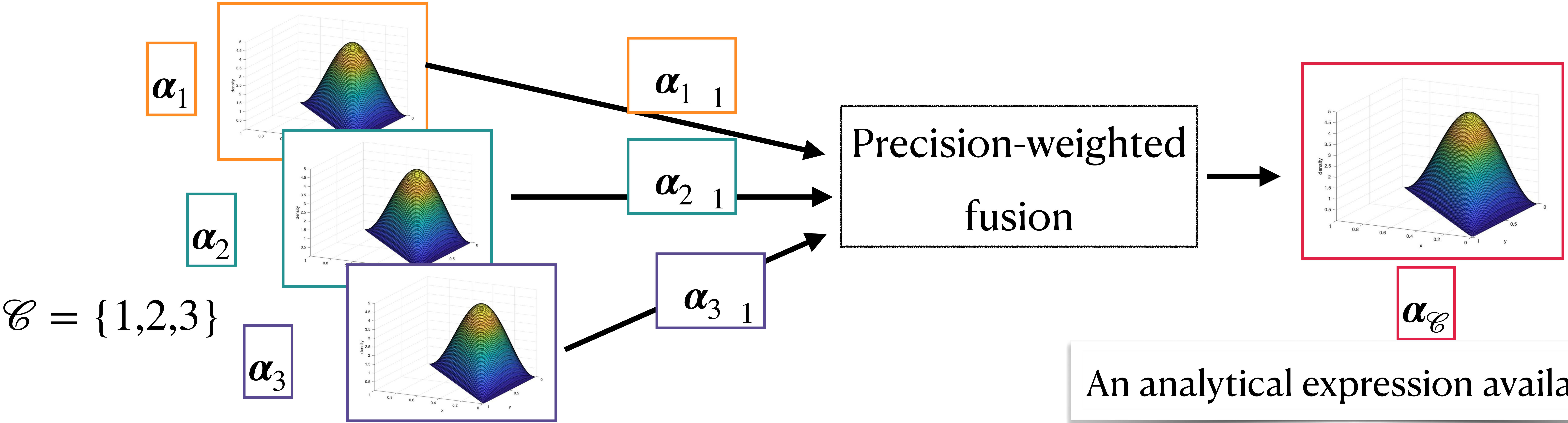
Since the oracle $\mathbb{Q}^*$ is the best (most valuable), by definition,

The value of $\mathbf{M}$ is defined to be the statistical similarity between its Dirichlet abstraction $\mathbb{Q}$ and $\mathbb{Q}^*$.

Model Shapley

Coalition: fusion of Dirichlet abstractions

Precision-weighted fusion (informal): *For a coalition $\mathscr{C} \subseteq [N] = \{1, \ldots, N\}$ of models, define an fused Dirichlet abstraction $\mathbb{Q}_{\mathscr{C}} := Dir(\boldsymbol{\alpha}_{\mathscr{C}})$, based on their respective precision (of the Dirichlet abstractions).*



An analytical expression available.

Model Shapley

Equitability from the Shapley value

Model Shapley value: $$\phi_i := \sum_{\mathscr{C} \subseteq [N] \setminus \{i\}} \omega_{\mathscr{C}} [\nu(\mathscr{C} \cup \{i\}) - \nu(\mathscr{C})]$$

where $\nu(\mathscr{C}) := -d_{\mathrm{H}}(\mathbb{Q}^*, \mathbb{Q}_{\mathscr{C}})$ and $\omega_{\mathscr{C}} := \ \mathscr{C}\ ! \times (N - \ \mathscr{C}\ ! - 1)!/N!$.

Equitability

- Null player
- Symmetry
- Linearity

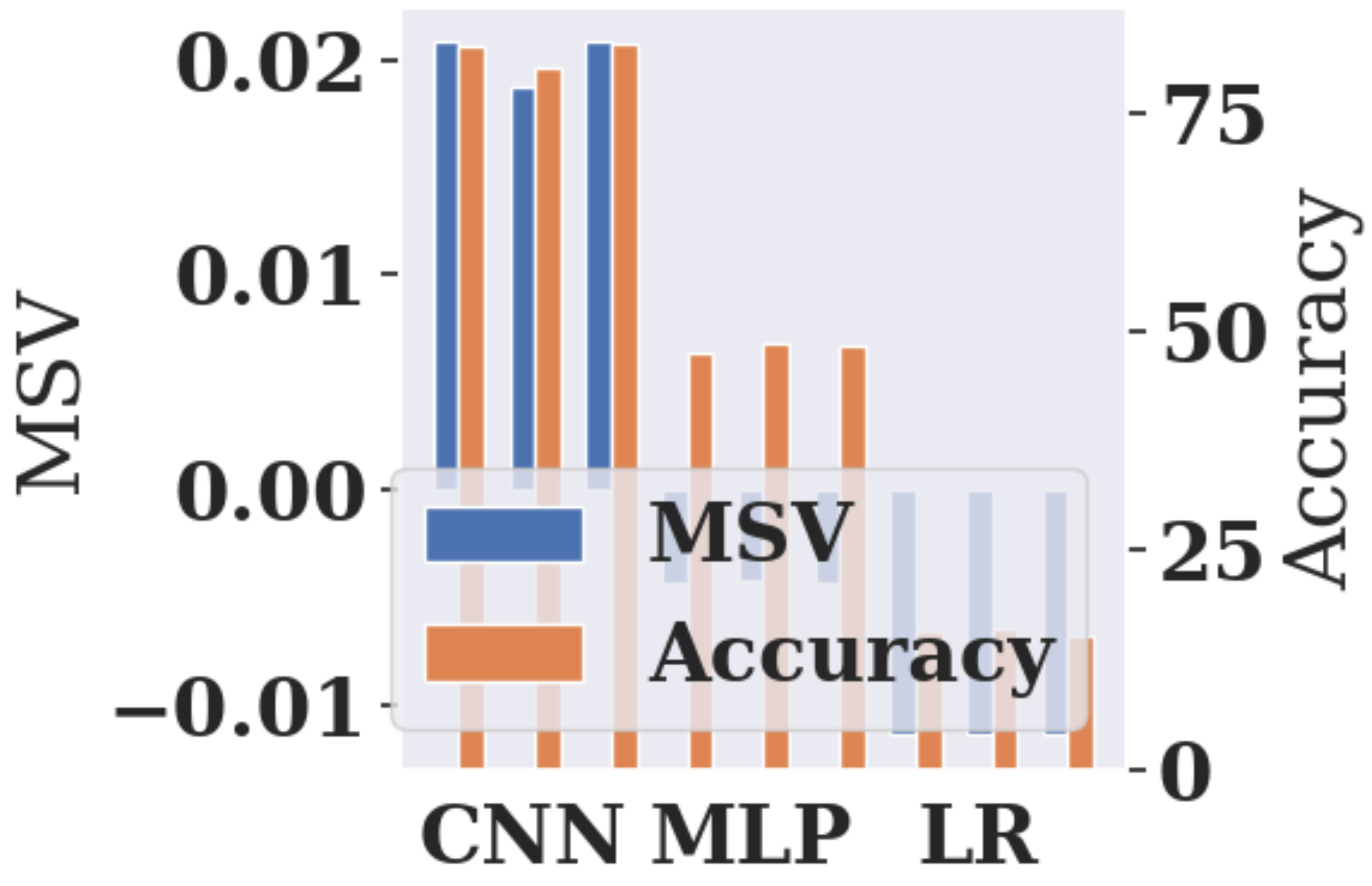
Analytic properties

For computational tractability.

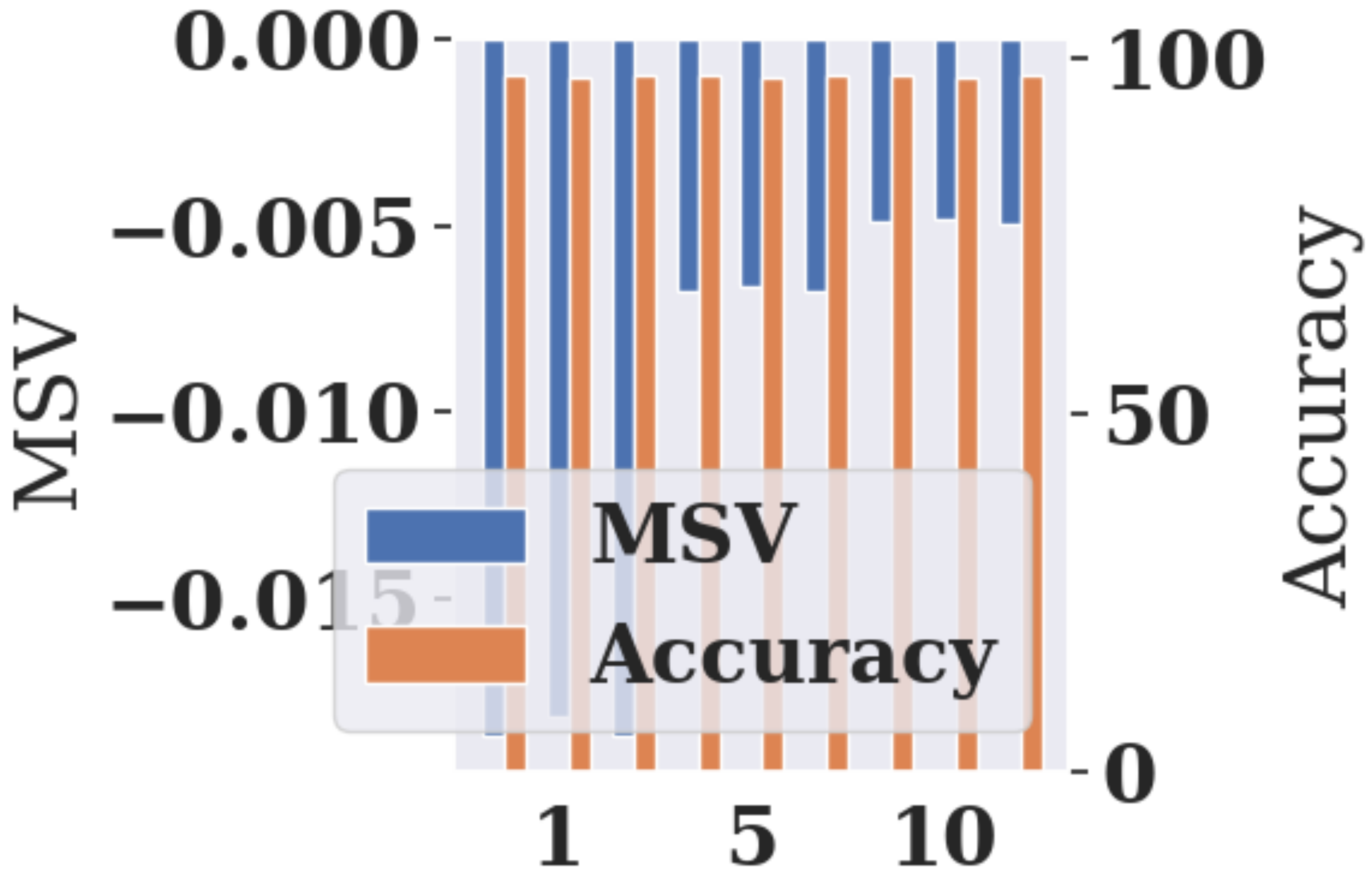
- ν is evaluated in closed-form
- $\mathbb{Q}_{\mathscr{C}}$ has analytic expression

Experiments

MSVs vs. common criteria



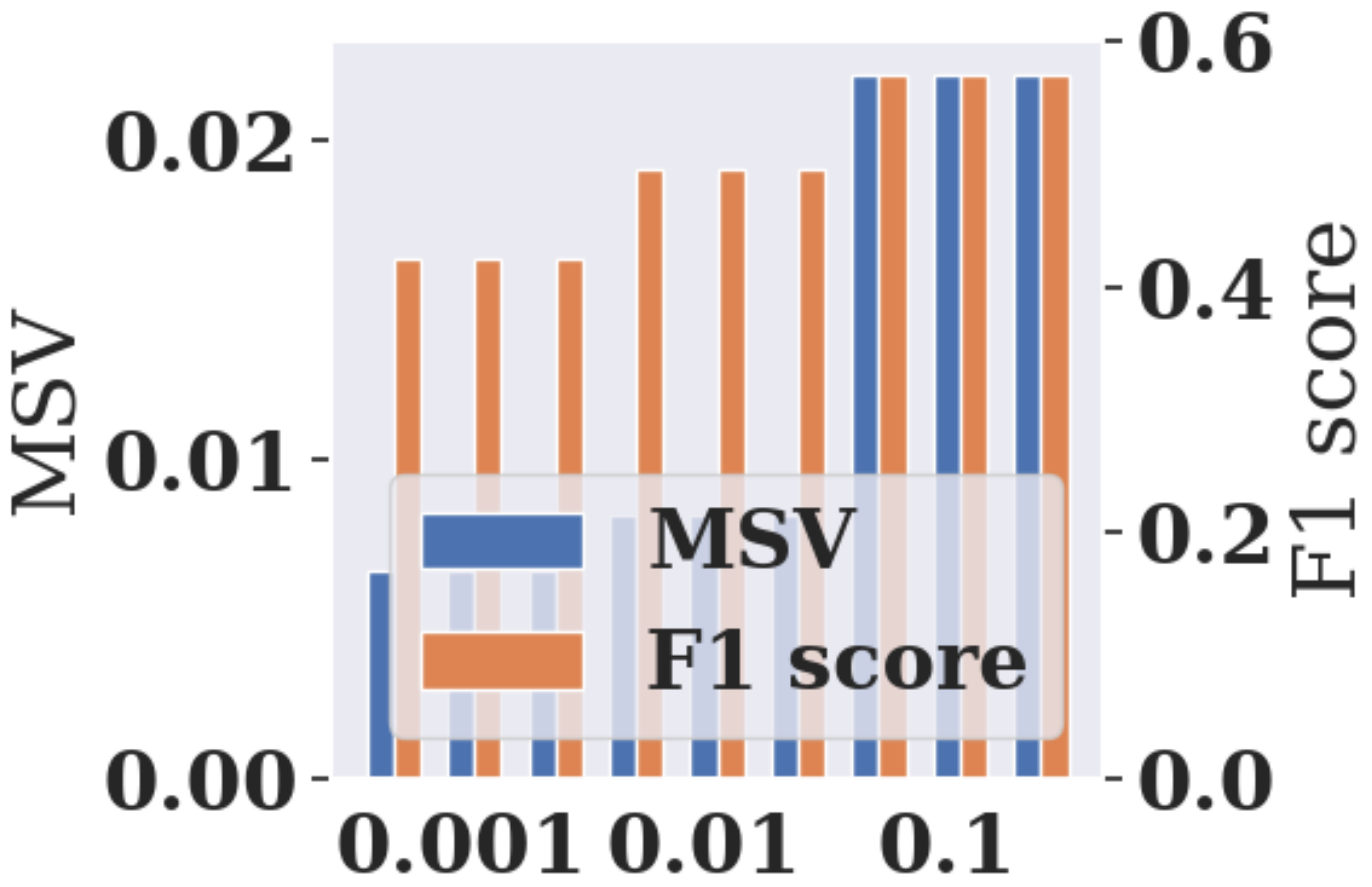
MSV vs. accuracy



MSVs vs. certainty (fixed accuracy)

Experiments

MSVs vs. common criteria



MSV vs. F1 score

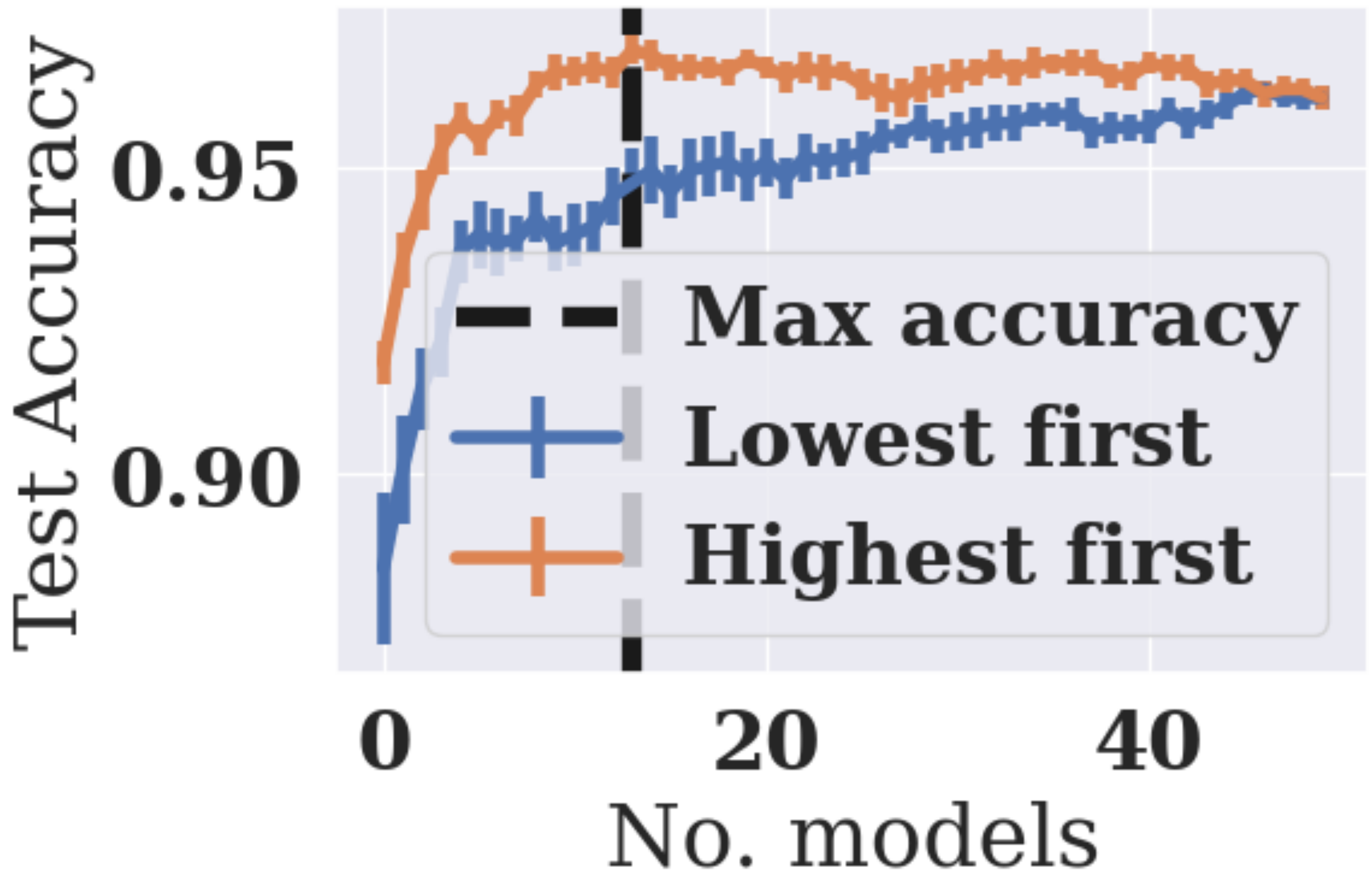
For highly imbalanced data.

Experiments

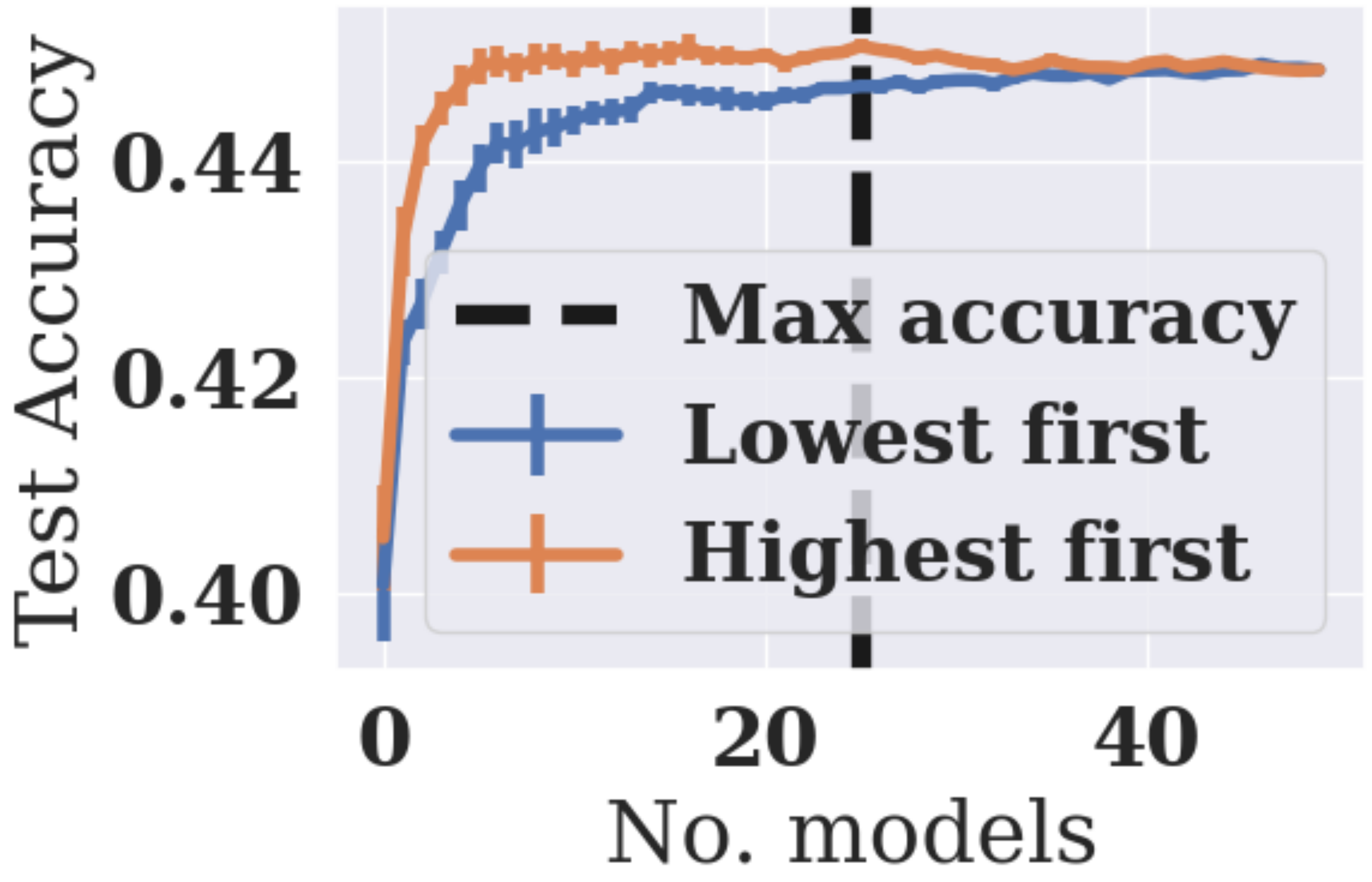
MSV for pruning in ensemble

Identifying the most valuable subsets of models for ensemble.

Max occurs *before* using all models.



Dataset: BreastCancer, Base learner: Decision Trees, Ensemble learner: Random forest.



Dataset: CIFAR-10, Base learner: LeNets, Ensemble learner: Voting.

Summary and conclusion

- **Model Valuation is a feasible alternative to Data Valuation**.

- Propose *Dirichlet abstractions* to compare different models;
- Define the *model Shapley* as an equitable valuation;
- Experiments show MSVs behave consistently with common evaluation criteria.

- Future works can consider extension to generative models.

See you at poster session

Great Hall & Hall B1+B2 #1710

Wed 13 Dec 6 p.m. EST — 8 p.m. EST

QR: Link to our NeurIPS poster page.

References

- K.Singhal, H.Sidahmed, Z.Garrett, S.Wu, J.K.Rush, and S.Prakash. Federated reconstruction: Partially local federated learning. In *Proc. NeurIPS*, 2021.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proc. NeurIPS*, 2020.
- A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proc. WWW*, 2007.
- J. M. Drazen, S. Morrissey, D. Malina, M. B. Hamel, and E. W. Campion. The importance — and the complexities — of data sharing. *New England Journal of Medicine*, 375(12):1182–1183, 2016.
- M.F.HaqueandR.Krishnan.Towardautomatedcyberdefensewithsecuresharingofstructured cyber threat intelligence. *Information Systems Frontiers*, 23(4):883–896, 2021.
- R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gu¨rel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *Proc. AISTATS*, pages 1167–1176, 2019.
- G. A. Kaissis, M. R. Makowski, D. Ru¨ckert, and R. F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2:305–311, 2020.
- T. Minka. Estimating a dirichlet distribution, 2000.