ACTIVE LEARNING AND APPLICATION TO SCIENTIFIC COMPUTING

Jiaxi Zhao Dec 2021

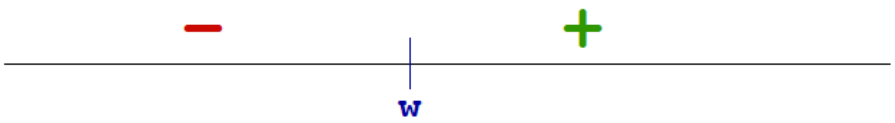
HEURISTIC EXAMPLE

Efficient search through hypothesis space

Threshold functions on the real line:

$H = \{h_w : w \in \mathbb{R}\}$

$h_w(x) = 1(x \geq w)$



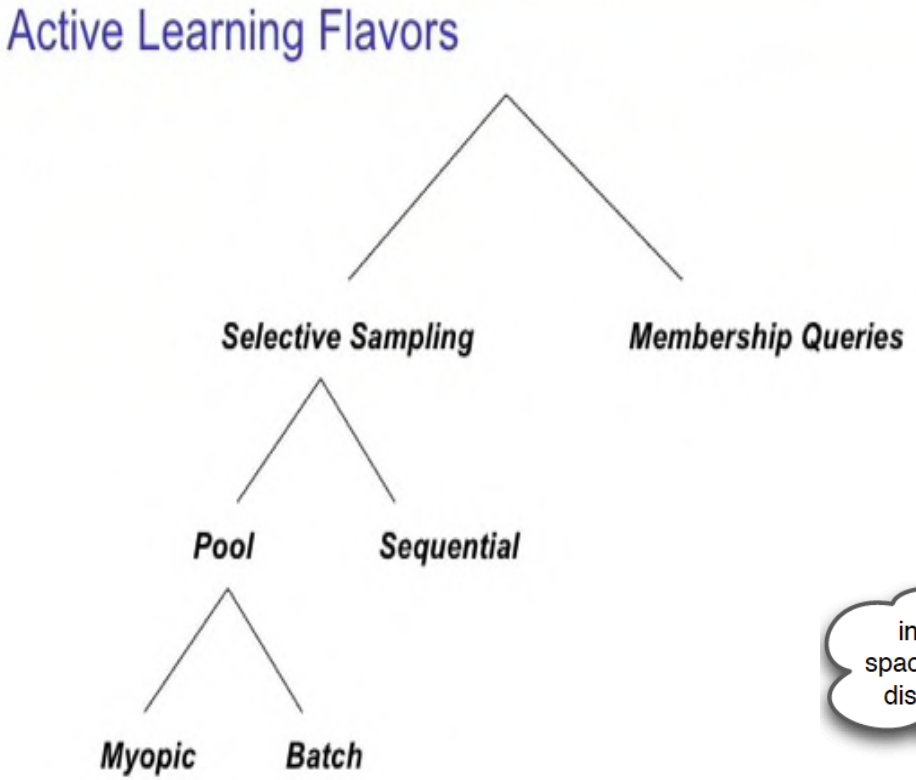
Supervised: for misclassification error $\leq \epsilon$, need $\approx 1/\epsilon$ labeled points.

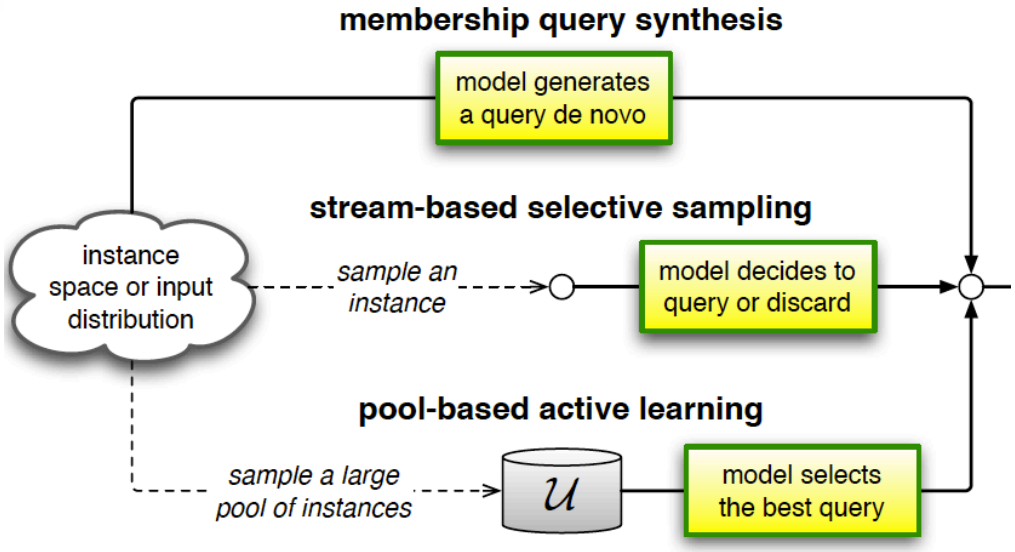
Active learning: instead, start with $1/\epsilon$ *unlabeled* points.

Binary search: need just $\log 1/\epsilon$ labels, from which the rest can be inferred. *Exponential improvement in label complexity!*

Challenges: Nonseparable data? Other hypothesis classes?

GENERAL FRAMEWORK OF ACTIVE LEARNING



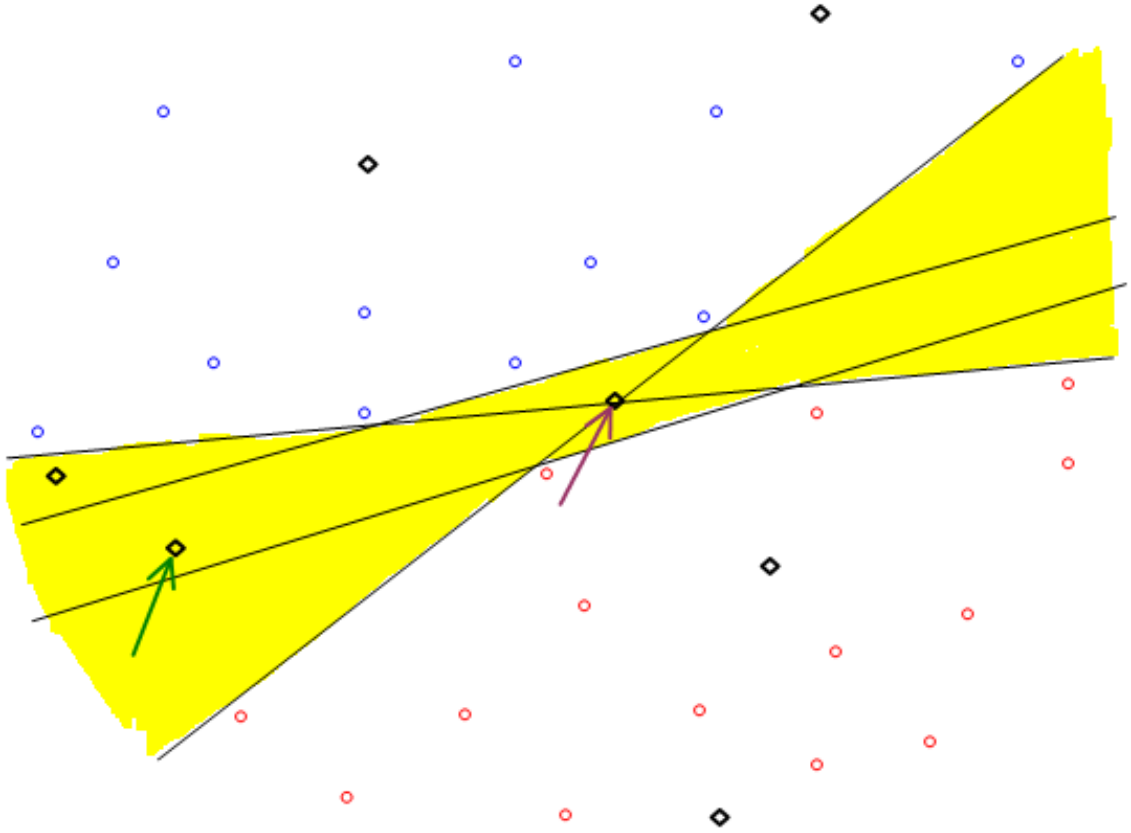


QUERY STRATEGY: UNCERTAINTY SAMPLING

- Query the event that the current classifier is most uncertain about
- Needs measure of uncertainty, probabilistic model for prediction
- Examples:
  - Entropy
  - Least confident predicted label
  - Euclidean distance (e.g. point closest to margin in SVM)

QUERY STRATEGY: QUERY-BY-COMMITTEE

- Yellow = valid hypotheses



QUERY STRATEGY: EXPECTED MODEL CHANGE

- Modeling the KL divergence of the posteriors measures the amount of information gain expected from query (where x' is the queried data):
- Goal: choose a query that *maximizes* the KL divergence between posterior and prior
- Basic idea: largest KL divergence between updated posterior probability and the current posterior probability represents largest gain

QUERY STRATEGY : VARIANCE REDUCTION AND FISHER INFORMATION RATIO

$$\sigma_o^2 \approx S(\mathcal{L};\theta)\left[\frac{\partial o}{\partial \theta}\right]^{\mathsf{T}}\left[\frac{\partial^2}{\partial \theta^2}S(\mathcal{L};\theta)\right]^{-1}\left[\frac{\partial o}{\partial \theta}\right],$$

$$x^*_{FIR} = \underset{x}{\operatorname{argmin}}\,\operatorname{tr}\left(\mathcal{I}_x(\theta)^{-1}\mathcal{I}_{\mathcal{U}}(\theta)\right),$$

COMPARISON

- Empirical Comparison with Supervised Learning:
  - It is the only method to be used in label-expensive setting
  - It truly show better sample complexity
  - It is not adaptive to model change

J. BALDRIDGE AND M. OSBORNE. ACTIVE LEARNING AND THE TOTAL COST OF ANNOTATION. IN PROCEEDINGS OF THE CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), PAGES 9–16. ACL PRESS, 2004.

APPLICATION TO SCIENTIFIC COMPUTING

- Labeled data is extremely expensive to obtain, due to time and money consumption
- We have enormous freedom to specify settings in scientific computing, i.e. IC & BC in PDE…

APPLICATION TO SCIENTIFIC COMPUTING

- DeePMD & DeePGEN:
  - Ab-initio calculation is extremely time-consuming
  - Experiment such as Cryo-SEM is extremely expensive
  - Query committee based active learning

ACTIVE LEARNING OF UNIFORMLY ACCURATE INTERATOMIC POTENTIALS FOR MATERIALS SIMULATIONL ZHANG, DY LIN, H WANG, R CAR, E WEINAN - PHYSICAL REVIEW MATERIALS, 2019

APPLICATION TO SCIENTIFIC COMPUTING

- In addition, dense sampling of the chemical composition space is not always necessary. For example, the initial ANI training set of 20 million molecules could be replaced with 5.5 million training points selected using an active learning method that added poorly predicted molecular examples from each training cycle

SMITH, J. S., B. NEBGEN, N. LUBBERS, O. ISAYEV, AND A. E. ROITBERG (2018), J. CHEM. PHYS. 148 (24), 241733.

APPLICATION TO SCIENTIFIC COMPUTING

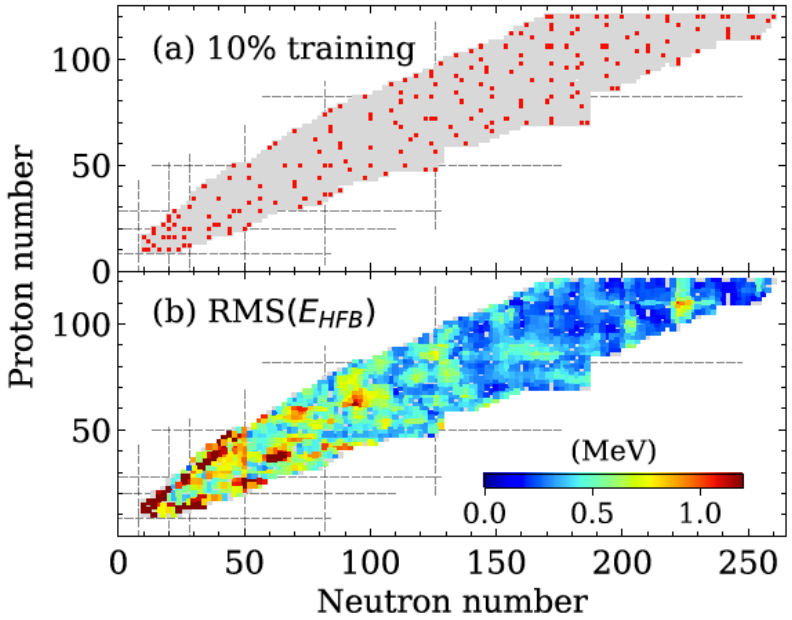


FIG. 3 **DFT emulator with ANN.** (a) The database nuclei (grey) as a function of N and Z. Nuclei included in the 10% training data set obtained by the active learning are marked in red. (b) The root mean square deviation between the total energy for the testing data set calculated in DFT (E_{HFB}) and with the committee of ANN. Adopted from (Lasseri *et al.*, 2020).

BOEHNLEIN, AMBER, ET AL. "ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN NUCLEAR PHYSICS." *ARXIV PREPRINT ARXIV:2112.02309*(2021).

FURTHER REFERENCE:

- Settles, Burr. "Active learning literature survey." (2009).
- Carleo, Giuseppe, et al. "Machine learning and the physical sciences." *Reviews of Modern Physics* 91.4 (2019): 045002.
- http://people.eecs.berkeley.edu/~jordan/courses/294-fall09/
- https://hunch.net/~active_learning/active_learning_icml09.pdf

THANK YOU