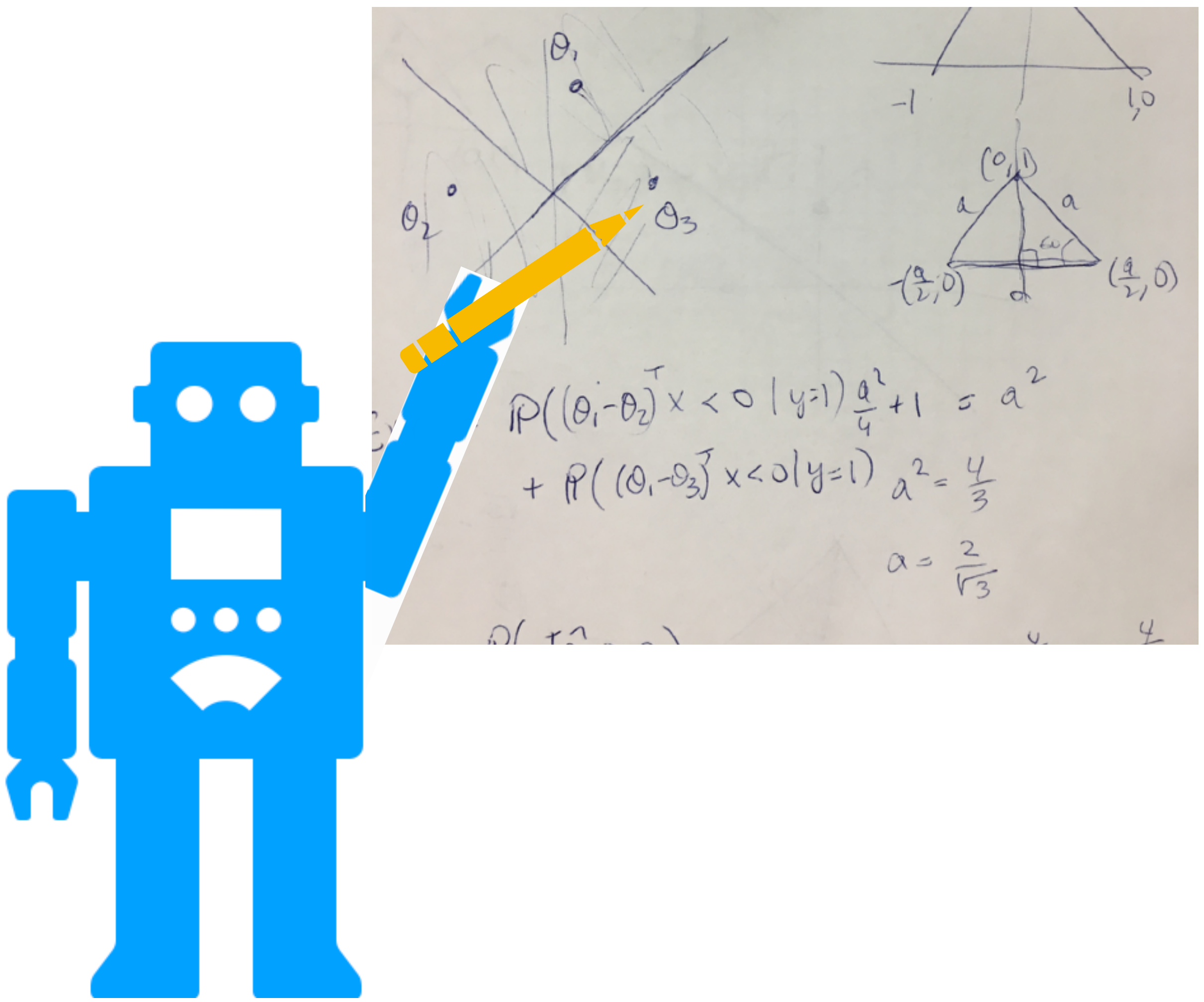
Nonparametric Active Learning



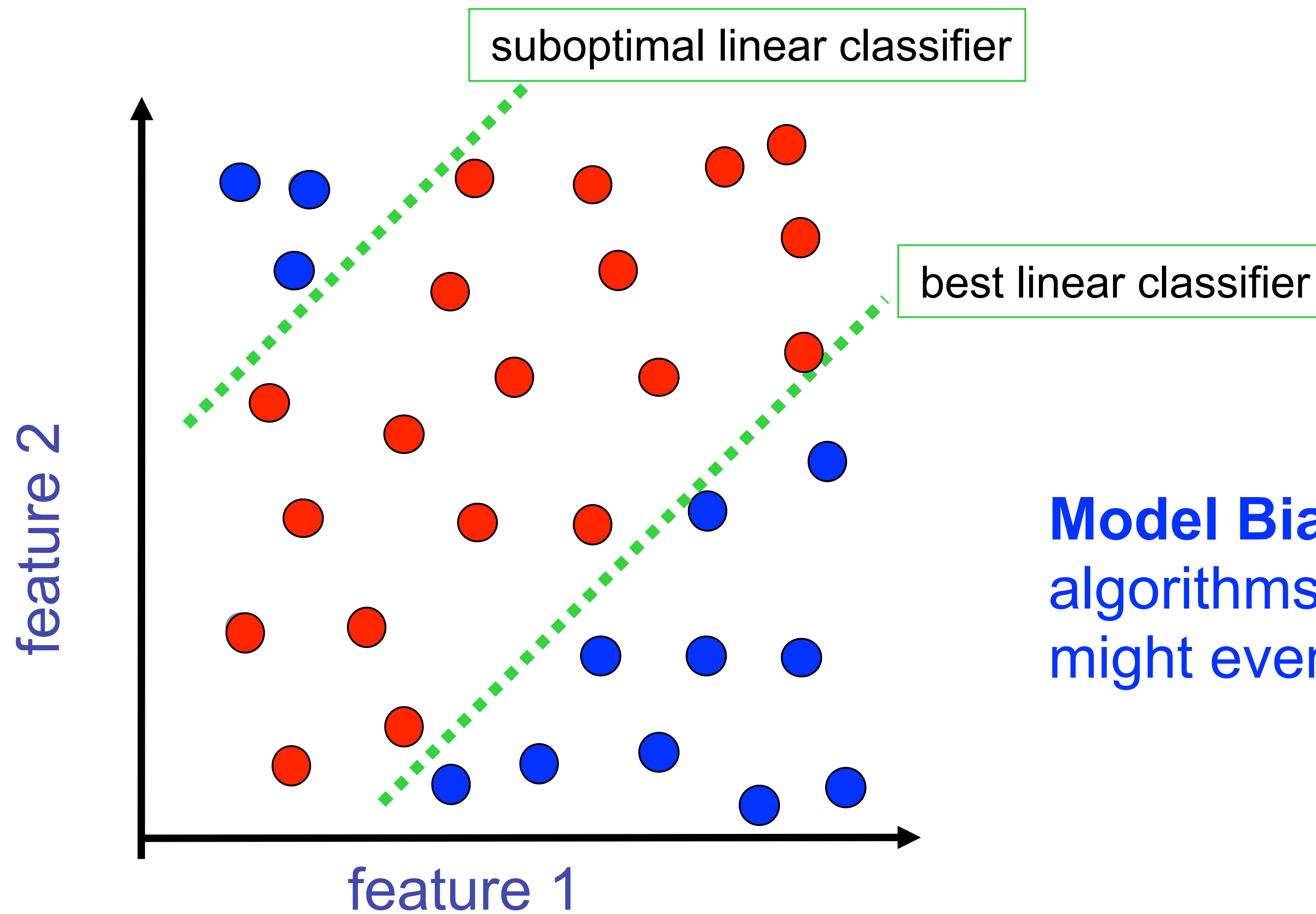
Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak
UW-Madison
rdnowak@wisc.edu

ICML | 2019

Thirty-sixth International Conference on Machine Learning

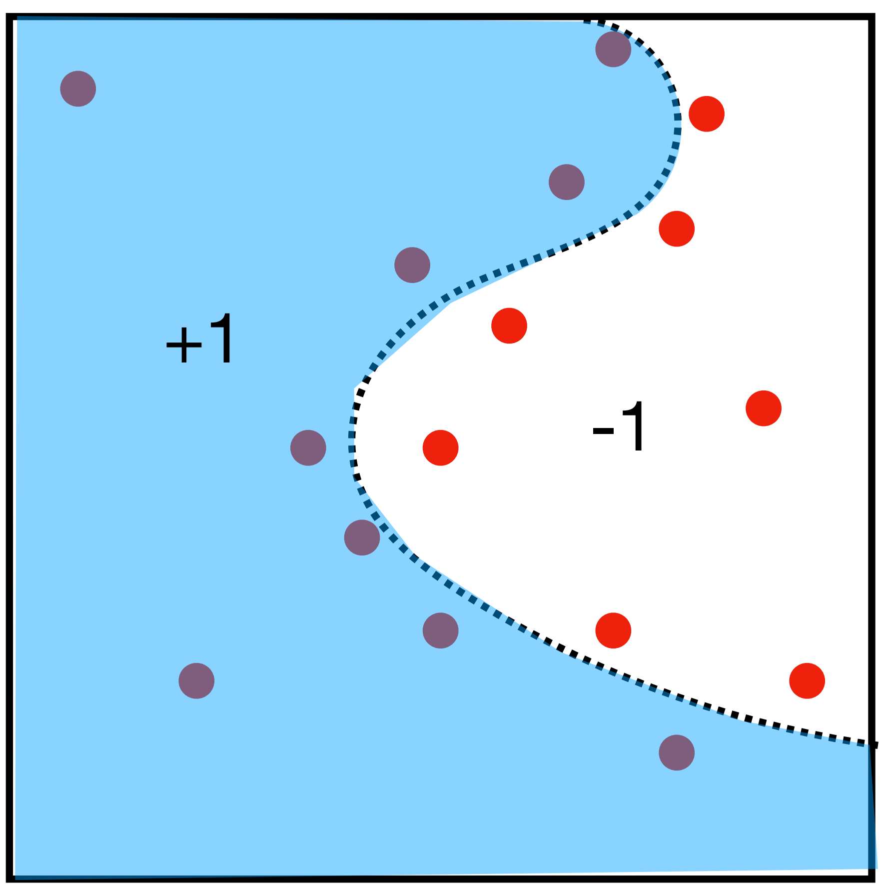
Active Learning and Inductive Bias



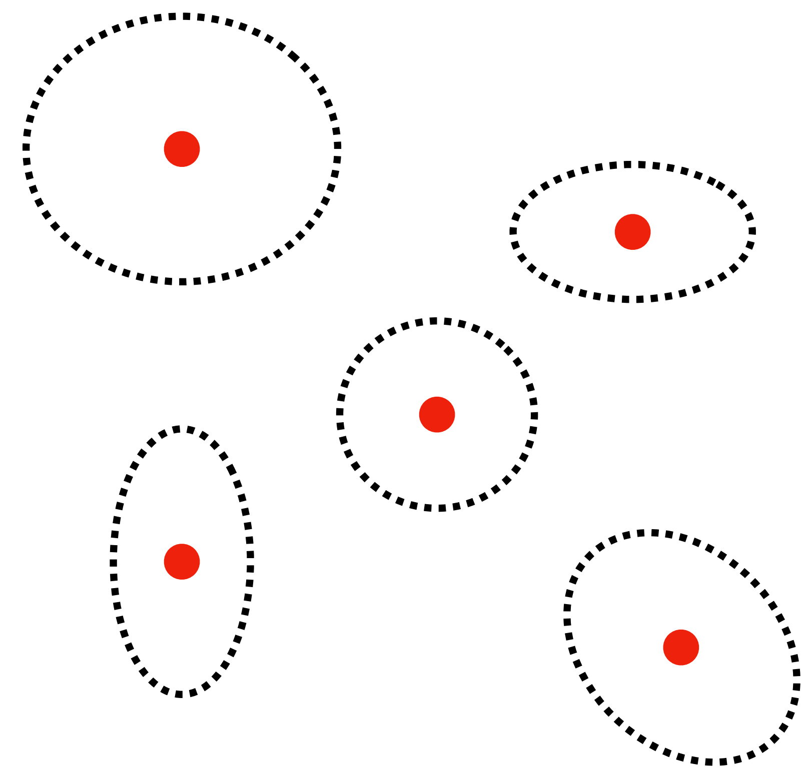
Model Bias: Standard active learning algorithms perform no better than passive or might even converge to suboptimal solutions!

Two Faces of Active Learning

Goal: Use nonparametric (or overparameterized) models to avoid bias and design active learning algorithms that exploit intrinsic structure in data



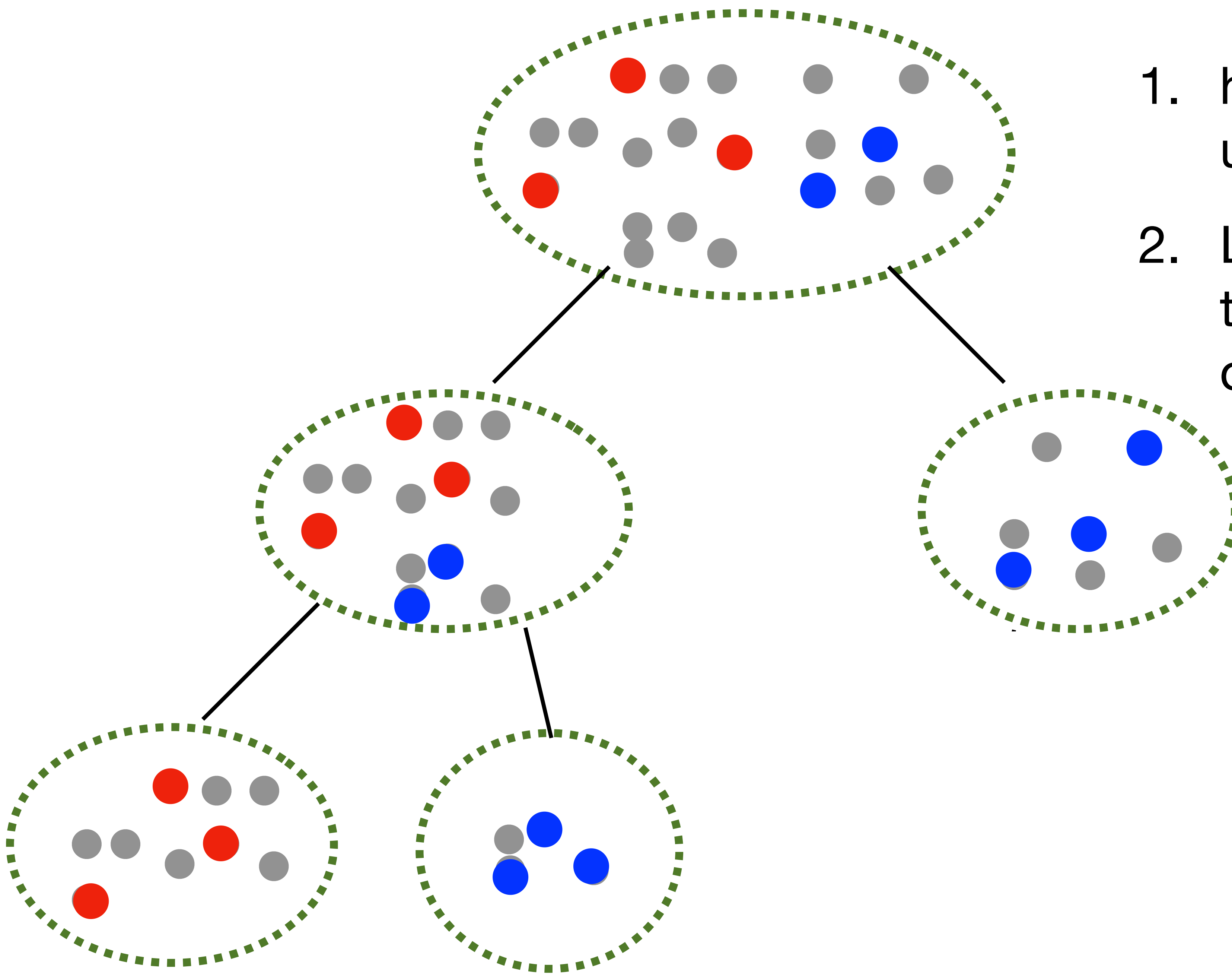
label examples close to estimated decision boundary



find clusters in unlabeled data and label one representative from each

Dasgupta, Sanjoy. "Two faces of active learning." *Theoretical computer science* 412.19 (2011): 1767-1781.

Hierarchical Clustering for Active Learning



1. hierarchically cluster unlabeled data
2. Label examples to test homogeneity of each cluster

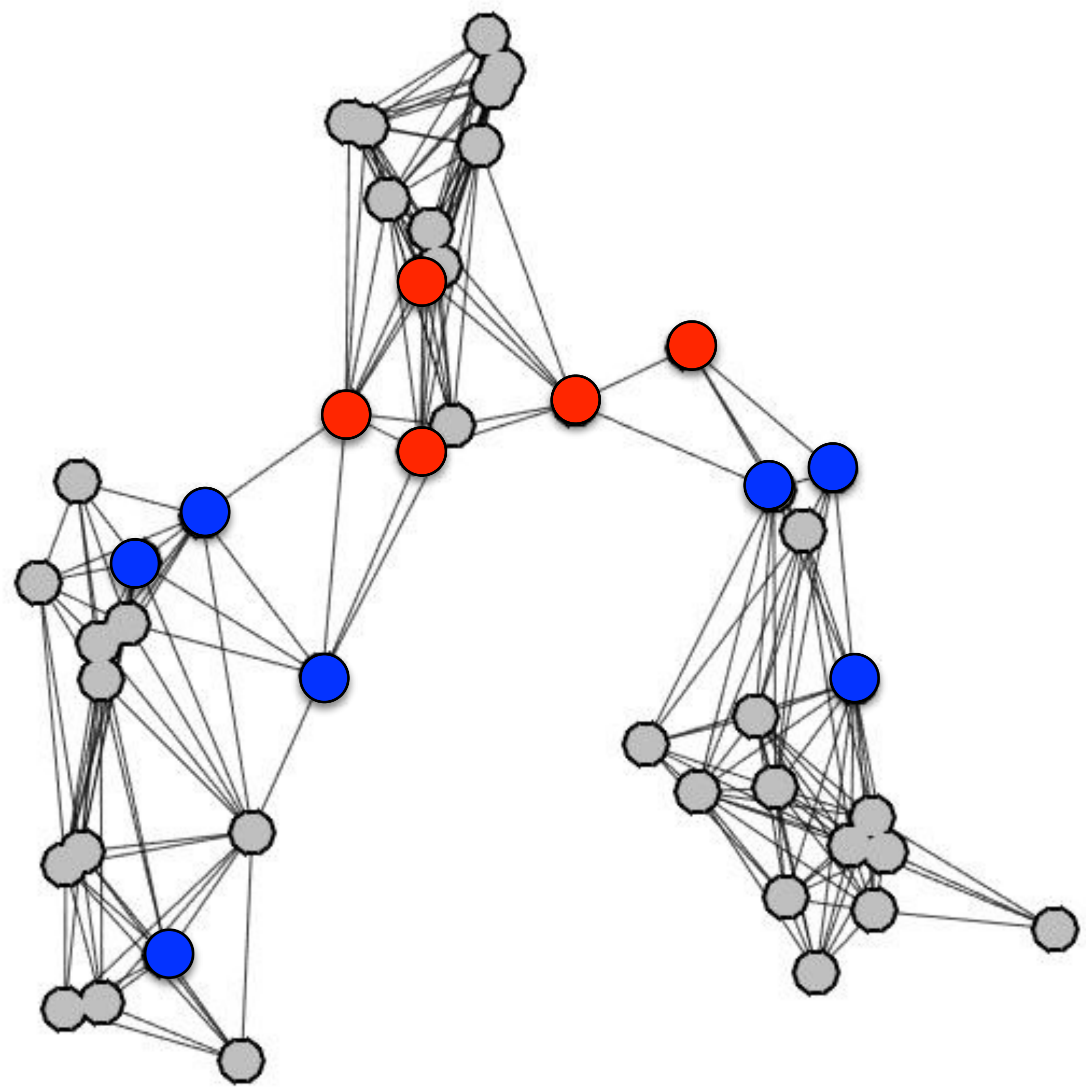
Theorem: If it is possible to prune the cluster tree to m leaves that are fairly pure in the labels of their constituent points, then O(m) labeled examples suffice to accurately label the entire dataset

Dasgupta and Hsu. "Hierarchical sampling for active learning." *ICML* 2008

Combining Active and Semi-Supervised Learning

Construct nearest neighbor graph of unlabeled dataset

Using prior model based on graph-Laplacian, select labels to minimize predicted risk



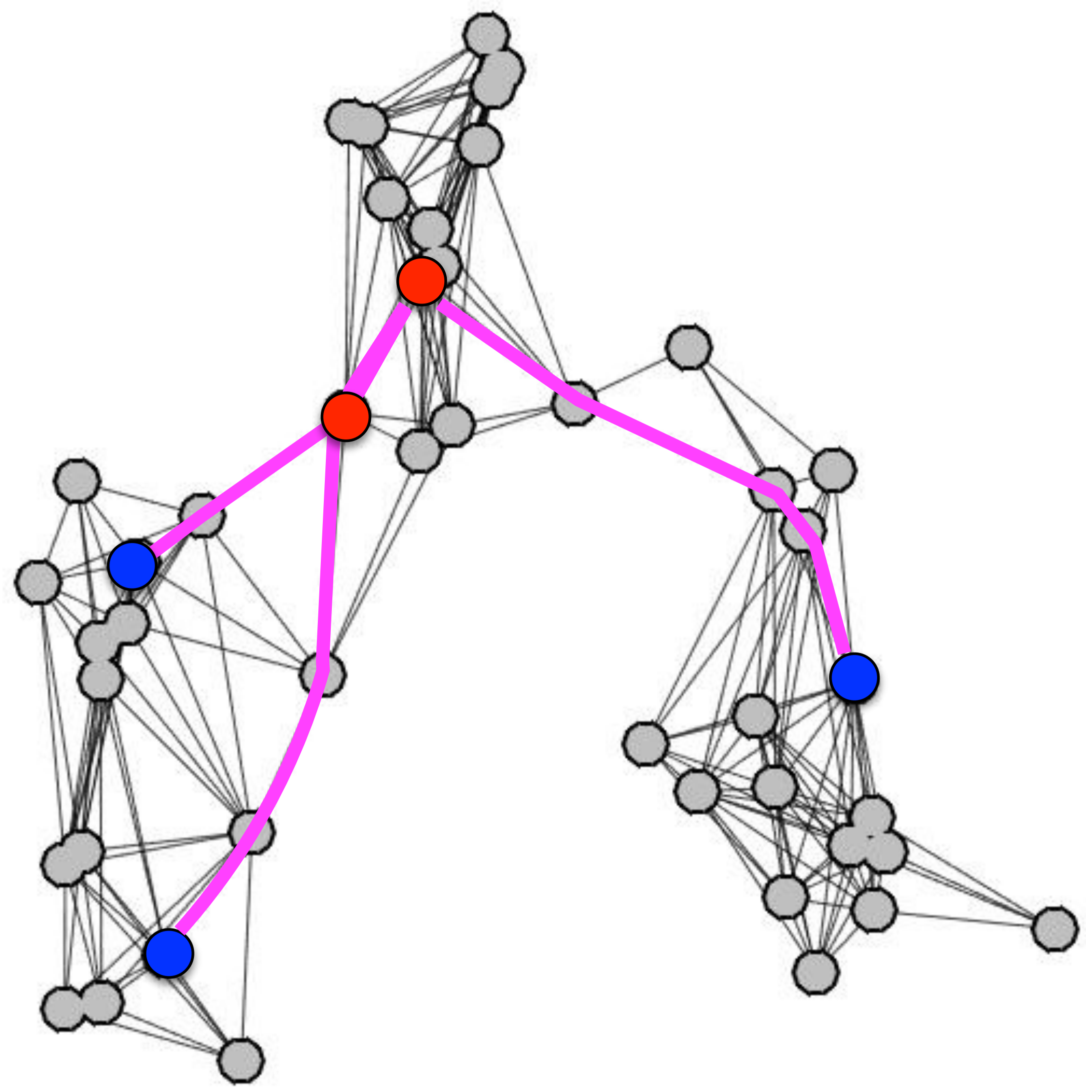
Propagate labels to rest of graph: e.g., nearest-neighbors, graph-Laplacian, etc

Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop*. Vol. 3. 2003.

Provably Correct Algorithm: Binary Search on Graph

basic idea:

then find all shortest paths between red and blue labeled nodes and bisect the shortest shortest path

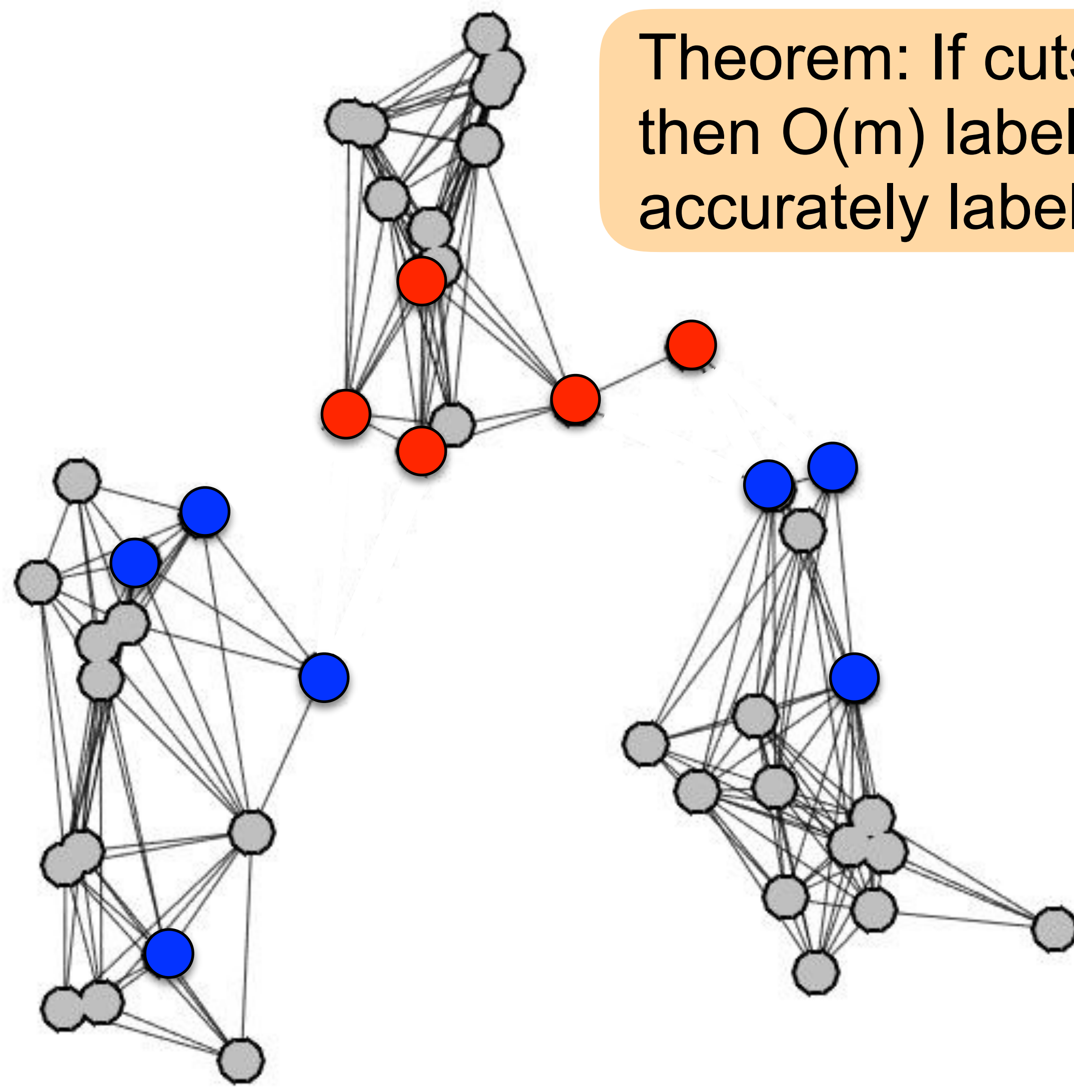


Binary Search on Graph

Recursive application of shortest-shortest path bisection efficiently identifies cutset, partitioning graph into pure-labeled components

Theorem: If cutset consists of m edges, then O(m) labeled examples suffice to accurately label the entire dataset

propagate labels to rest of graph: e.g., nearest-neighbors, graph-Laplacian, etc



Dasarathy, Gautam, RN, and Xiaojin Zhu. *Conference on Learning Theory*. 2015.

Active Learning with Kernels and Neural Nets

Active learning based on nearest neighbor graphs and clustering can be effective, but require two separate steps

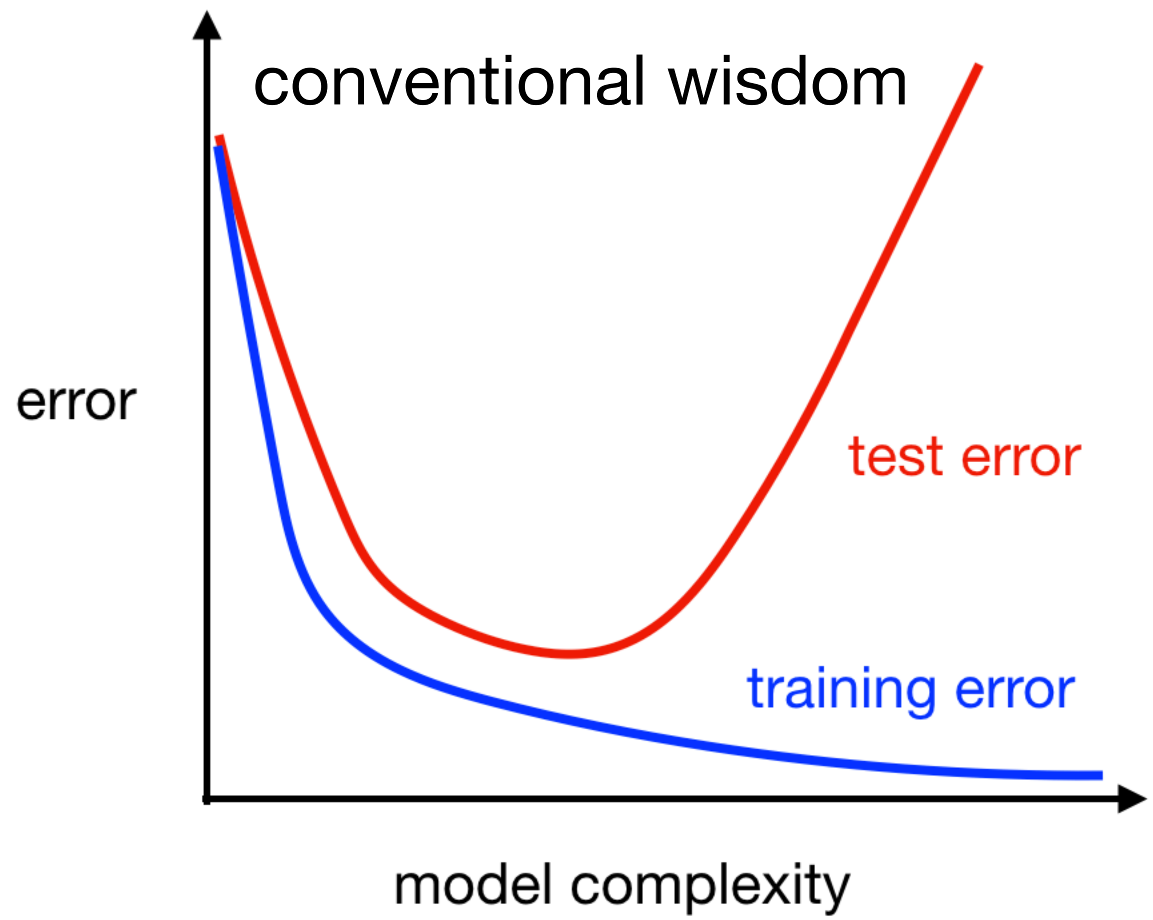
1. build graph or partition on unlabeled dataset
2. exploit graph/cluster structure for active learning

Can we develop similar procedures that can be applied directly to popular classifiers like kernel methods and neural networks?

multilayer neural net $y = \boldsymbol{W}_L\, f(\boldsymbol{W}_{L-1} \cdots f(\boldsymbol{W}_1 \boldsymbol{x}) \cdots))$

kernelized classifier $y = \sum_{i=1}^{n} \alpha_i\, k(\boldsymbol{x}_i, \boldsymbol{x})$

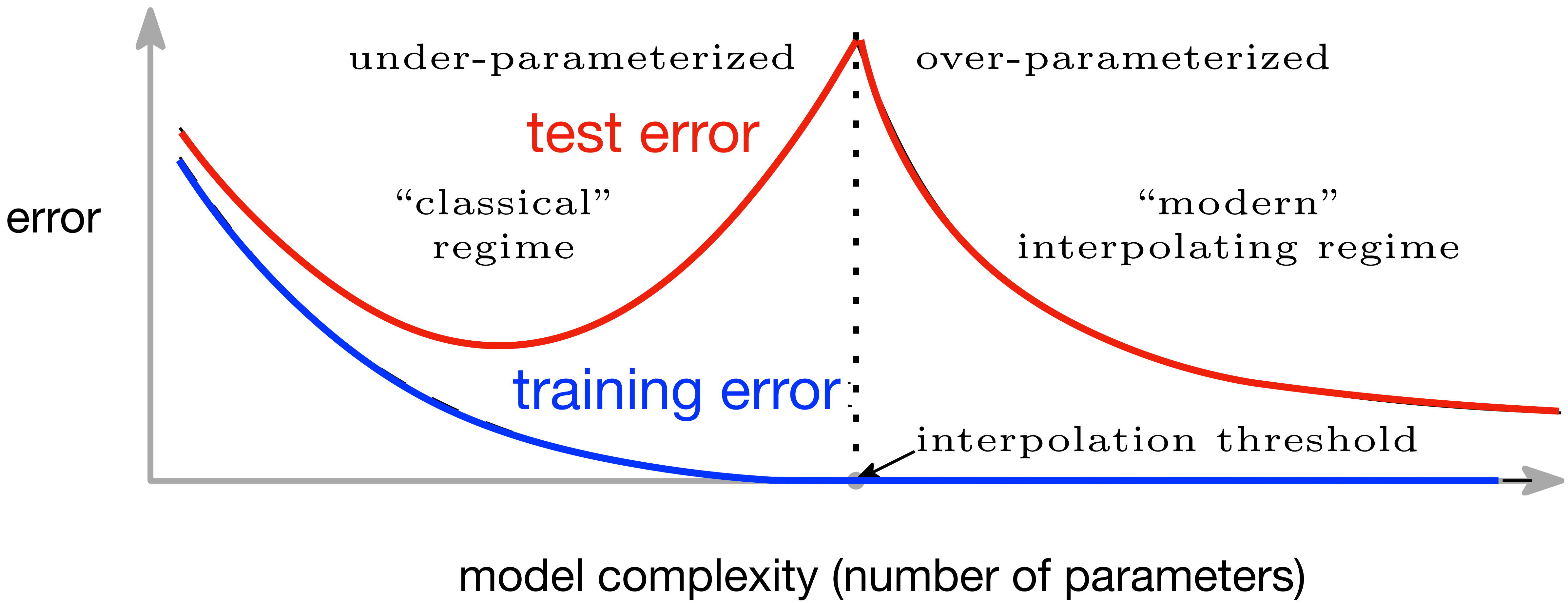
Rethinking Conventional Wisdom



deep nets are trained to perfectly fit training data, yet still generalize well

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530* (2016).

Double-Descent

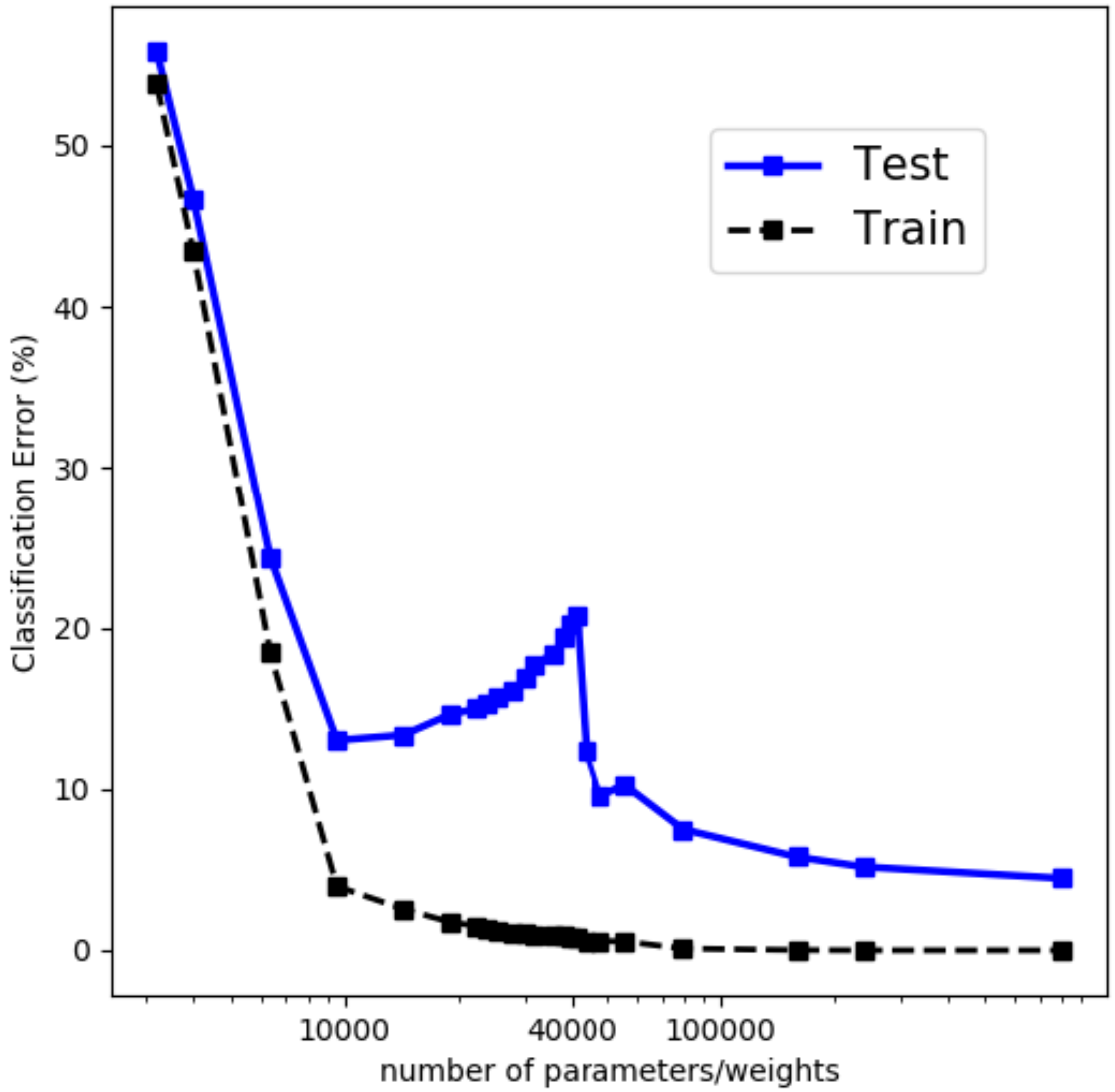


Maximizing model *smoothness* subject to the interpolation constraints is a form of the Occam's razor

from Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning and the bias-variance trade-off." *arXiv preprint arXiv:1812.11118* (2018).

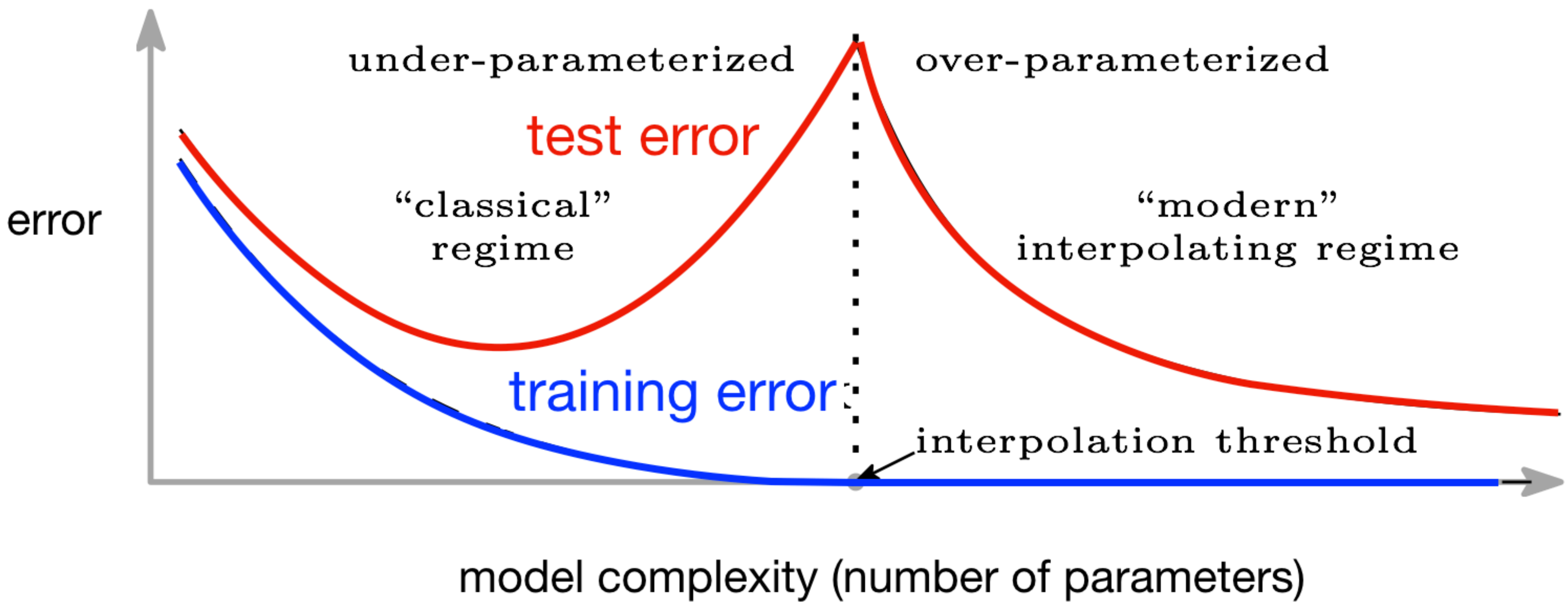
Theory Meets Practice

MNIST experiments with kernels, random features, and neural nets



from Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning and the bias-variance trade-off." *arXiv preprint arXiv:1812.11118* (2018).

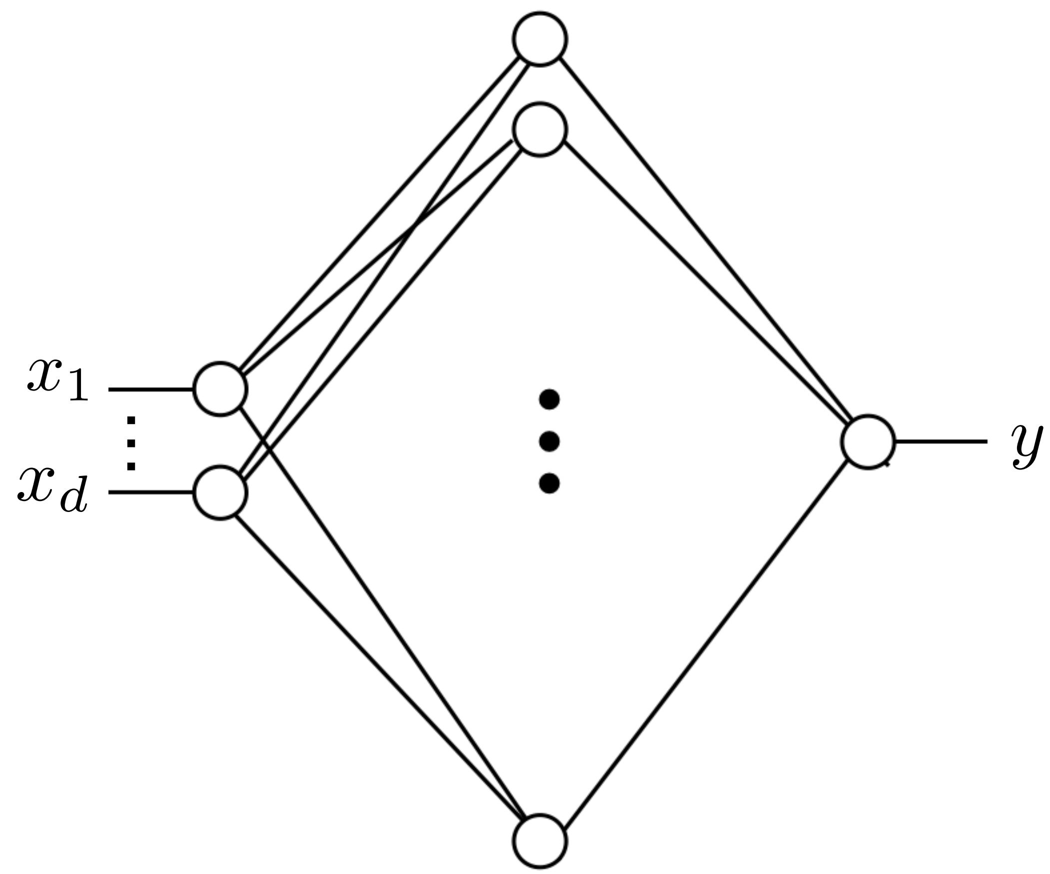
Rethinking Active Learning



Standard active learning theory and methods are based on bounding test error in terms of training error (e.g., VC theory)

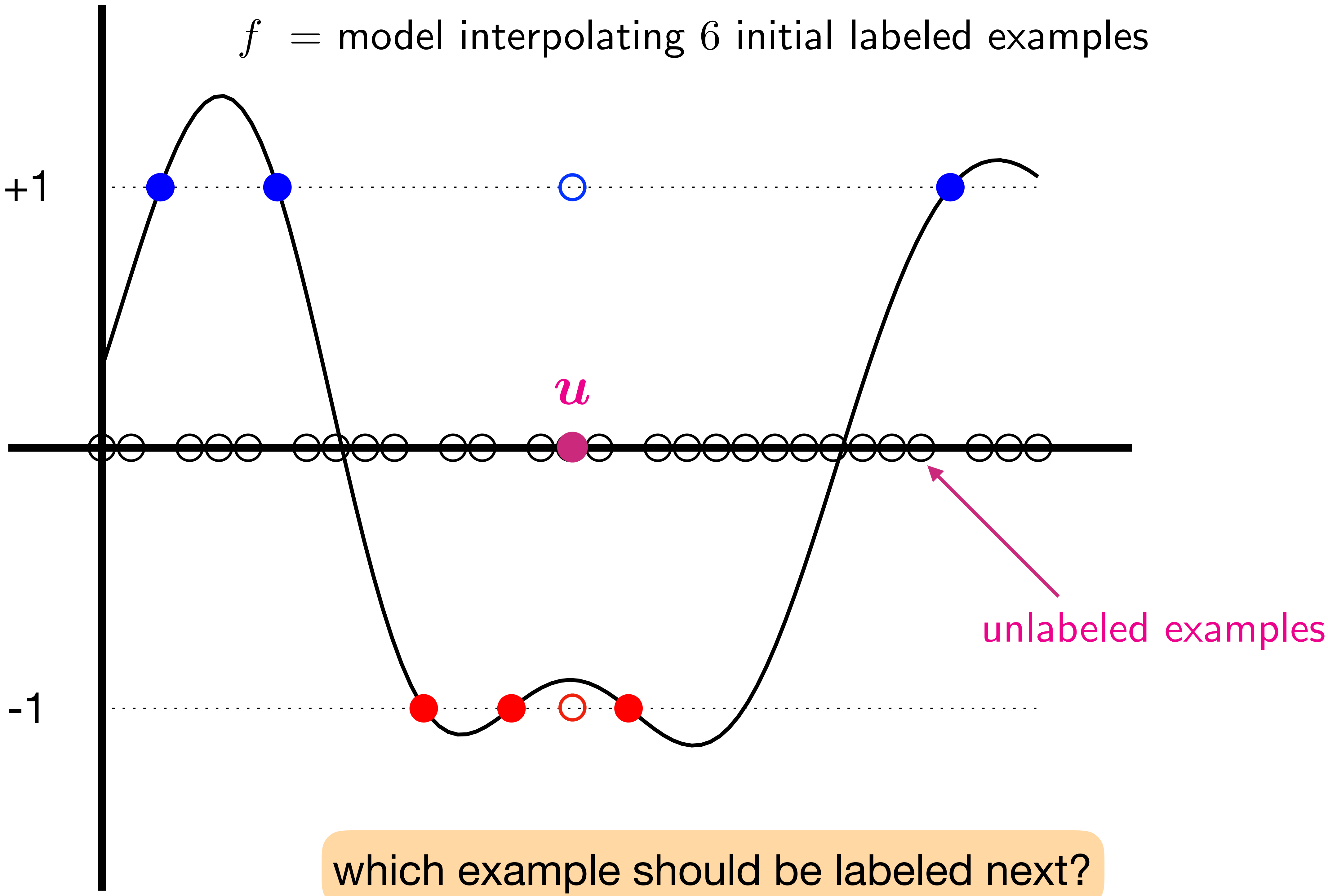
In the interpolating regime, the training error is identically zero and yields no information about the expected loss

Kernel Machines and Neural Networks



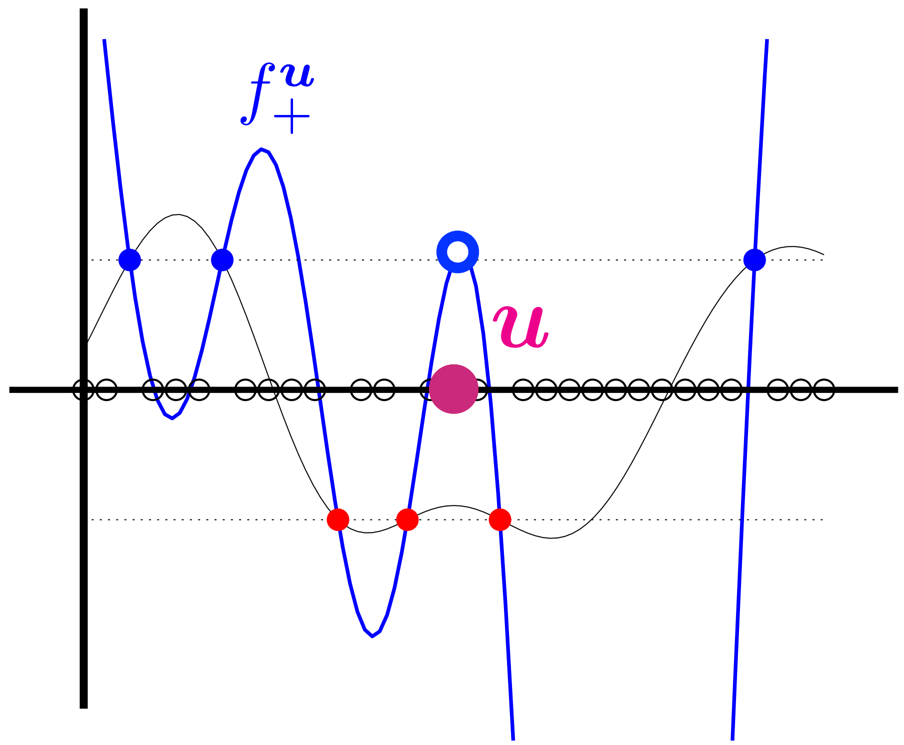
- kernel machine is a single hidden-layer neural network
- interpolation possible with infinite dimensional Reproducing Kernel Hilbert Space (RKHS) or overparameterized neural net

Active Learning in Overparameterized Setting



which example should be labeled next?

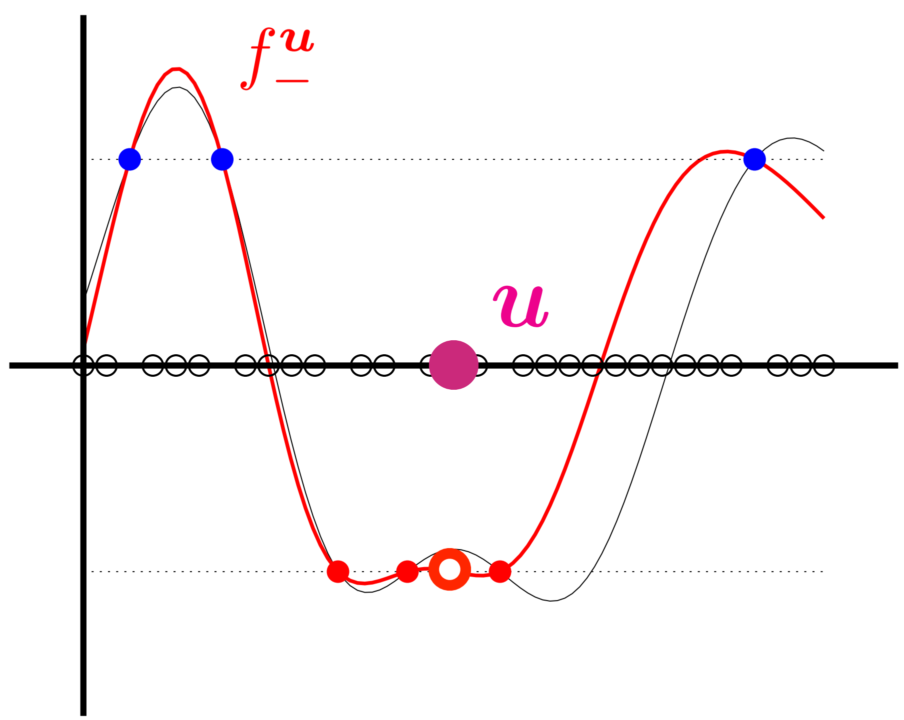
Active Learning in Overparameterized Setting



New example in between
identically labeled examples

difficult to interpolate, large $\|f_+^{\boldsymbol{u}}\|$

less smooth

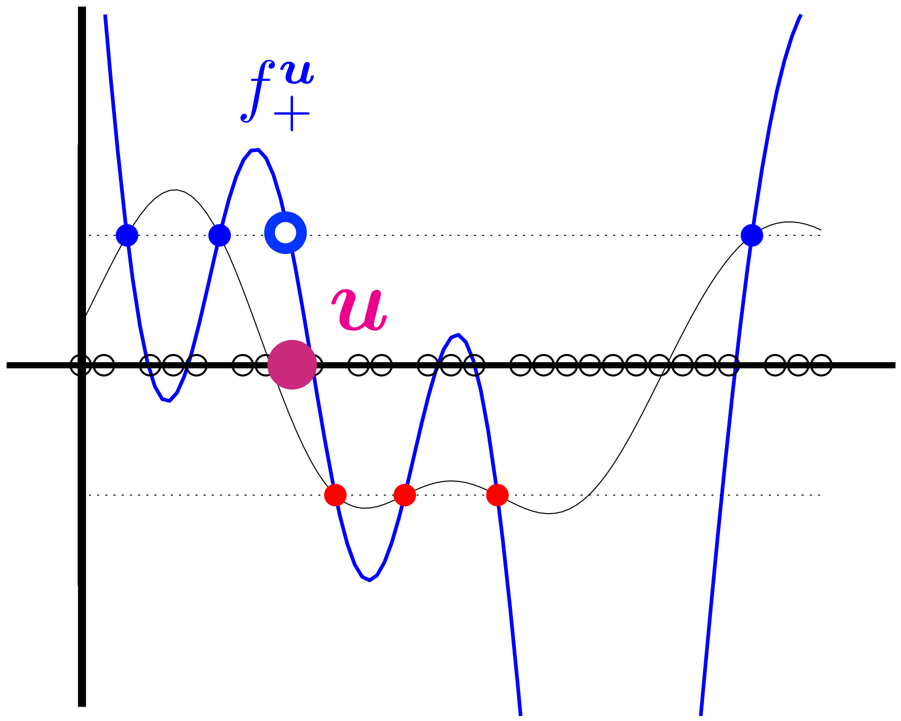


easy to interpolate, small $\|f_-^{\boldsymbol{u}}\|$

smoother

$\|\cdot\| \;=\;$ RKHS norm or norm of neural network weights

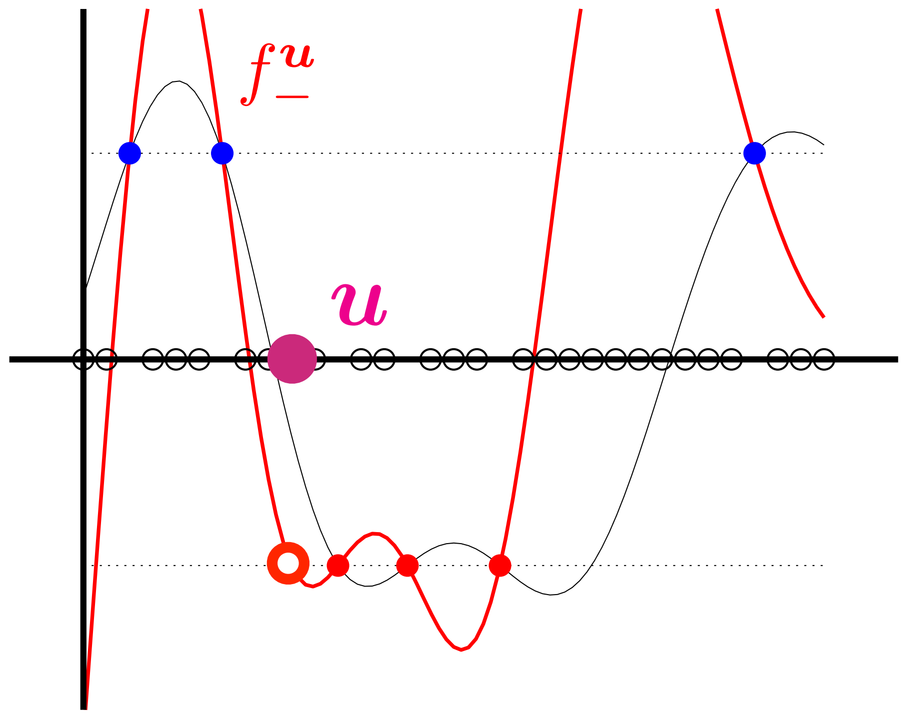
Active Learning in Overparameterized Setting



New example in between
oppositely labeled examples

difficult to interpolate, large $\|f_+^{\boldsymbol{u}}\|$

less smooth



difficult to interpolate, large $\|f_-^{\boldsymbol{u}}\|$

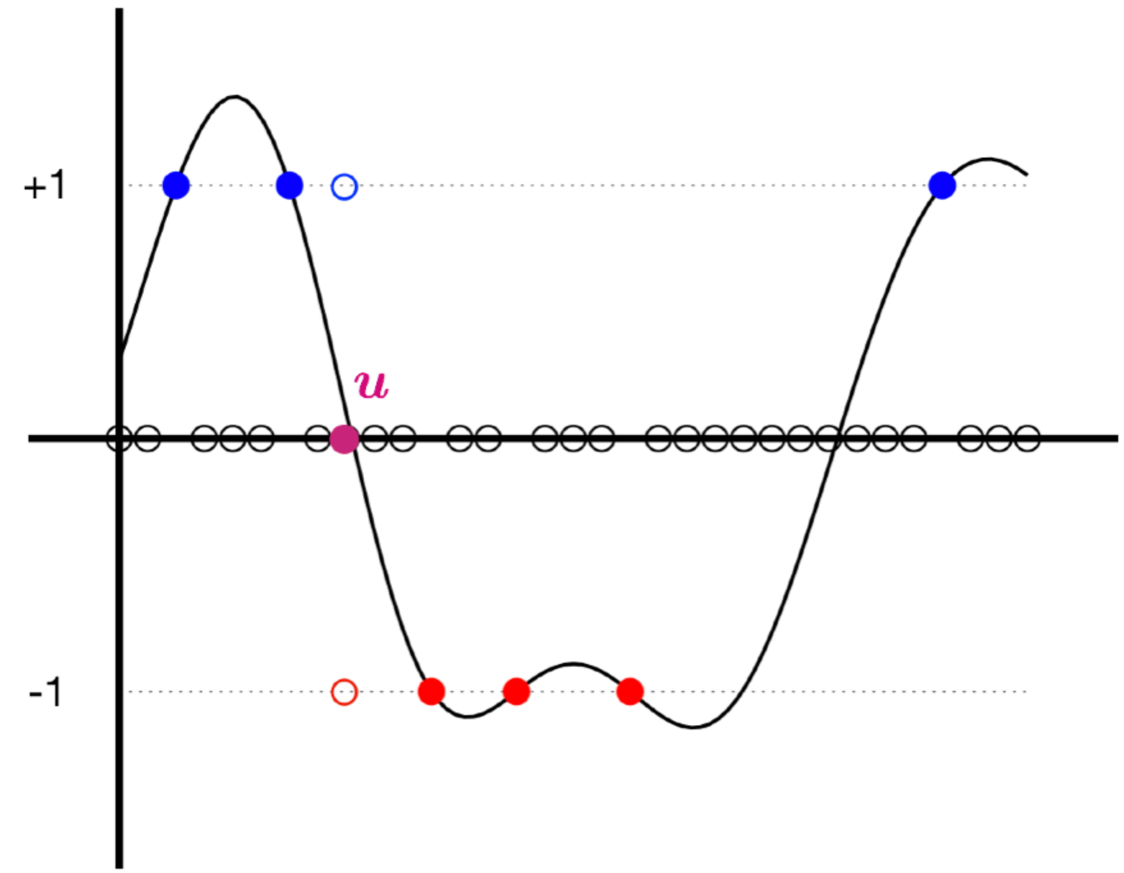
less smooth

$\|\cdot\|$ = RKHS norm or norm of neural network weights

Max-Min Sampling Criterion

Selection of next example to label

$$\boldsymbol{u}^\star \;=\; \arg\max_{\boldsymbol{u}\in U}\min\left\{\|f_-^{\boldsymbol{u}}\|\,,\;\|f_+^{\boldsymbol{u}}\|\right\}$$



Intuition: attacking the most challenging points in the input space first may eliminate the need to label other “easier” examples later

Mina Karzand and RN. “Active Learning in the Overparameterized and Interpolating Regime.” arXiv preprint arXiv:1905.12782 (2019).

Properties of Max-Min Sampling in RKHS

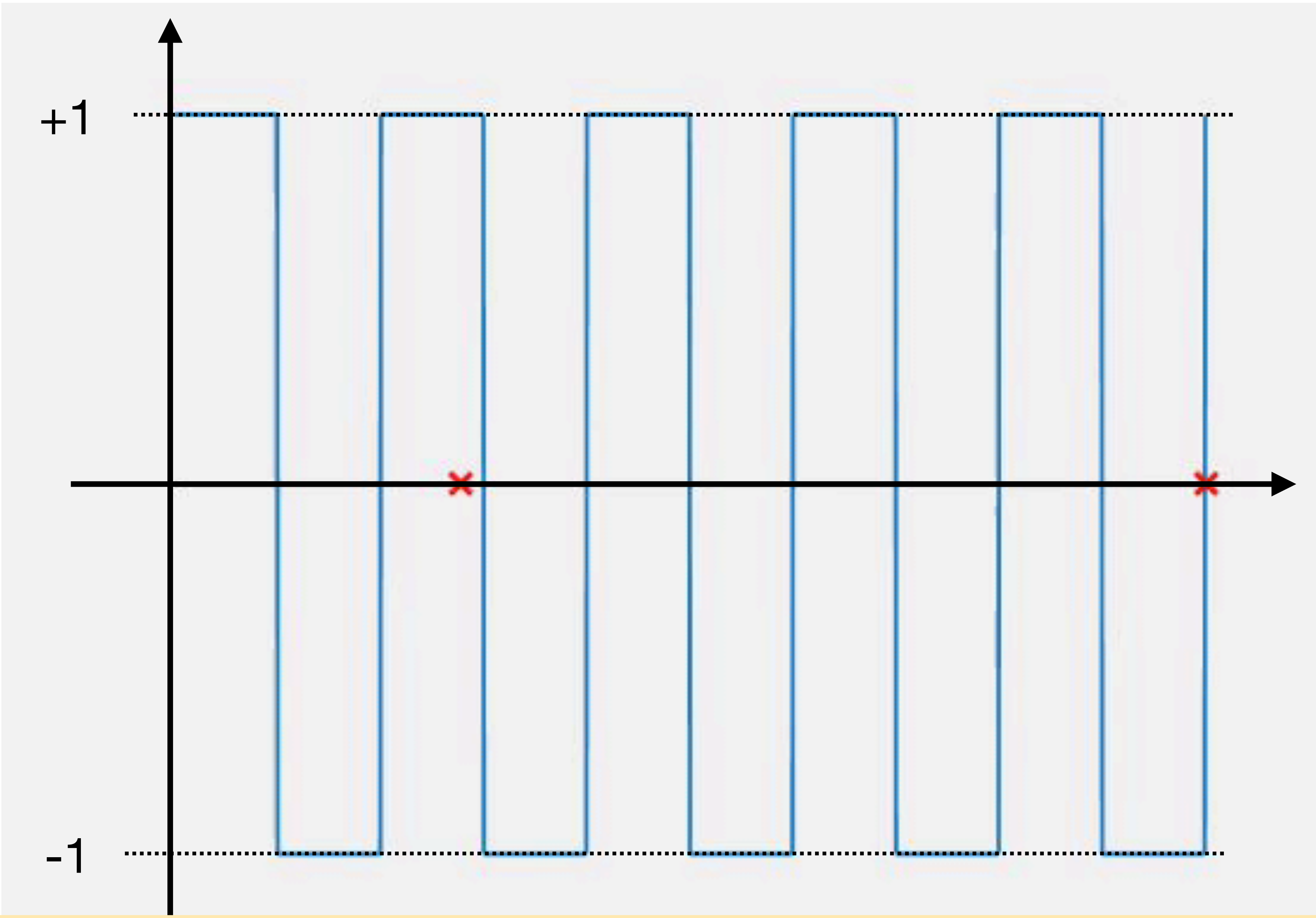
Assume that $f \in \mathcal{H}$, $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), and f is the minimum RKHS-norm interpolator of the labeled examples

$$\boldsymbol{u}^{\star} \;=\; \arg\max_{\boldsymbol{u}\in U} \min \left\{ \|f_{-}^{\boldsymbol{u}}\| \,,\; \|f_{+}^{\boldsymbol{u}}\| \right\}$$

Key Properties of RKHS Active Learner:

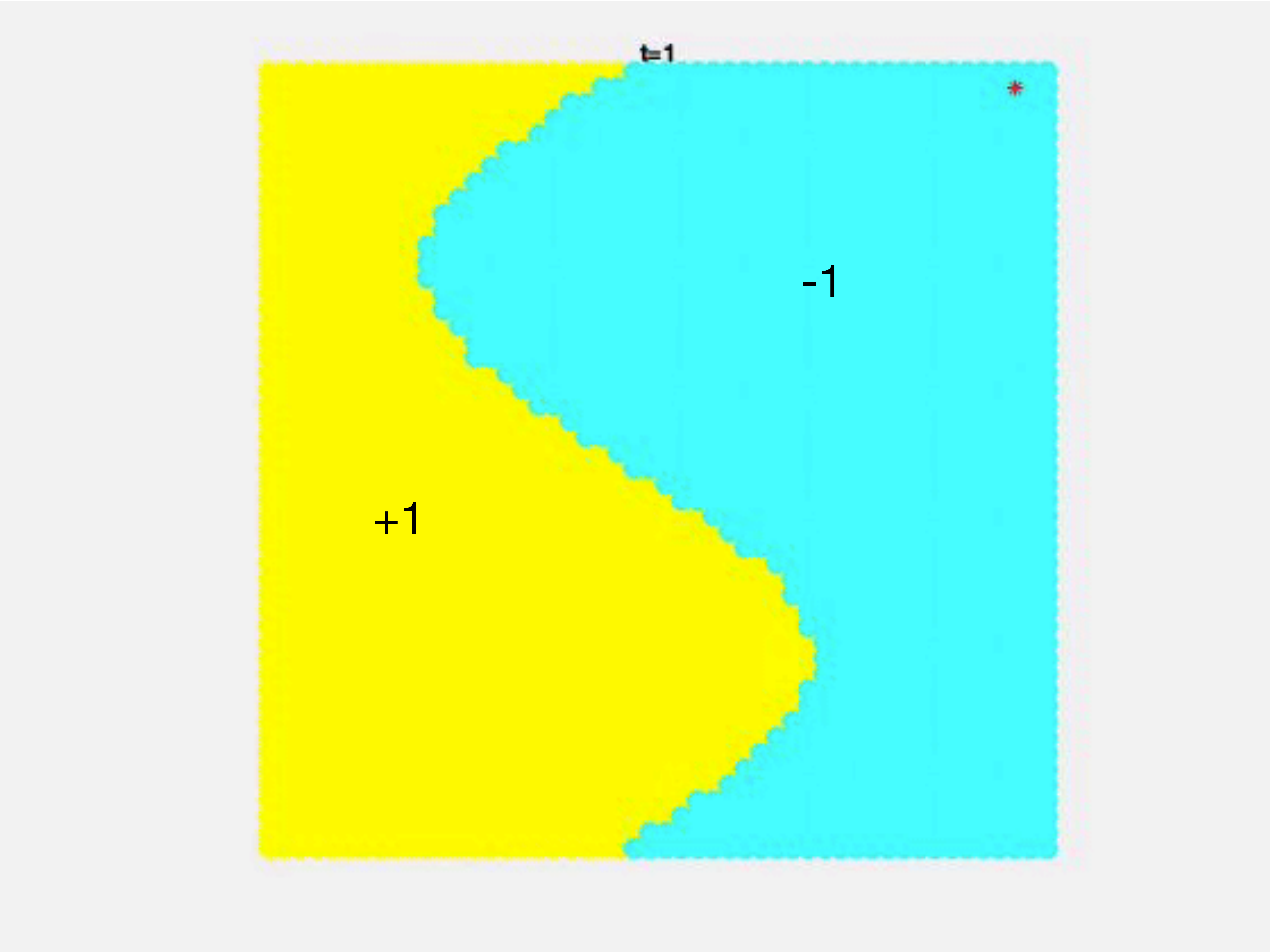
- Minimum norm labeling of new example $\boldsymbol{u}$ is given by sign of current interpolator f
- Selects samples near the current decision boundary *and* closest to oppositely labeled examples
- Yields optimal binary search behavior in one-dimension

Kernel Active Learner in One Dimension

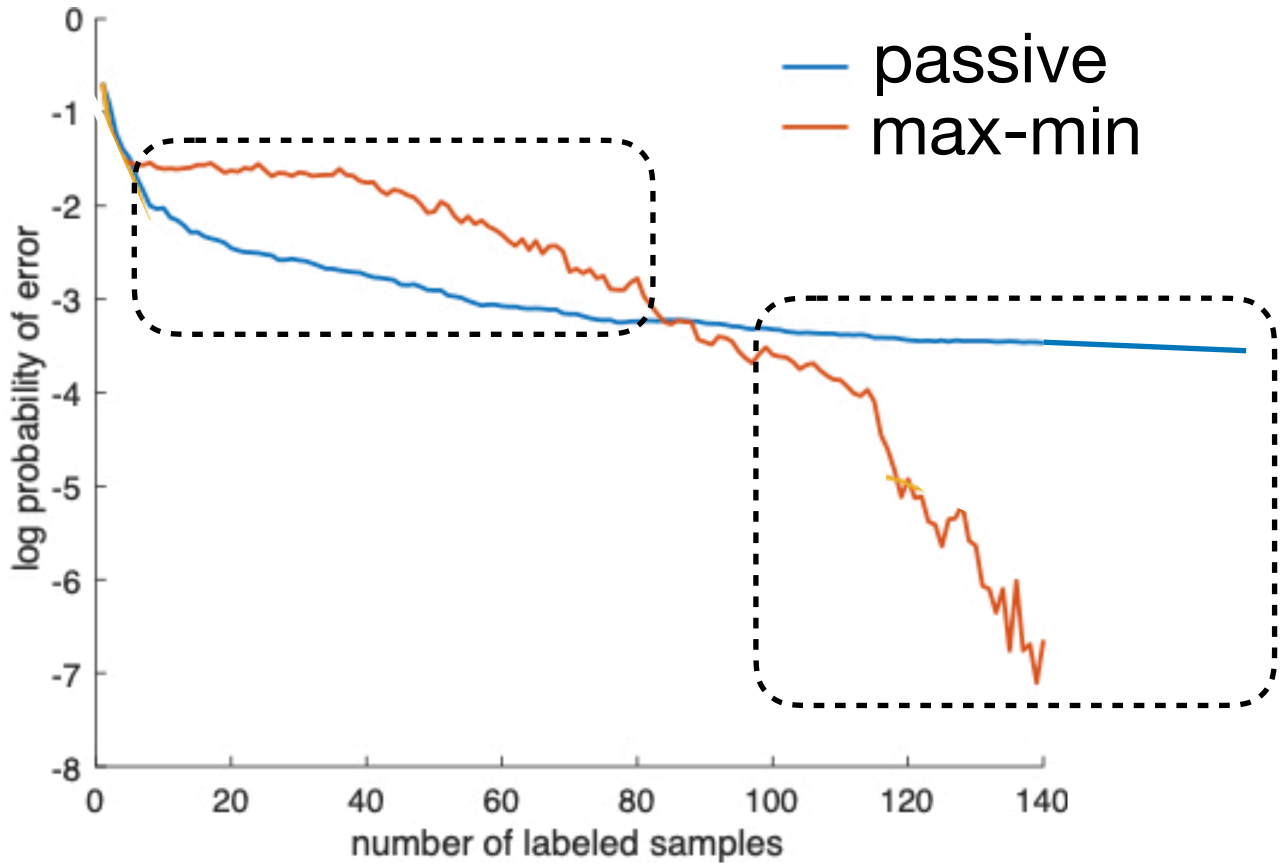


Theorem: Consider N points uniformly distributed in $[0, 1]$ and labeled according to piecewise constant binary-valued function $g(x)$ with k pieces. Then the Laplace kernel active learner perfectly predicts the labels of all N points after labeling $O(k \log N)$ examples.

Kernel Active Learner in Multiple Dimensions



Strength and Weakness of Max-Min Criterion



Limitation of max-min criterion:

- Can be too myopically focused on learning decision boundary
- RKHS norm is insensitive to data distribution

Data-Based Norm

Data-based Criterion:

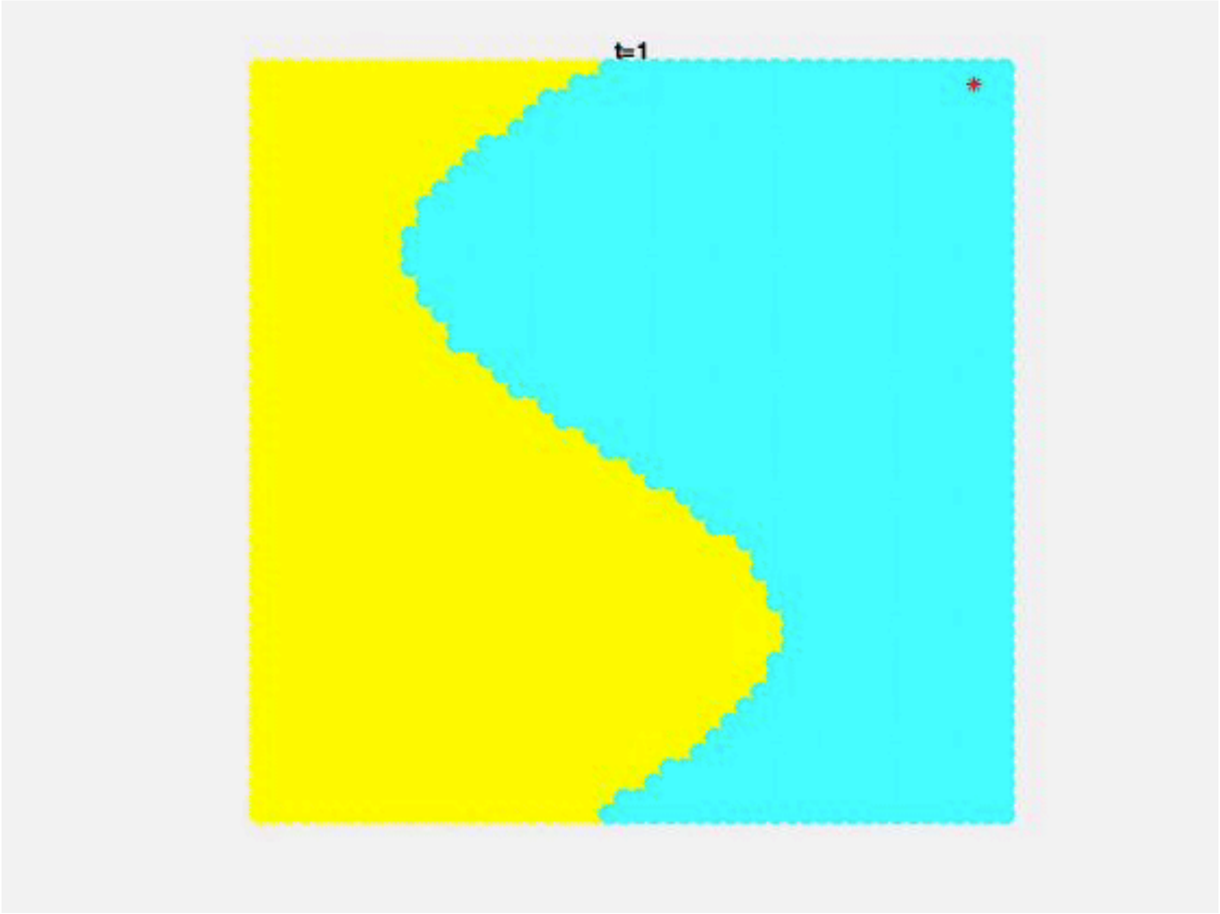
$$\begin{aligned} f &= \text{min RKHS norm interpolator of } \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \\ f^{\boldsymbol{u}} &= \text{min RKHS norm interpolator adding } \boldsymbol{u} \\ \boldsymbol{u}^* &= \arg\max_{\boldsymbol{u}\in\mathcal{U}} \sum_{x\in\mathcal{U}} \Big(f^{\boldsymbol{u}}(x) - f(x)\Big)^2 \end{aligned}$$

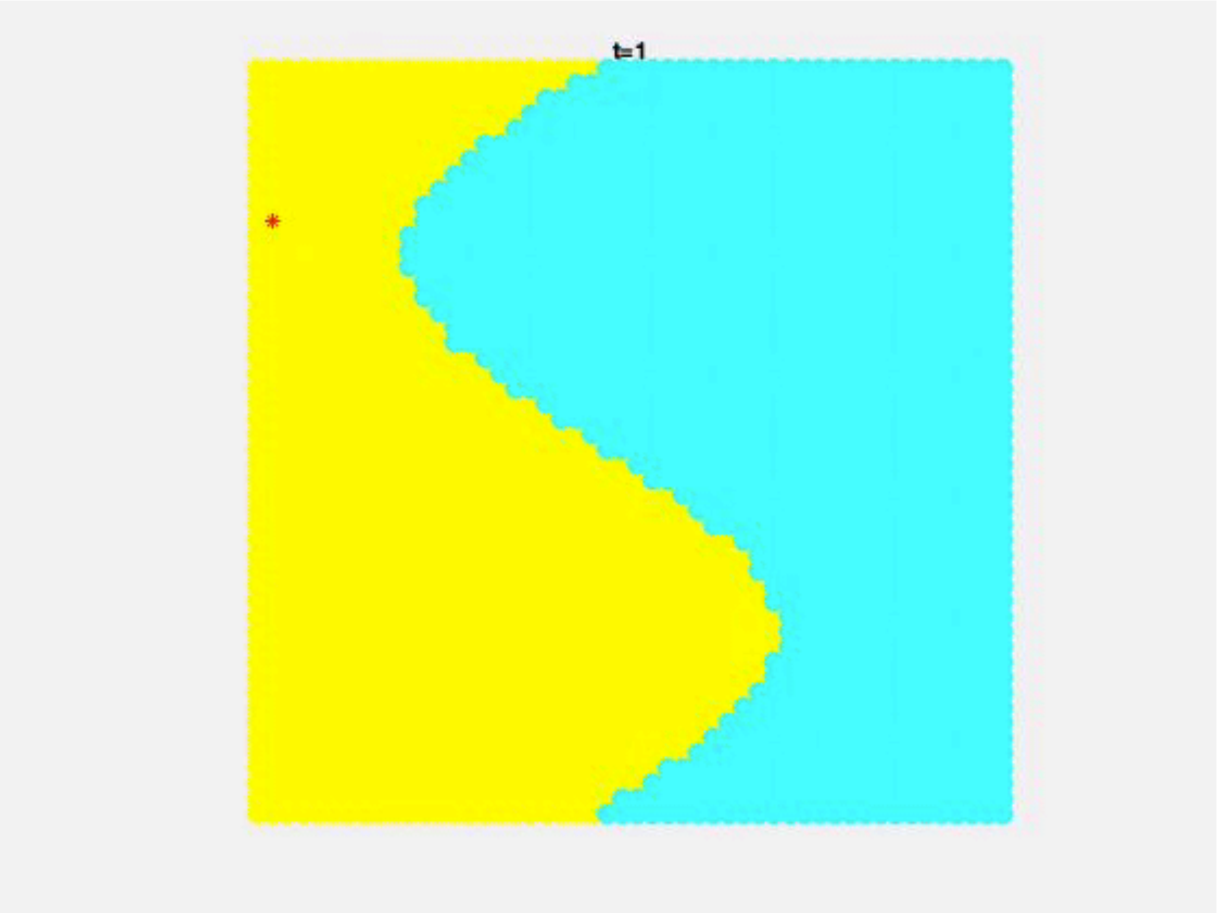
select new example that leads to greatest change
in interpolating function *on the dataset*

Max-Min RKHS norm vs. Data-based norm

$$\boldsymbol{u}^\star = \arg\max_{\boldsymbol{u}\in U}\min\left\{\|f_-^{\boldsymbol{u}}\|\,,\;\|f_+^{\boldsymbol{u}}\|\right\}$$

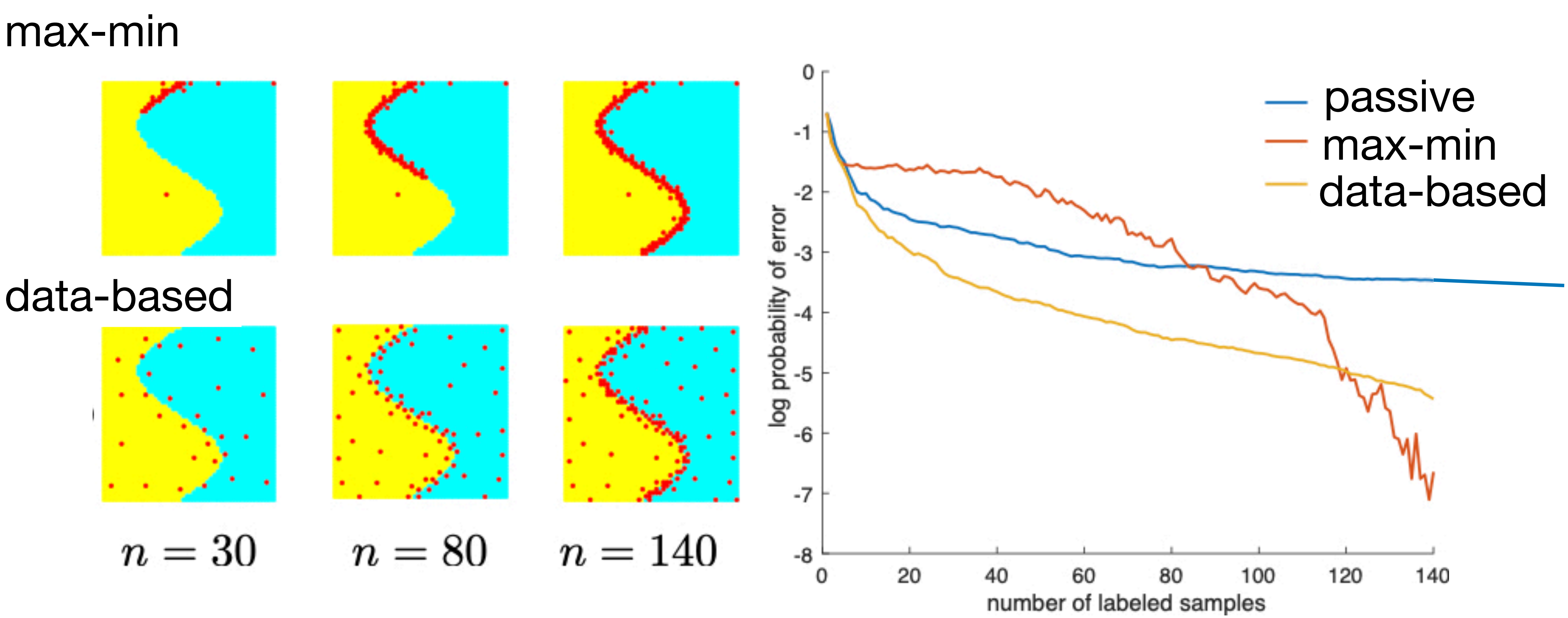
$$\boldsymbol{u}^\star = \arg\max_{\boldsymbol{u}\in\mathcal{U}}\sum_{x\in\mathcal{U}}\left(f^{\boldsymbol{u}}(x)-f(x)\right)^2$$





selection using on data-based norm strikes balance
between focusing boundary and exploring more globally

Min-Max and Data-Based Criteria

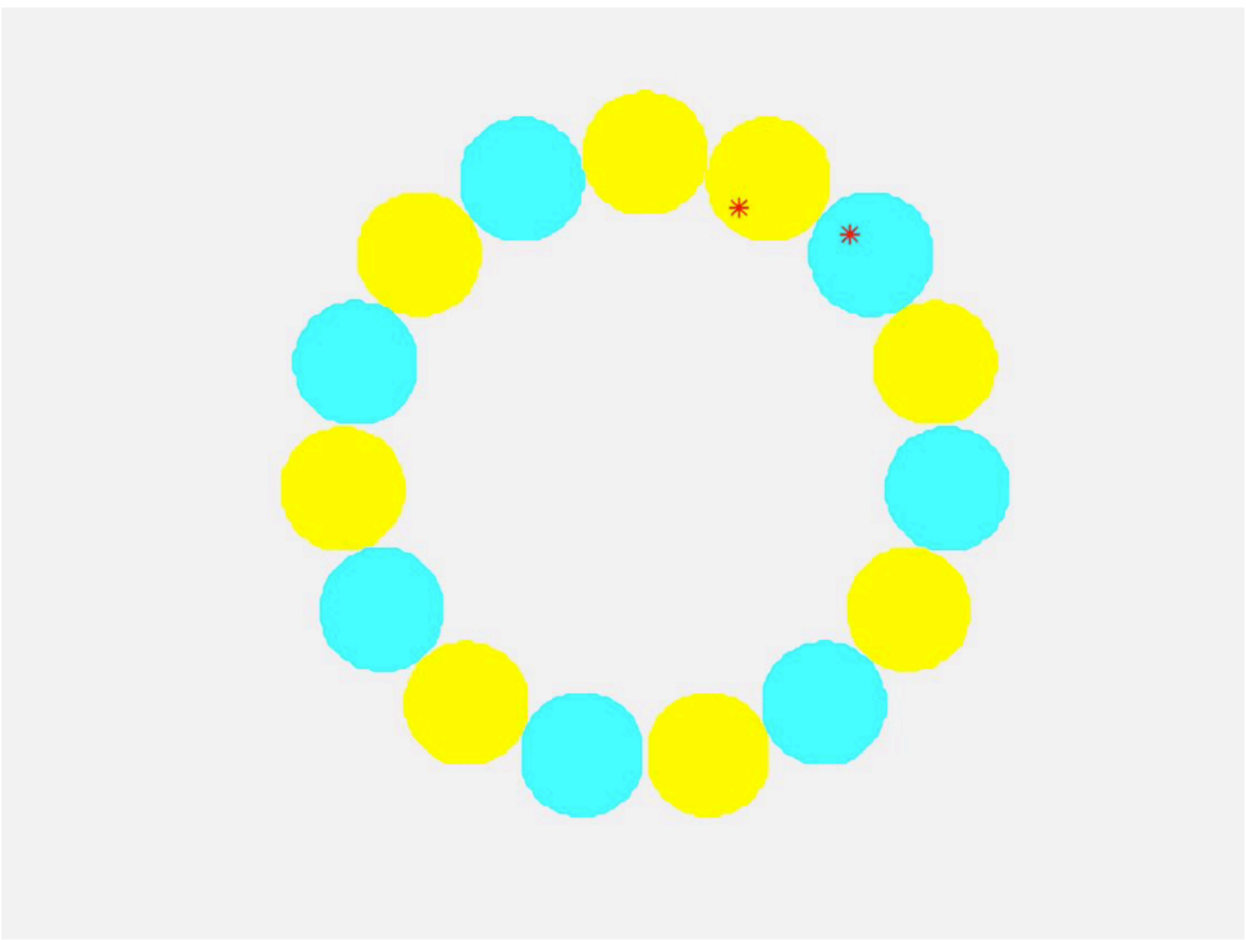


data-based criterion has more graceful error decay

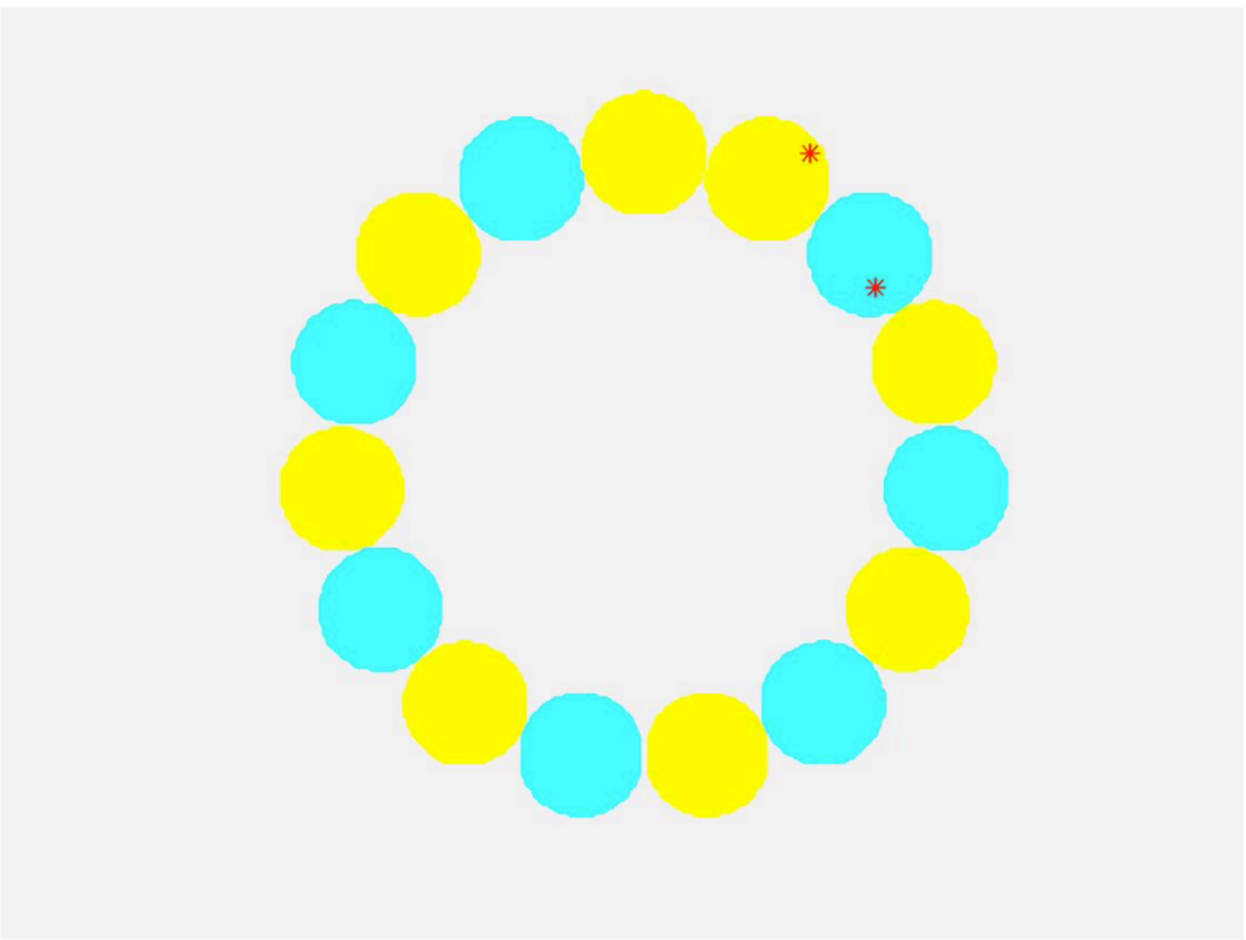
Cluster-Seeking Nature of Data-Based Criterion

$$\boldsymbol{u}^\star = \arg\max_{\boldsymbol{u}\in U} \min \left\{ \|f_-^{\boldsymbol{u}}\|,\ \|f_+^{\boldsymbol{u}}\| \right\}$$

$$\boldsymbol{u}^\star = \arg\max_{\boldsymbol{u}\in \mathcal{U}} \sum_{x\in\mathcal{U}} \left(f^{\boldsymbol{u}}(x) - f(x) \right)^2$$

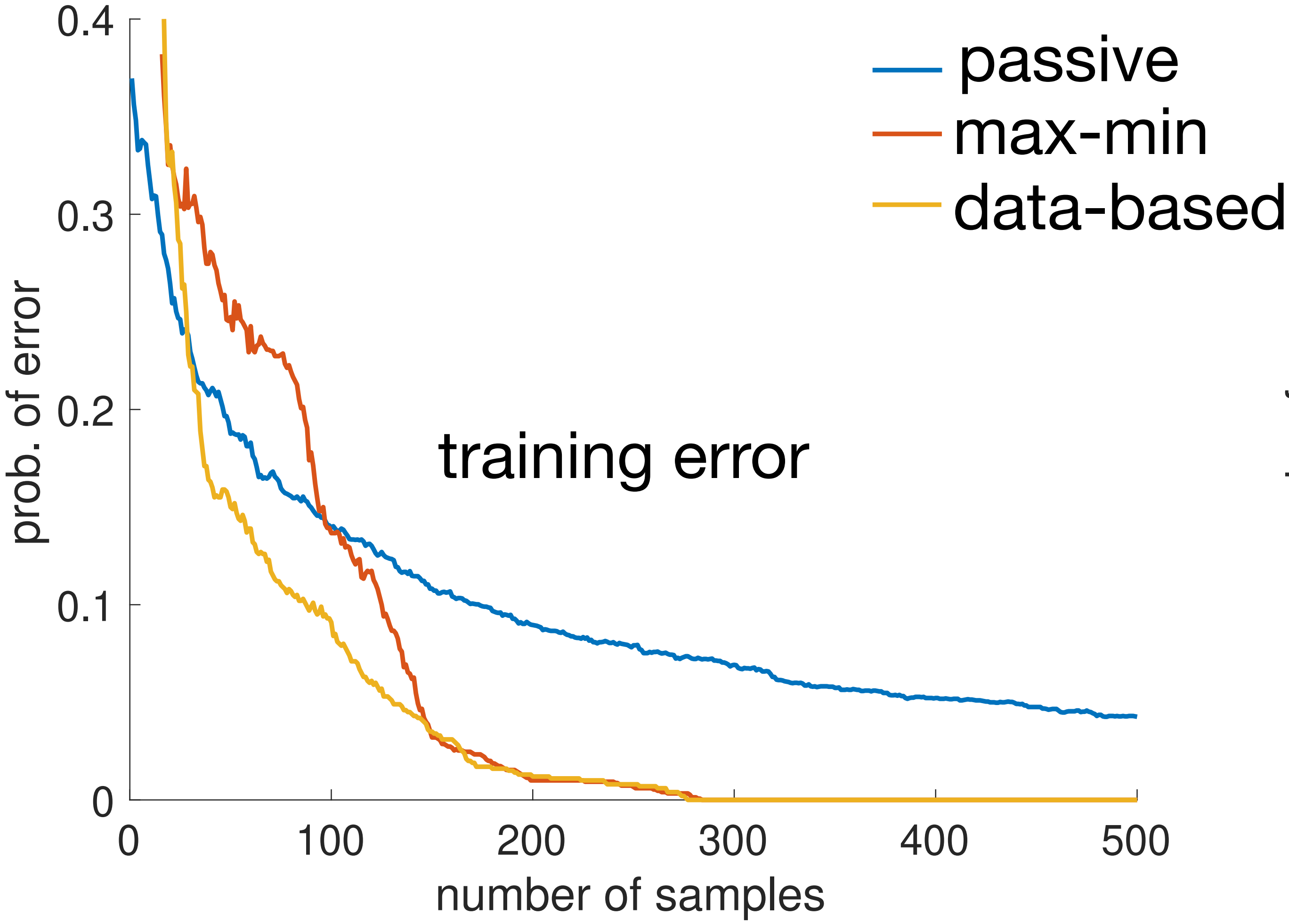


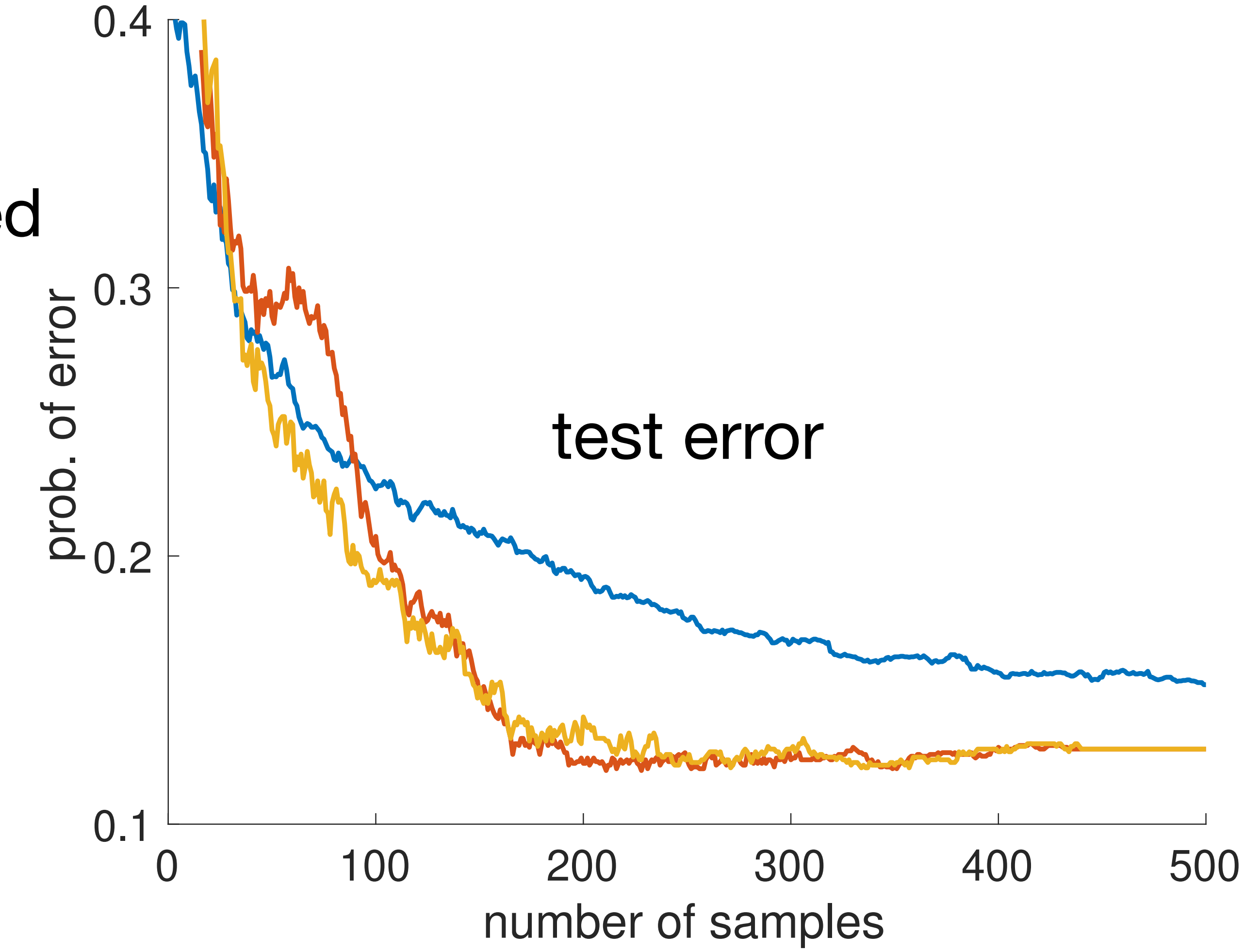
focuses more on finding boundaries



focuses more on finding clusters

MNIST Experiment using Laplace Kernel





Conclusions

- theory and methods of active learning are well-developed in the classical statistical learning framework (e.g., VC theory)
- classical theory may not be applicable in overparameterized regime
- new framework for active learning based on minimum norm interpolators shows promise in theory and practice, for both kernel machines and neural networks
- many opportunities to develop new theory for modern deep learning methods and new computationally efficient algorithms for active learning

Thanks!

slides: http://nowak.ece.wisc.edu/ActiveML.html

Recommended Reading

Dasgupta and Hsu. "Hierarchical sampling for active learning." *ICML* 2008

Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*. Vol. 3. 2003.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530* (2016).

Dasarathy, Gautam, Robert Nowak, and Xiaojin Zhu. "S2: An efficient graph based active learning algorithm with application to nonparametric classification." *Conference on Learning Theory*. 2015.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning and the bias-variance trade-off." *arXiv preprint arXiv:1812.11118* (2018).

Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research* 2.Nov (2001): 45-66.

Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." *arXiv preprint arXiv:1708.00489* (2017).

Savarese, Pedro, Itay Evron, Daniel Soudry, and Nathan Srebro. "How do infinite width bounded norm networks look in function space?." *arXiv preprint arXiv:1902.05040* (2019).

Mina Karzand and RN. "Active Learning in the Overparameterized and Interpolating Regime." *arXiv preprint arXiv:1905.12782* (2019).