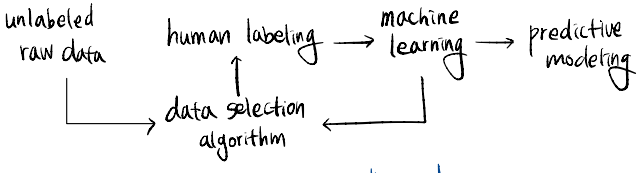
[Strategies for Active Machine Learning]

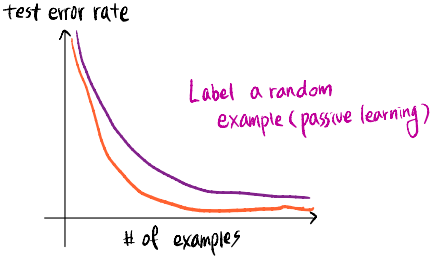
- Conventional (passive) machine learning

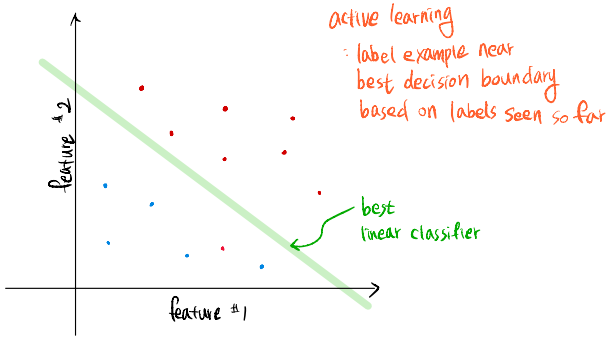
unlabeled raw data → human labeling → machine learning → predictive modeling

- Active machine learning

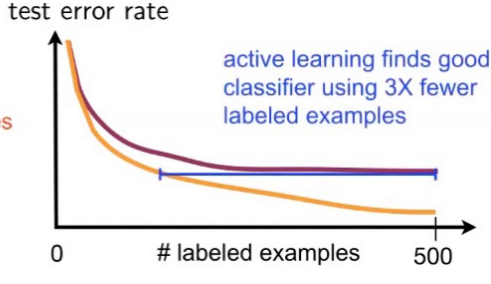


machine automatically and adaptively selects most informative data for labeling
– reduces time and effort of human supervisor(s)

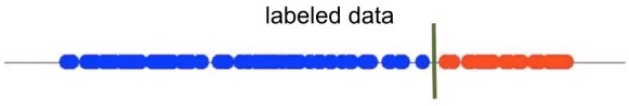


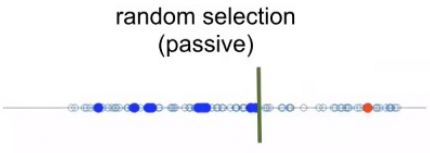


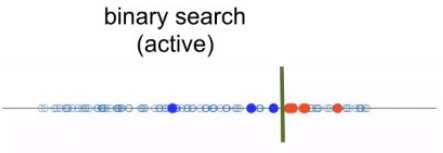
- cataracts classification based on 6182 EHR features
- linear classifiers



Learning a 1-D Classifier







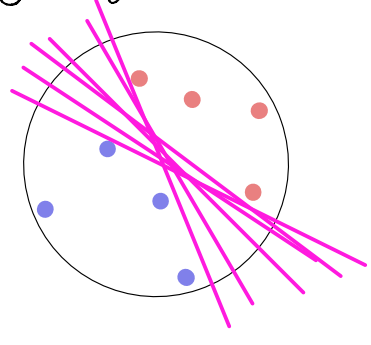
binary search quickly finds **decision boundary**

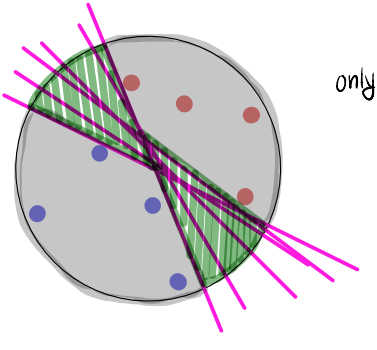
passive: err $\sim n^{-1}$
active : err $\sim 2^{-n}$

} can be generalized to handle any situation where optimal classifier is a single threshold

Linear Classifiers in Multiple Dimensions

Consider examples uniformly distributed in unit ball and linear classifiers passing through origin.





only label points in RoD

Data-consistent classifiers agree in gray region, but disagree in green region, the region of disagreement (RoD)

Linear Classifiers in Multiple Dimensions

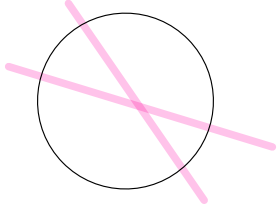
Consider examples uniformly distributed in unit ball and linear classifiers passing through origin.

Data-consistent classifiers agree in gray region, but disagree in green region,

the Region of Disagreement (RoD)

Only label points in RoD

Iterate the process to efficiently learn a good classifier.



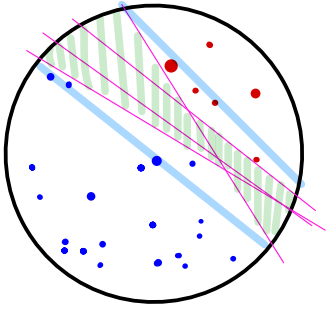
ϵ-optimal classifier requires

- passive: $O(d/\epsilon)$ labeled examples
- active: $O(d \log \frac{1}{\epsilon})$ labeled examples

Challenge: Shape of RoD above is deceptively simple...
much more complicated in higher dimensions

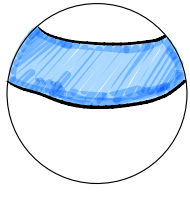
[Approximating Region of Disagreement]

general form of a linear classifier: $w^T x + b$



Computing RoD may be computationally expensive

(combinatorial object)



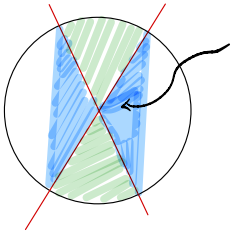
Approch: Approximate RoD with a sphere segment.

[Homogeneous Linear Classifiers (passing through origin)]

Property of high-dimensional ball : most volume near equator

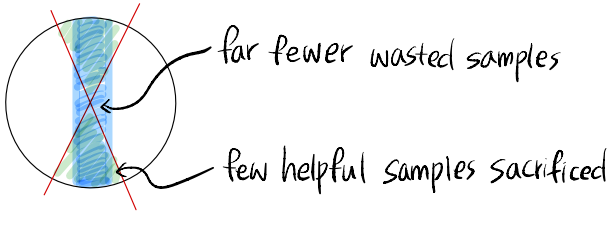
Problem : Overbounding RoD may be wasteful

⇒ active learning using overbound requires $O(d^{3/2} \log(1/\epsilon))$ labeled examples



more aggressive localization is necessary to reduce wasteful labeling

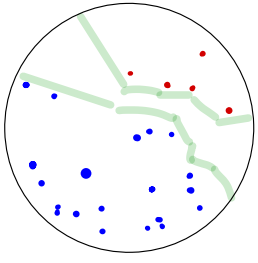
⇒ aggressive active learning requires $O(d \log(1/\epsilon))$ labeled examples



Caution : Two-dimensional RoD is deceptively simple...
more complex in higher dimensions

[General Linear Classifiers]

General form of a linear classifier : $w^T x + b$



relevant in real-world applications

- optimal separator does not pass through origin in general
- classes are often imbalanced
  (e.g., more healthy people than diseased)

Challenges

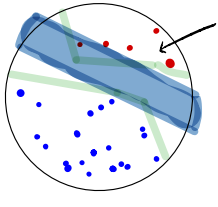
- RoD can be much complicated, even in two dimensions.
- Spherical caps are exponentially smaller in volume than segments through equator.

[Approximating RoD for General Linear Classifiers]

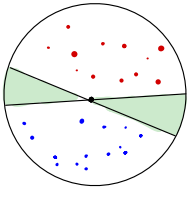
How to approximate RoD with spherical segment?

general linear classifier

$w^T x + b$



problem:
exponentially less
volume in sphere cap



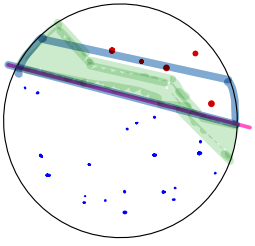
much simpler in case of
homogeneous classifiers
(pass through origin)

- can disproportionately favor majority class
- potential huge waste in sampling near equator

Spherical segment further from equator can better balance classes and reduces waste

Solution: construct spherical segment based on data-consistent classifier closest to the origin, the maximum-volume separator

maximum-volume separator



assume learning algorithm initialized with at least one example from each class.

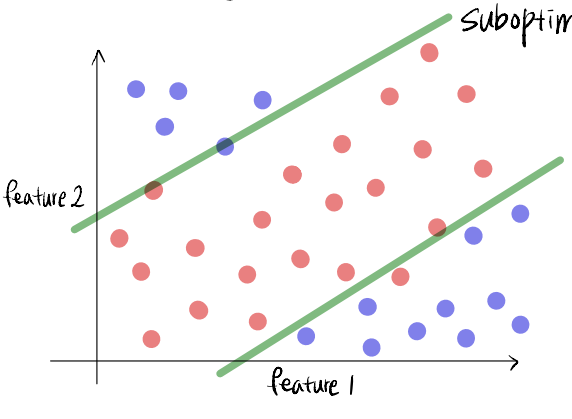
ϵ-optimal classifier requires

- passive: $O(d/\epsilon)$ labeled examples
- active: $O(d \log \frac{1}{\epsilon})$ labeled examples

Open problem: efficient methods to find maximum-volume separator

$$\min_{b,\ \|w\|=1} b^2 \text{ such that } y_i(w^T x_i + b) \geq 0,\ \forall i$$

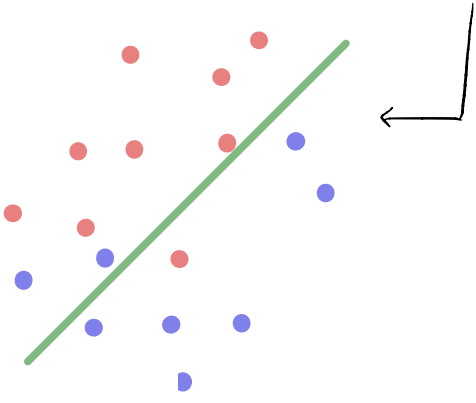
Active Learning Can Breakdown When Models Are Wrong

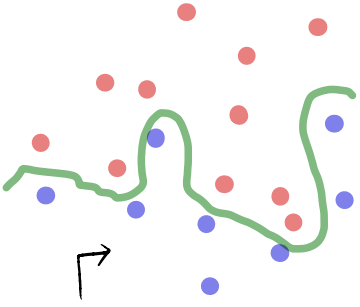


Model Bias: Standard active learning algorithms perform no better than passive because model does not match reality, and may even converge to suboptimal solutions

Nonparametric Active Learning

active learning suffers when models are misspecified



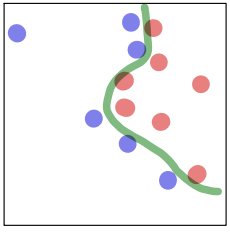


nonparametric models can handle arbitrary learning problems

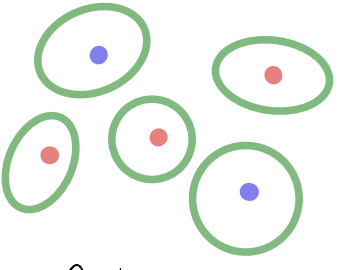
past theory work indicates potential, but does not provide practical nonparametric active learning algorithms.

Two Faces of Nonparametric Active Learning

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

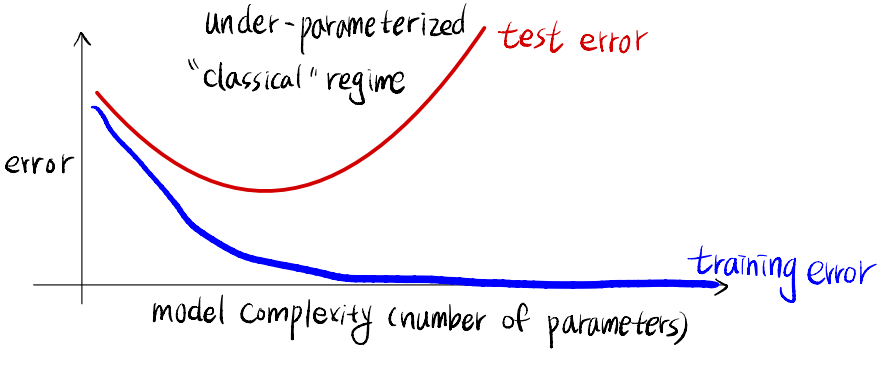


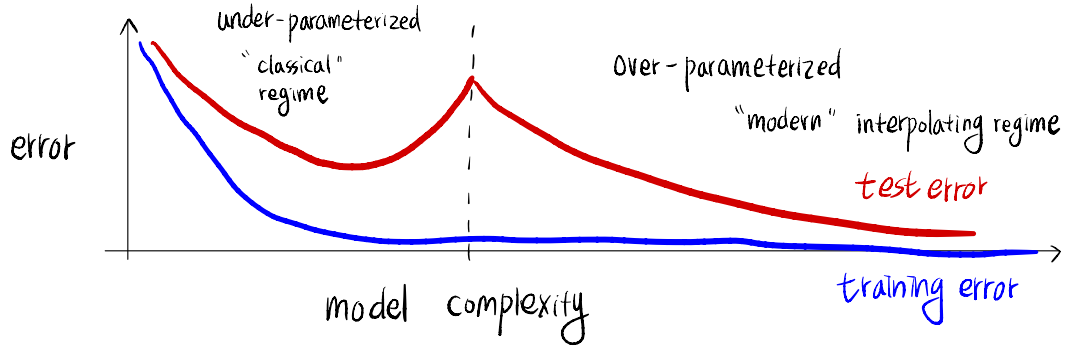
label examples close to estimated decision boundary



find clusters in unlabeled data and label one representative each

Generalization Error in Function Space





Key idea: don't obsess about number of parameters, focus on function space $\mathcal{F}$ associated with overparameterized models

$\Rightarrow$ minimize $\|f\|_{\mathcal{F}}$ subject to fitting data

MaxiMin Active Learning Heuristic

Selection of next example to label

$$u^* = \arg \max_{u \in U} \min \left\{ \|f_-^u\|_{\mathcal{F}}, \|f_+^u\|_{\mathcal{F}} \right\}$$

$u \in U$: unlabeled examples

norm of new learned function depends on choice of label +1 or -1

label example that maximizes norm of "optimistic" new function learned by adding it to training set.

Intuition: attacking the most challenging examples first may eliminate the need to label other "easier" examples later

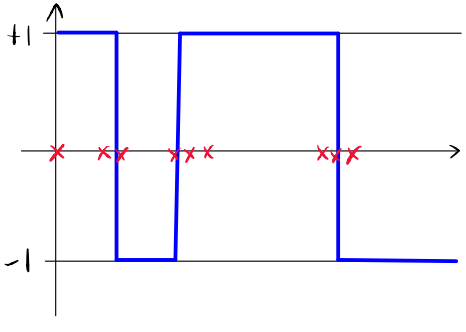
Theory: automatically selects an example near the current decision boundary and closest to oppositely labeled examples.

MaxiMin Active Learner Performs Bisection

MaxiMin Criterion

$$u^* = \arg\max_{u \in U} \min \{ \|f^u_-\|, \|f^u_+\| \}$$

selects next example near midpoint of close and oppositely labeled examples.



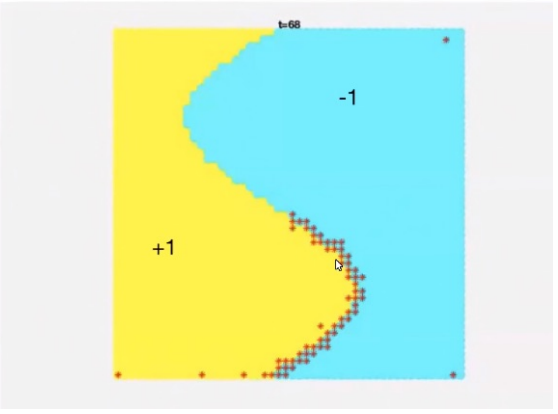
← selects near decision boundary

target function

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

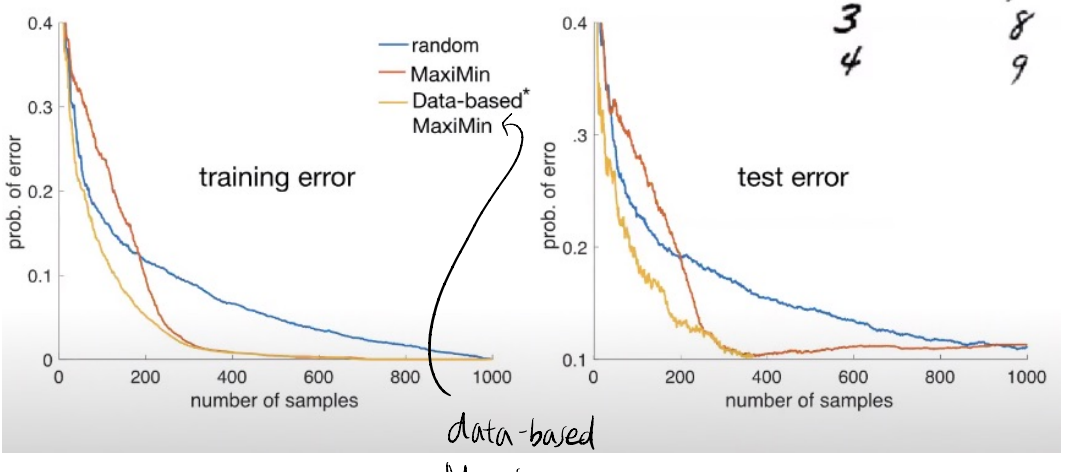
MaxiMin Sampling in Multiple Dimensions

(unlabeled) examples
uniformly distributed



MNIST Experiment using Laplace Kernel

0 1 2 3 4 vs. 5 6 7 8 9



data-based MaxiMin is a variant that is sensitive to the distribution of data

MaxiMin Active Learner Performs Bisection

Laplace Kernel case:

- RKHS representer theorem shows that optimal interpolator is superposition of kernel representers.
- Kernel Gram matrix has special block-diagonal structure.
- increase in RHKS norm maximal at midpoint between closest oppositely labeled points.

ReLU Network Case:

- *neural network representer theorem* shows that optimal neural network is equivalent to classic linear spline
- network weight norm corresponds to *total variation norm* of second-derivative of interpolator
- total variation norm increases the most when new example is added at midpoint between closest oppositely labeled points

Representer Theorems

A *representer theorem* tells us that the solutions to certain learning problems in *infinite-dimensional function spaces* can be expressed in terms of *finite-dimensional parametric functions*

Grace Wahba

Smoothing Splines, Kimeldorf and Wahba (1970)

Kernel Machines: generalization to wide classes of machine learning problems in Reproducing Kernel Hilbert Spaces (1990-present)

Banach Spaces: generalized splines, reproducing kernel Banach spaces, TV-regularization & atomic representations (2010-present)

Classical Representer Theorem

RKHS Representer Theorem: Let $\mathcal{F}$ be a Reproducing Kernel Hilbert Space with norm $\|\cdot\|_{\mathcal{F}}$ and kernel k. Then for any training dataset $\{\boldsymbol{x}_i, y_i\}$, any loss function ℓ, and any $f \in \mathcal{F}$ minimizing the regularized empirical risk

$$\sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i)) \; + \; \lambda \|f\|_{\mathcal{F}} \; , \; \lambda > 0$$

admits a representation of the form

$$f(\boldsymbol{x}) \; = \; \sum_{i=1}^{n} \alpha_i \, k(\boldsymbol{x}, \boldsymbol{x}_i) \; , \; \text{for some } \alpha_1, \ldots, \alpha_n \in \mathbb{R}$$

Neural Network Representer Theorem

Rahul Parhi

Is there a representer theorem for neural networks?

Answer: Yes! But not in a Hilbert Space

Neural Network Representer Theorem (Parhi & N, 2020): There is nonparametric (Banach) space $\mathcal{F}$ with norm $\|\cdot\|_{\mathcal{F}}$ such that for any training dataset $\{\boldsymbol{x}_i, y_i\}$ and any convex and coercive loss function ℓ, there exists a solution to

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i)) \; + \; \lambda \|f\|_{\mathcal{F}}$$

with a representation in the form of a single hidden-layer neural network

$$f(\boldsymbol{x}) \; = \; \sum_{k=1}^{K} v_k \, \phi(\boldsymbol{w}_k^T \boldsymbol{x} - b_k) + c(\boldsymbol{x}) \qquad K \leq n$$

Banach Spaces and Neural Networks

Neural Network Banach Spaces (informal):

$\|f\|_{\mathcal{F}_m}$ measures "sparsity" of f in Radon domain

$$\mathcal{F}_m \; = \; \left\{ f : \mathbb{R}^d \to \mathbb{R}, \; \left\| \Lambda^{d-1} R \, \Delta^m f \right\|_{L^1} < \infty \right\} \qquad m = 1, 2, \ldots$$

filtered Radon transform — Laplacian operator

$$f(\boldsymbol{x}) \; = \; \sum_{k=1}^{K} v_k \, \phi_m(\boldsymbol{w}_k^T \boldsymbol{x} - b_k) + c(\boldsymbol{x})$$

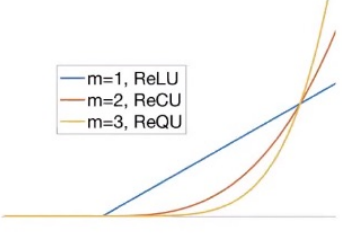
minimizes $\left\| \Lambda^{d-1} R \, \Delta^m f \right\|_{L^1}$ subject to data-fitting

$\{v_k, \boldsymbol{w}_k\}$ = network weights

c = generalized bias term (polynomial)

$\phi_m(\cdot) = \dfrac{\max\{\cdot, 0\}^{2m-1}}{(2m-1)!}, \; m = 1, 2, \ldots,$

ReLU / truncated power function



Weight Norms and Generalization Bounds

For any neural network $f(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \phi_m(\boldsymbol{w}_k^T \boldsymbol{x} - b_k)$ its $\mathcal{F}_m$-norm is

$$\|f\|_{\mathcal{F}_m} := \left\| \Lambda^{d-1} R \, \Delta^m f \right\|_{L^1} = \sum_{k=1}^{K} |v_k| \, \|\boldsymbol{w}_k\|_2^{2m-1}$$

a network "path" norm

Neural Network Generalization Theorem: The $\|\cdot\|_{\mathcal{F}_m}$ norm controls the *Rademacher complexity* of neural nets. Let $\widehat{f}$ be a solution to

$$\min_{f \in \mathcal{F}_m} \sum_{i=1}^{n} \ell(y_i f(x_i)) \text{ subject to } \|f\|_{\mathcal{F}_m} \leq \mathbf{B}$$

where ℓ is a Lipschitz loss. Then

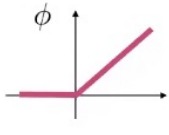
$$\text{test-error}(\widehat{f}) \leq \text{train-error}(\widehat{f}) + O\Big(\frac{\mathbf{B}}{\sqrt{n}}\Big)$$

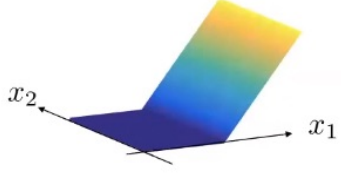
Why the Radon Transform?

neurons are **ridge functions:**

$$\phi(w_1 x_1 + w_2 x_2 - b)$$

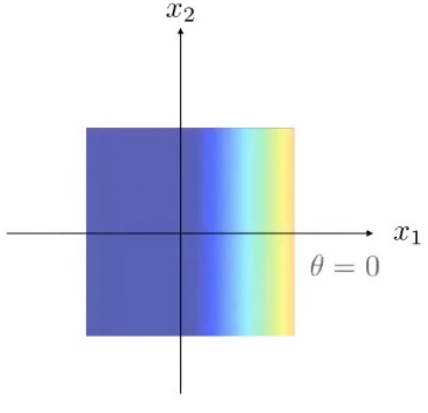
where $\phi : \mathbb{R} \to \mathbb{R}$
is "activation" function





ridge functions are parameterized by *angle* θ and *offset* b

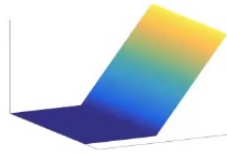
$$\boldsymbol{w} = (\sin\theta, \cos\theta)$$



Ridge Functions and Radon Transform

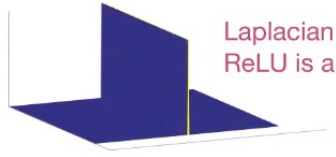
single-hidden-layer neural nets are superpositions of ridge functions

ReLU function: $\phi_1(w_1 x_1 + w_2 x_2 - b_0)$



ReLU is a special ridge

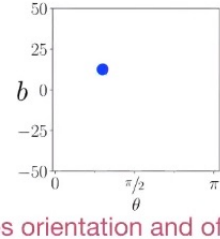
Laplacian of ReLU: $\Delta\phi = \frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2}$



Laplacian of ReLU is a "line"

filtered Radon transform:

$$\Lambda R \Delta\phi = \delta(b - b_0)\,\delta\big(\theta - \tan^{-1}\tfrac{w_1}{w_2}\big)$$



filtered Radon transform of "line" is an impulse

indicates orientation and offset of ridge

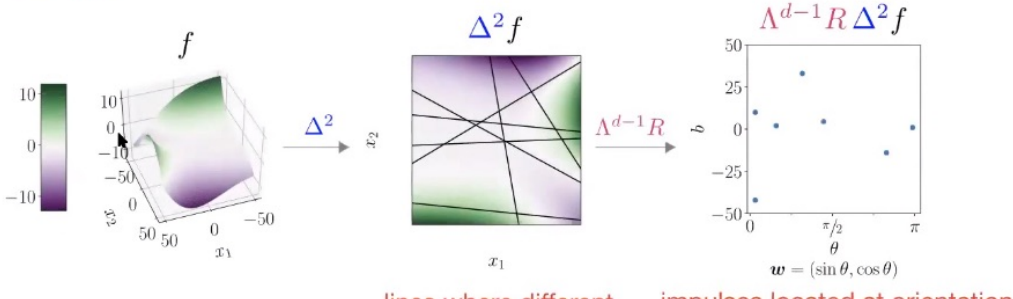
Single-Hidden-Layer Neural Networks

Differentiation: Δ^m annihilates polynomial surfaces, leaving only linear boundaries of activation thresholds

Radon Transform: $\Lambda^{d-1} R$ "extracts" orientations and offsets of each neuron, producing impulses in Radon domain

$\left\| \Lambda^{d-1} R \Delta^m f \right\|_{L^1}$ measures sparsity in Radon domain

Example: 7 neuron network with Rectified Cubic Units (ReCU)



lines where different neurons turn on/off

impulses located at orientation and offset of each neuron

Proving the Neural Network Representer Theorem

Single-hidden-layer networks with $K \leq n$ neurons are solutions to

$$\min_{f \in \mathcal{F}_m} \sum_{i=1}^{n} \ell\big(y_i, f(\boldsymbol{x}_i)\big) \; + \; \lambda \|f\|_{\mathcal{F}_m}$$

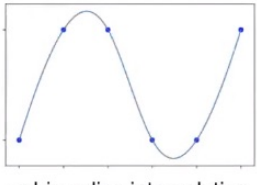
- properties of operator $\mathcal{R}_m = \Lambda^{d-1} R \, \Delta^m$: *Green's functions* are truncated-power ridges and null space is polynomials of degree $< 2m$
- every $f \in \mathcal{F}_m$ can be expressed in terms of finite measure μ i.e., $\mathcal{R}_m f = \mu$ and $f = \mathcal{R}_m^+ \mu + c$, for some polynomial c
- minimizing $\|\mathcal{R}_m f\|_{L^1}$ equivalent to minimizing $\|\mu\|_{TV}$, total-variation norm in sense of measures
- $\min \|\mu\|_{TV}$ subject to finite number of linear constraints is classical measure recovery problem – solution is superposition of Dirac impulses
- applying pseudoinverse $\mathcal{R}_m^+$ to solution yields a superposition of ridge functions, i.e., a single-hidden-layer neural net

Take-Away Messages

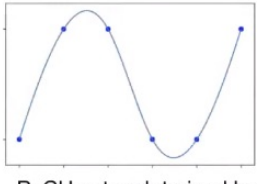
- nonparametric learning problems with $\|\cdot\|_{\mathcal{F}_m}$-regularization have *sparse* atomic solutions $\Rightarrow$ neural networks with truncated power activations
- $\|\cdot\|_{\mathcal{F}_m}$ controls Rademacher complexity and generalization error
- adding a training example increases $\|\cdot\|_{\mathcal{F}_m}$ and requires adding a neuron

key idea behind active learning heuristic

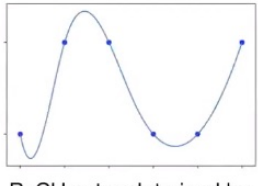
- $\|\cdot\|_{\mathcal{F}_m}$-regularization is equivalent to forms of "weight decay" in SGD
- one-dimensional case reduces to classical polynomial splines



cubic spline interpolation



ReCU network trained by SGD with weight decay



ReCU network trained by SGD w/o weight decay

Conclusions

- theory and methods of active learning for linear classifiers now well understood, but computational challenges persist

  open problem: efficient methods for computing maximin-volume separator

- classical theory inappropriate for understanding modern (nonparametric/ overparameterized) machine learning systems
- new framework for nonparametric active learning based on minimum norm solutions in appropriate function spaces shows promise in theory and practice

  open problems:
  - efficient methods for maximin sample selection
  - representer-like theorem for deep networks

Karzand and N. "Active learning of general linear separators" *in preparation*

Karzand and N. "MaxiMin Active Learning in Overparameterized Model Classes." *IEEE Journal on Selected Areas in Information Theory* (2020)

Parhi and N. "Neural Networks, Ridge Splines, and TV Regularization in the Radon Domain." *arXiv preprint arXiv:2006.05626* (2020).