Greedy Convex Ensemble

Tan Nguyen (Queensland University of Technology)

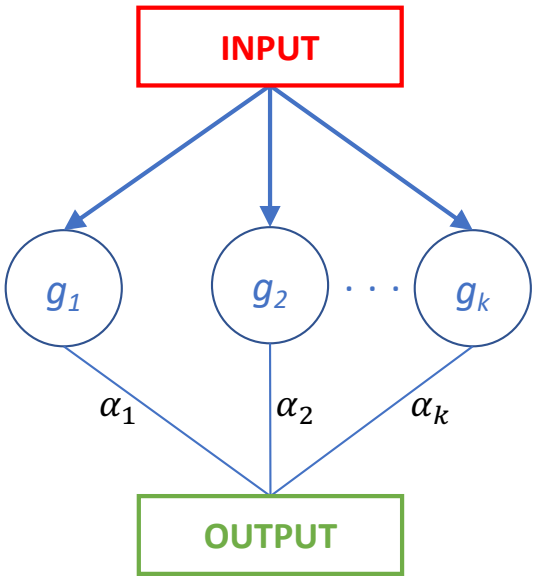
Nan Ye (University of Queensland)

Peter Bartlett (UC Berkeley)

December 15, 2020

The Problem

"Learning a convex ensemble of basis models"



Given some set $\mathcal{G}$ of basis models.

Linear hull, $\text{lin}(\mathcal{G})$, is the set of all possible linear combinations of models in $\mathcal{G}$

- $\text{lin}(\mathcal{G})$ of even simple basis models is an universal approximator

Convex hull, $\text{co}(\mathcal{G})$, is the set of all possible convex combinations of models in $\mathcal{G}$

- Convex hull is a subset of the linear hull
- Theoretical: Capacity? Generalization?
- Empirical: How to learn from a convex hull? Any advantages?

Capacity of Convex Hulls

Proposition 2: For linear threshold basis models,

$$\text{i.e. } \mathcal{G} = \{\mathbb{I}(\theta^\top x \geq t) : \theta \in \mathbf{R}^d, t \in \mathbf{R}\},$$

- Linear hulls: infinite pseudodimension and Rademacher complexity
- Convex hulls: infinite pseudodimension, finite Rademacher complexity

Implication:

- Linear hulls: unbounded capacity, thus prone to overfitting.
- Convex hulls: rich but bounded capacity, can be seen as a regularized version of linear hulls.

Generalization Bound

Theorem 2: With probability at least $1-\delta$, for all $f \in \text{co}(\mathcal{G})$,

$$R(f) - R(f^*) \leq R_n(f) - R_n(f^*) + \frac{c}{\sqrt{n}},$$

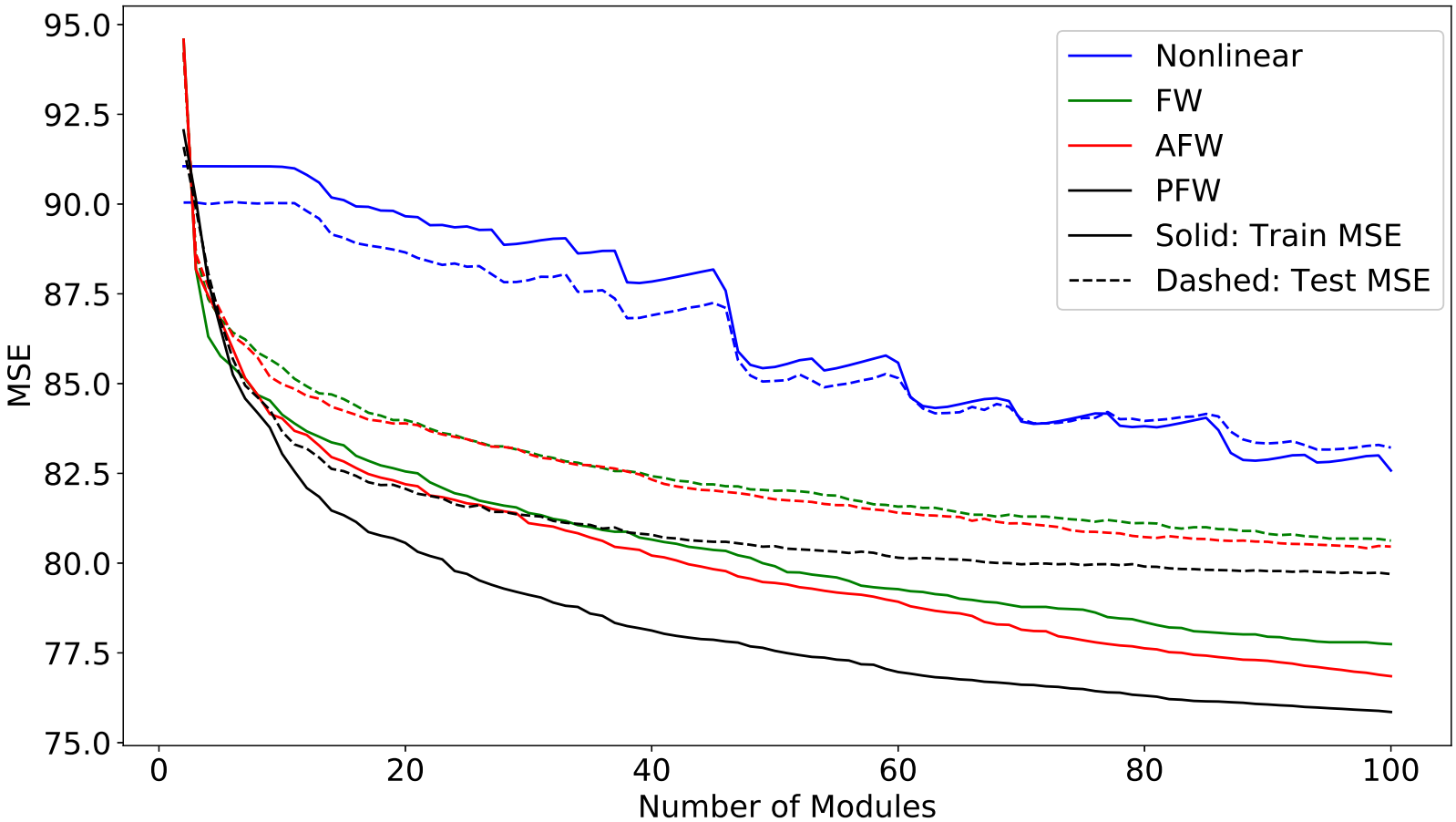
where $c = 2c_\phi B\left(\sqrt{2\ln(1/\delta)} + D\sqrt{p} + 2\right)$ and f^* is Bayes optimal.

$\Longrightarrow$ *Minimizing empirical risk $R_n(f)$ over the convex hull $\text{co}(\mathcal{G})$ results in minimizing the bound of the expected risk $R(f)$.*

Theorem 3: Let $\hat{f} = \arg\min_{f \in \text{co}(\mathcal{G})} R_n(f)$, and $h^* = \arg\min_{f \in \text{co}(\mathcal{G})} R(f)$, then with probability at least $1-\delta$, $\quad R(\hat{f}) \leq R(h^*) + \frac{c}{\sqrt{n}}$.

$\Longrightarrow$ *In $\text{co}(\mathcal{G})$, the empirical risk minimizer $\hat{f}$ converges to the expected risk minimizer h^* at the rate $O(1/\sqrt{n})$.*

Algorithms to Learn GCE



Performance Comparison

Datasets	#Samples	GCE	XGBoost	RForest	NN	ConvNet
diabetes	442	**42.706**	46.569	49.519	43.283	44.703
boston	506	**2.165**	2.271	2.705	2.217	2.232
ca_housing	20,640	0.435	**0.393**	0.416	0.440	0.437
msd	515,345	**6.084**	6.291	6.462	6.186	7.610
iris	150	**0.00**	6.67	6.67	3.33	10.00
wine	178	**0.00**	2.78	2.78	**0.0**	**0.0**
breast_cancer	569	**3.51**	4.39	8.77	3.51	4.39
digits	1,797	2.78	3.06	**2.50**	3.33	3.06
cifar10_f	60,000	**4.86**	5.40	5.16	5.00	4.92
mnist	70,000	1.22	1.66	2.32	1.24	**1.11**
covertype	581,012	26.70	26.39	27.73	26.89	**26.56**
kddcup99	4,898,431	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**

Greedy Convex Ensemble

Tan Nguyen (Queensland University of Technology)

Nan Ye (University of Queensland)

Peter Bartlett (UC Berkeley)

December 15, 2020