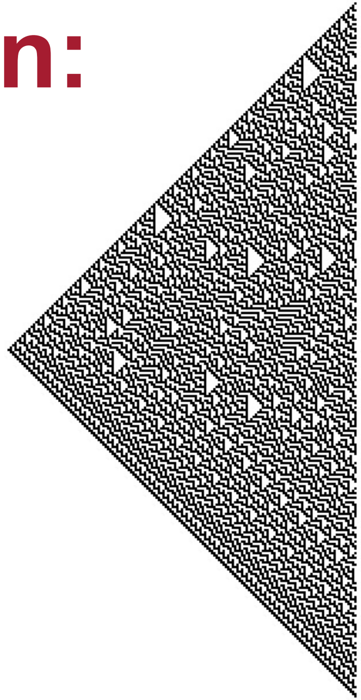
Distributional Generalization:
A New Kind of Generalization

Preetum Nakkiran*
Harvard → UCSD

Yamini Bansal*
Harvard



arXiv:2009.08092

Sept 24, 2021 @ DLCT

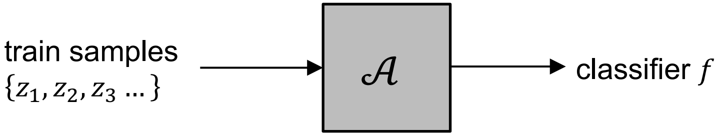
Motivation

Setting: Supervised classification $(x, y) \sim D$

Objects:

Classifiers (NN, decision trees,…): $f: \mathcal{X} \to \mathcal{Y}$

Learning Procedures (SGD,…): $(\mathcal{X} \times \mathcal{Y})^n \mapsto f$



Q: *Often study only Test Error. Can we hope to know more about f?*

(eg: many ways to get 40% error…)

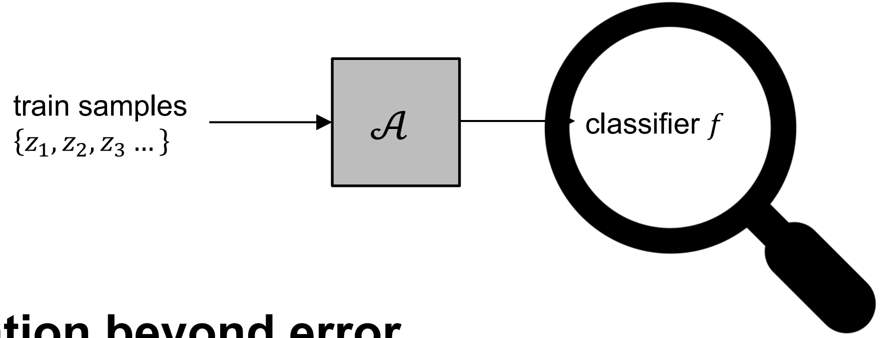
Motivation

Setting: Supervised classification $(x, y) \sim D$

Objects:

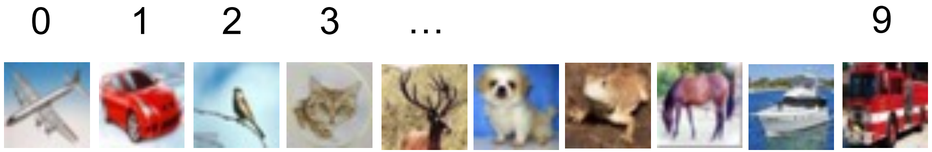
Classifiers (NN, decision trees,…): $f : \mathcal{X} \rightarrow \mathcal{Y}$

Learning Procedures (SGD,…): $(\mathcal{X} \times \mathcal{Y})^n \mapsto f$



This talk: Generalization beyond error…

Experiment

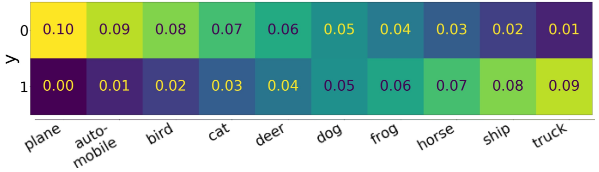


Distribution on (x, y):

$x \sim$ { random CIFAR-10 image }

$y|x$ ~ Bernoulli(type(x) / 10)

Train Set (x, y)



Sample n=50K from this distribution.

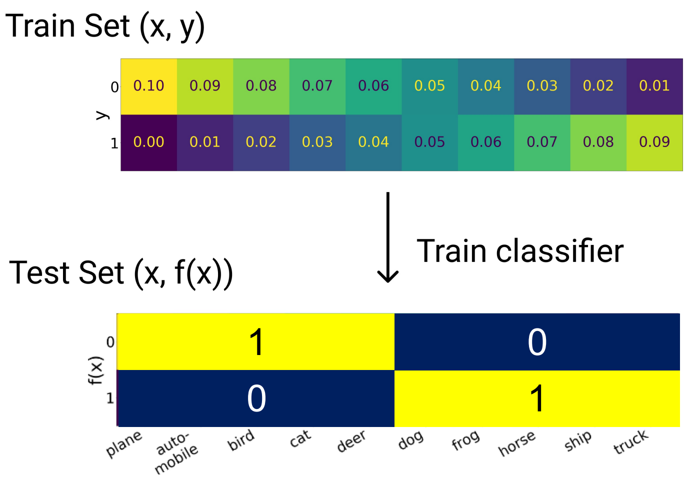
Train a ResNet to interpolation,

to predict $f: \mathcal{X} \to \mathcal{Y}$

Q: What happens at test time?

A: ~Same distribution!

We use a method for **classification.**
We **don't get** a good classifier: high test error!



We get an approximate **sampler**:

$$f(x) \sim p(y\,|x)$$

<u>Happens for:</u>

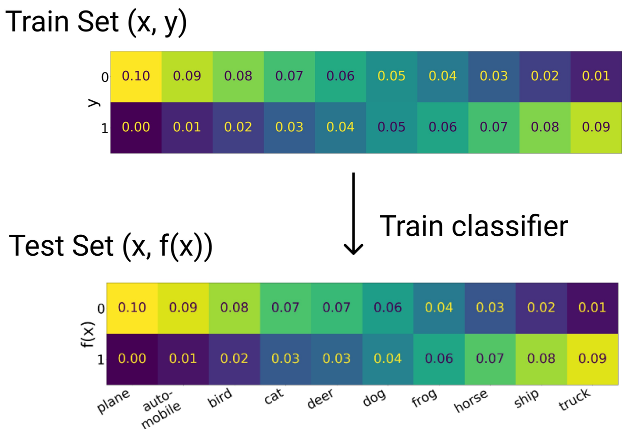
- Interpolating **neural networks**
- Interpolating **kernel regressors**
- Interpolating **decision trees**

Best thought of as samplers.

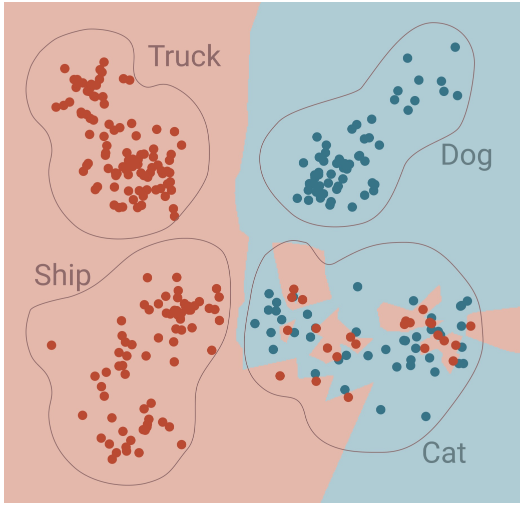
Large generalization gap. But "distributional generalization"
Classical generalization insufficient language.

Locality?

Classifier sensitive to *subclass-structures*



1-Nearest-Neighbors (in a well-clustered space) would have the same behavior.



… but why do ResNets?

Distributional Generalization (informal):

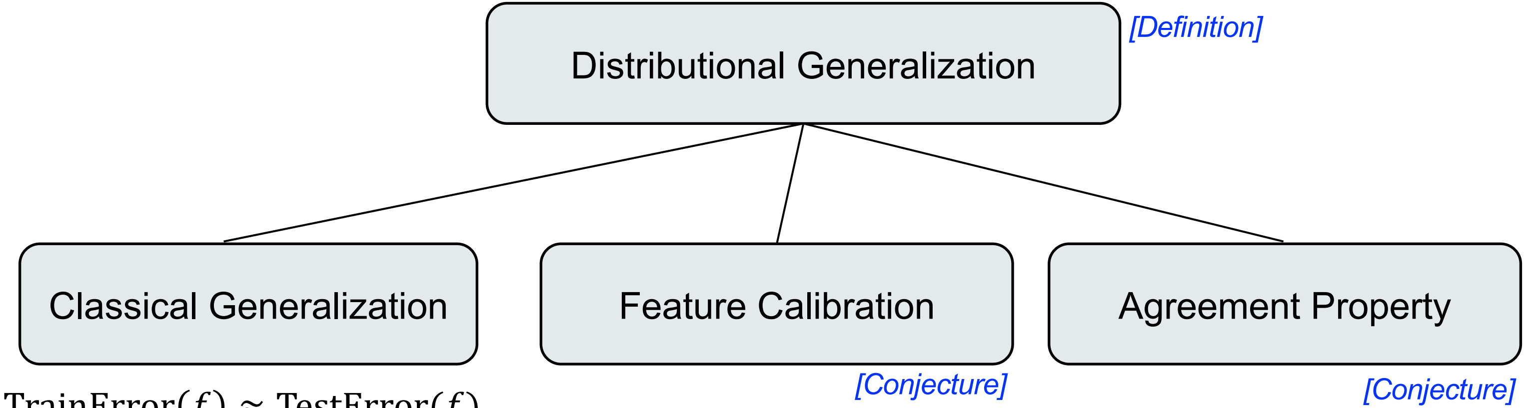
" Test and train outputs of classifiers are close as ***distributions*** *"*

$$(x, f(x))_{x \in \mathrm{TrainSet}} \approx (x, f(x))_{x \in \mathrm{TestSet}}$$

Distributional Generalization (informal):

*“ Test and train outputs of classifiers are close as **distributions** ”*

$$(x, f(x))_{x \in \mathrm{TrainSet}} \approx (x, f(x))_{x \in \mathrm{TestSet}}$$



$$\mathrm{TrainError}(f) \approx \mathrm{TestError}(f)$$

Classical Framework of Generalization

Classical Generalization: $\mathrm{Error}_{\mathrm{TrainSet}}(f) \approx \mathrm{Error}_{\mathrm{Test}}(f)$

$$\Updownarrow$$

$$\mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\ \mathbb{1}\{\ \hat{y} \neq y(x)\}\ ] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[\ \mathbb{1}\{\ \hat{y} \neq y(x)\}\ ]$$

Expectation of the same function under different distributions.

$$\mathrm{T}_{err}(x,\hat{y}) := \mathbb{1}\{\ \underset{\text{predicted label}}{\underset{\uparrow}{\hat{y}}} \neq \underset{\text{true label}}{\underset{\uparrow}{y(x)}}\}$$

Joint Distributions: D_{Train}, D_{Test} over $\mathcal{X} \times \mathcal{Y}$

Classical Framework of Generalization

Classical Generalization: $\mathrm{Error}_{\mathrm{TrainSet}}(f) \approx \mathrm{Error}_{\mathrm{TestSet}}(f)$

$$\Updownarrow$$

$$\mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\,\mathbb{I}\{\,\hat{y} \neq y(x)\}\,] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[\,\mathbb{I}\{\,\hat{y} \neq y(x)\}\,]$$

classifier output ($\hat{y}$) true label ($y(x)$)

To sample from D_{Train}:

- Sample $\mathrm{TrainSet} \sim D^n$
- Train classifier $f \leftarrow \mathrm{Learn(TrainSet)}$
- Sample train pt $x \sim \mathrm{TrainSet}$
- Output $(x, f(x))$

To sample from D_{Test}:

- Sample $\mathrm{TrainSet} \sim D^n$
- Train classifier $f \leftarrow \mathrm{Learn(TrainSet)}$
- Sample test pt $x \sim D$
- Output $(x, f(x))$

Distributional Generalization

Defn. *A trained classifier f satisfies* **classical generalization** *if:*

$$\mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\,T_{\mathrm{err}}(x,\hat{y})\,] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[T_{\mathrm{err}}(x,\hat{y})]$$

Defn. *A trained classifier f satisfies* **distributional generalization** *for a family of tests $\mathcal{T}$*

$$\mathcal{T} \subseteq \{T: \mathcal{X} \times \mathcal{Y} \to [0,1]\}$$

if

$$\forall T \in \mathcal{T}: \quad \mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\,T(x,\hat{y})\,] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[T(x,\hat{y})]$$

Ex:

1. $\mathcal{T} = \{T_{\mathrm{err}}\} \Leftrightarrow$ classical generalization
2. $\mathcal{T} = \{all\ bounded\ tests\} \Leftrightarrow$ *TV-closeness*

Distributional Generalization

Defn. *A trained classifier f satisfies* **classical generalization** *if:*

$$\mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\,T_{\text{err}}(x,\hat{y})\,] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[T_{\text{err}}(x,\hat{y})]$$

Defn. *A trained classifier f satisfies* **distributional generalization** *for a family of tests $\mathcal{T}$*

$$\mathcal{T} \subseteq \{T : \mathcal{X} \times \mathcal{Y} \rightarrow [0,1]\}$$

if

$$\forall T \in \mathcal{T} : \; \mathbb{E}_{(x,\hat{y})\sim D_{Train}}[\,T(x,\hat{y})\,] \approx \mathbb{E}_{(x,\hat{y})\sim D_{Test}}[T(x,\hat{y})]$$

Intuition (DG):

"Train and Test outputs are close ***as distributions****"*

$$(x, f(x))_{x\in \text{TrainSet}} \approx (x, f(x))_{x\in \text{TestSet}}$$

Interpolating Classifiers

Special case of Distributional Generalization:

$$(x, f(x))_{x \in \mathrm{TestSet}} \approx (x, f(x))_{x \in \mathrm{TrainSet}} \equiv (x, y)_{x,y \sim \mathcal{D}}$$

Distributional Generalization

for interpolating classifiers

PART I: Feature Calibration

Roadmap

We want to formalize the closeness:

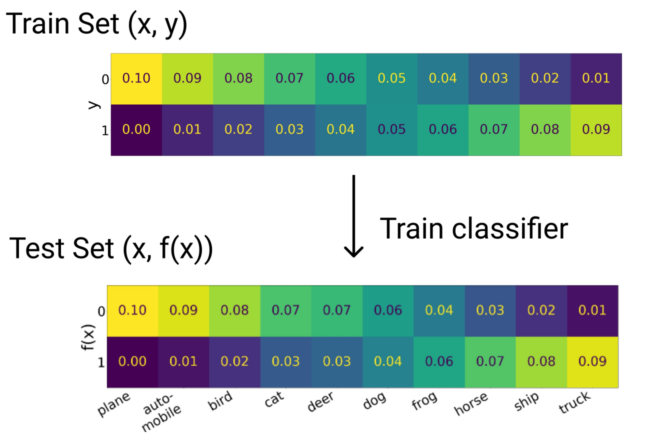
$$(x, f(x)) \approx (x, y)$$

Claim: For some "good" partitions $L: \mathcal{X} \to [M]$,

$$(L(x), f(x)) \approx_{TV} (L(x), y)$$

Which partitions are "good"?

- Depends on architecture, distribution, num samples…
- Intuitively, "partitions which can be learnt"



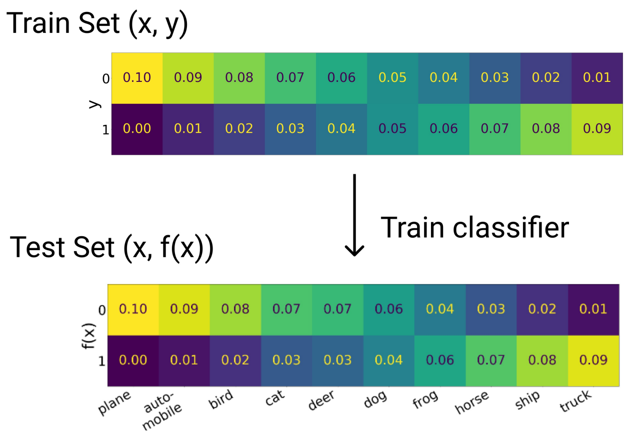
x is "coarsened" into a partition L(x)

Feature Calibration

<u>Conjecture (informal)</u>:

Marginal distributions of $\boldsymbol{f(x)}$ *and* $\boldsymbol{y}$ *match,*
when conditioned on any "good" subgroup $\boldsymbol{L(x) \in \{0, 1\}}$

Eg: $p(f(x) \mid x \in CAT) \approx p(y \mid x \in CAT)$



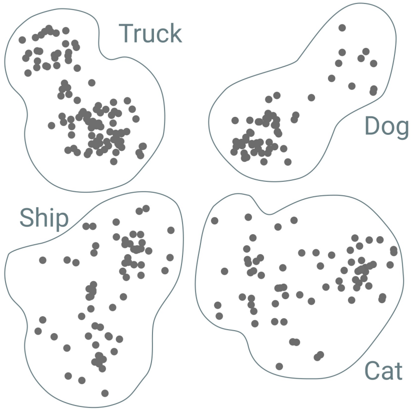
What is a "good" subgroup?
- Subgroups which are *themselves learnable*

- Many "good subgroups"! (cats, animals, objects,…)

- Training procedure doesn't know about subgroups...

Definition: Distinguishable Feature

Given: Training procedure $\mathcal{F}$, distribution $(x, y) \sim \mathcal{D}$, num train samples n.



<u>Defn:</u> *An* $(\epsilon, \mathcal{F}, \mathcal{D}, n)$-**distinguishable feature** *is a labeling* $L: \mathcal{X} \rightarrow [M]$ *of the domain* $\mathcal{X}$ *that is learnable to test accuracy* $\geq 1 - \epsilon$.

1. Sample unlabeled $\{x_i\} \sim D^n$.
2. Label as $y_i := L(x_i)$
3. Train classifier $f \leftarrow \text{Train}_{\mathcal{F}}(\{x_i, y_i\})$
4. Check Test Accuracy: $\Pr_{x \sim D}[f(x) = L(x)] \geq 1 - \epsilon$

eg: L: X → {cat, dog, plane…} is a ResNet-distinguishable feature for CIFAR with n=50K samples.

L: X → {animal, object} is an MLP-dist feature

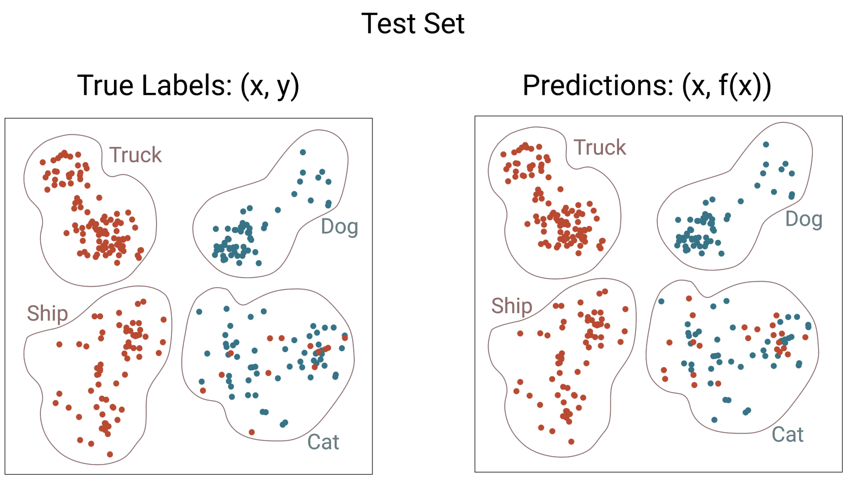
Main Conjecture: Feature Calibration

Conjecture: *For all natural distributions $\mathcal{D}$, family of interpolating models $\mathcal{F}$, and train samples $n \in \mathbb{N}$ the following holds.*

Let $f \leftarrow \mathrm{Train}_{\mathcal{F}}(\mathcal{D}^n)$ *be a trained classifier. Then* $\forall(\epsilon, \mathcal{F}, \mathcal{D}, n)$*-distinguishable features* L:

$$\boxed{\left(L(x), f(x)\right)_{x\sim\mathcal{D}} \approx_{\epsilon} \left(L(x), y\right)_{x,y\sim\mathcal{D}}}$$

compare to: $\left(x, f(x)\right) \approx (x, y)$

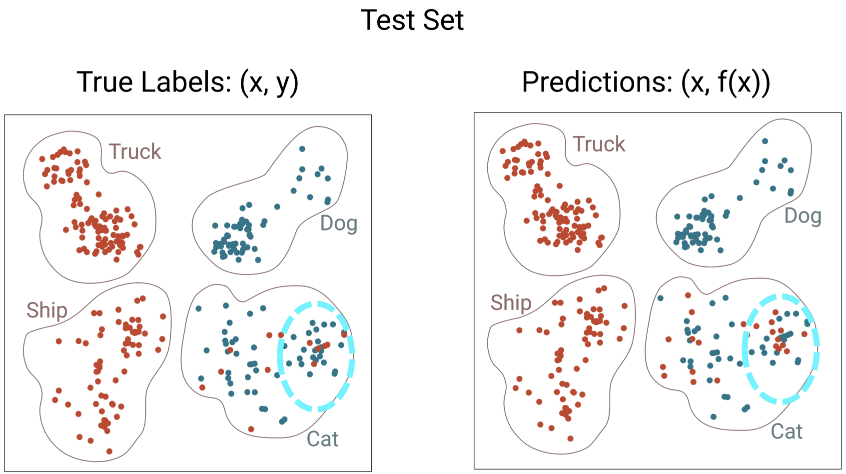


Main Conjecture: Feature Calibration

<u>**Conjecture:**</u> *For all natural distributions $\mathcal{D}$, family of interpolating models $\mathcal{F}$, and train samples $n \in \mathbb{N}$ the following holds.*

*“Marginal distributions of **f(x)** and **y** match, when conditioned on any distinguishable-feature **L**”*

Eg: $p(f(x) | \, x \in CAT) \approx p(y \,|\, x \in CAT)$



Main Conjecture: Feature Calibration

<u>Conjecture:</u> *For all natural distributions $\mathcal{D}$, family of interpolating models $\mathcal{F}$, and train samples $n \in \mathbb{N}$ the following holds.*

*"Marginal distributions of **f(x)** and **y** match,*
*when conditioned on any distinguishable-feature **L**"*

Eg: $p(f(x)|\ x \in CAT) \approx p(y \mid x \in CAT)$

Remarks:

- Train one interpolating classifier f. Holds "automatically" for **all** distinguishable features L.
- Formally true for $\mathcal{F}$ = 1-Nearest-Neighbors.
- Statement of density approximation:

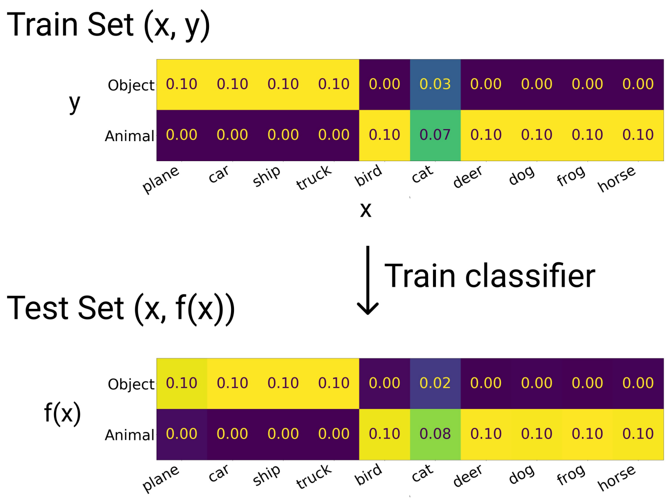
$\boldsymbol{f(x)}$ ***"looks like" sample from*** $\boldsymbol{p(y|x)}$

Experiments to test Conjecture 1: $(L(x), f(x)) \approx_{\epsilon} (L(x), y)$

All experiments: Pick a distribution $\mathcal{D}$, model $\mathcal{F}$, and distinguishable feature L.

Compute joint distribution $(L(x), f(x))$ vs. $(L(x), y)$ on test set.

Joint distribution (L(x), y) and (L(x), f(x))



L(x) := CIFAR-10 class of x

ResNets on CIFAR-10: Arbitrary Confusion Matrix

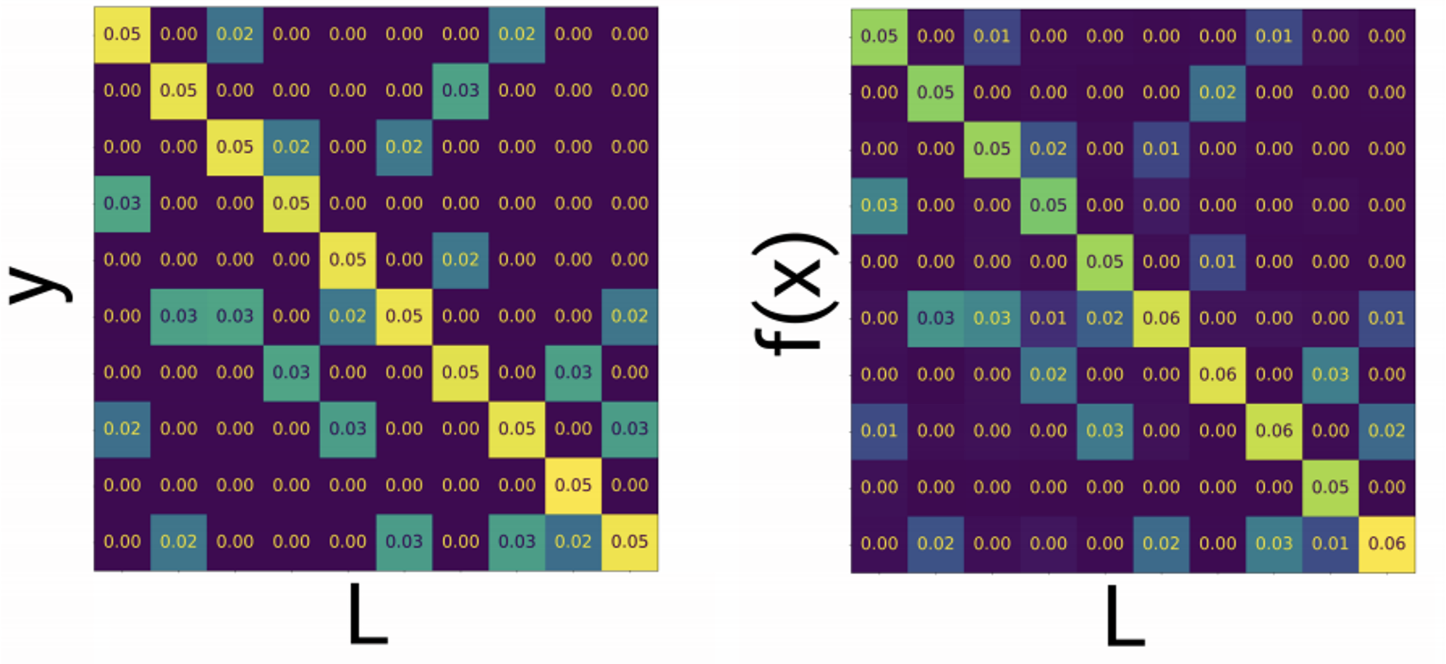
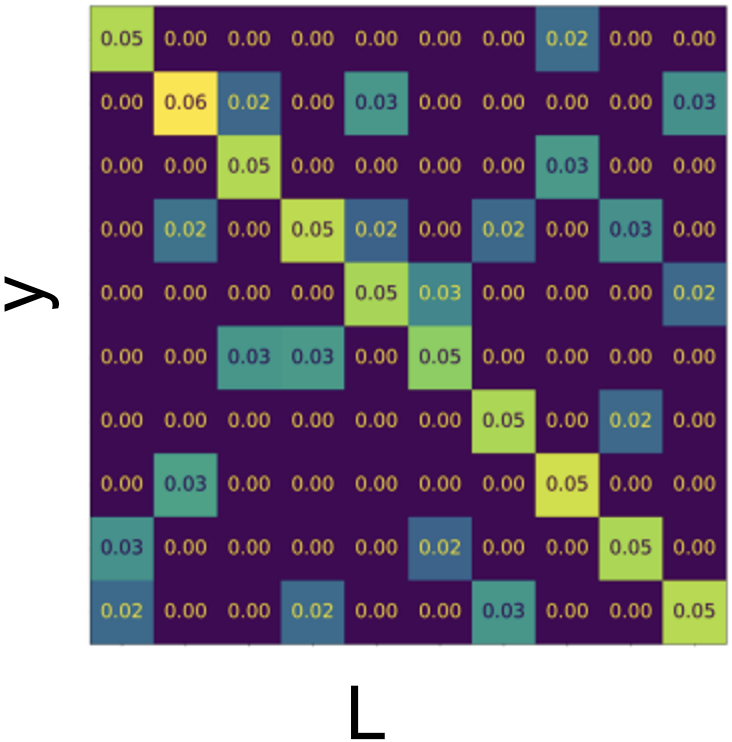
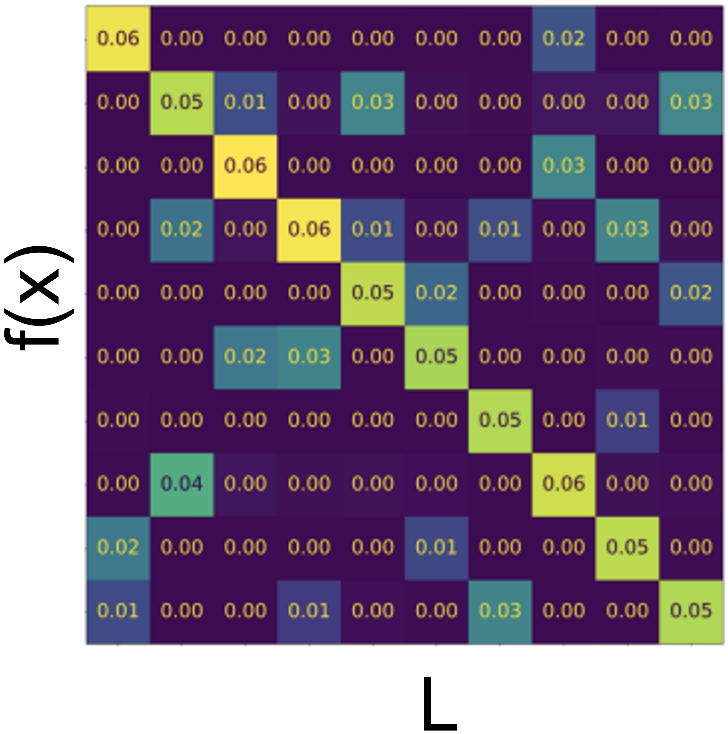


Figure 10: Train/test confusion matrices (left/right) for WideResNet28-10 on CIFAR-10 mislabeled according to a random confusion matrix

RBF kernel on MNIST (λ=0)





Decision Trees on UCI

"Decision trees ≈ adaptive nearest-neighbors"

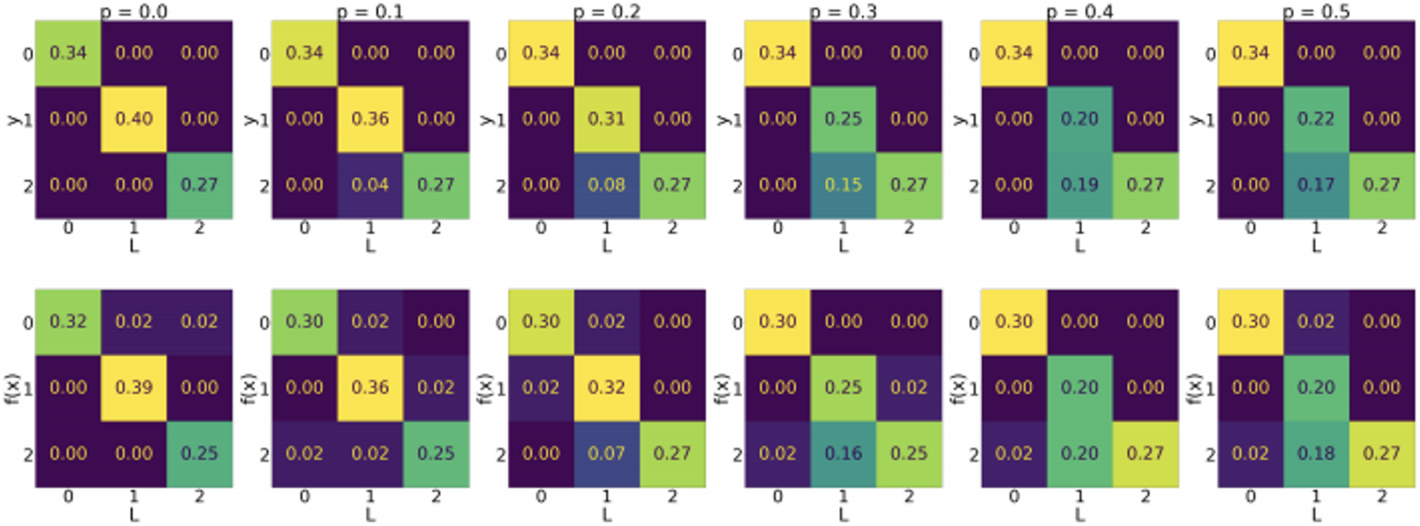


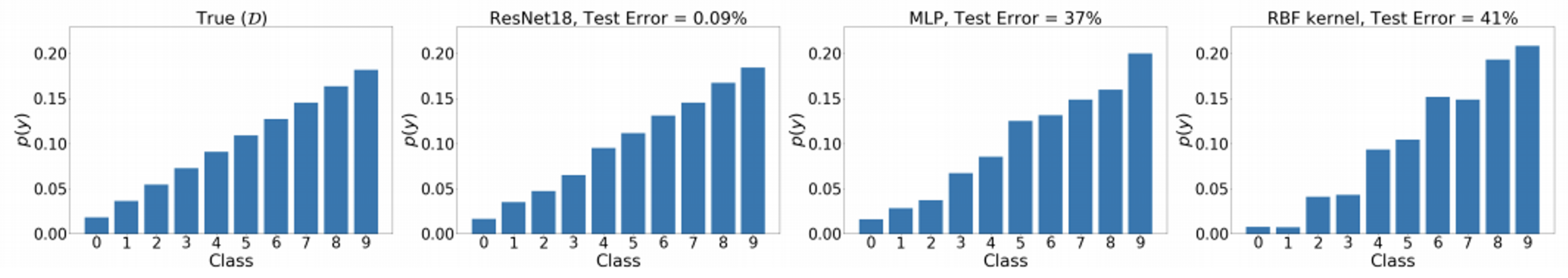
Figure 12: Decision trees on UCI (wine). We add label noise that takes class 1 to class 2 with probability $p \in [0, 0.5]$. Each column shows the test and train confusion matrices for a given p. Note that this decision trees achieve high accuracy on this task with no label noise (leftmost column).

For deterministic distributions

Consider "constant feature" $L(x) = 0$ in the conjecture $(L(x), f(x)) \approx_\epsilon (L(x), y)$

Conjecture ⇒ Interpolating classifiers have the right marginal distribution of labels:

$$\boldsymbol{p(f(x)) \approx p(y)}$$



Beyond Error

ImageNet: Image classification. 1000-classes, 116 dogs.

AlexNet (f) gets 56% test accuracy.

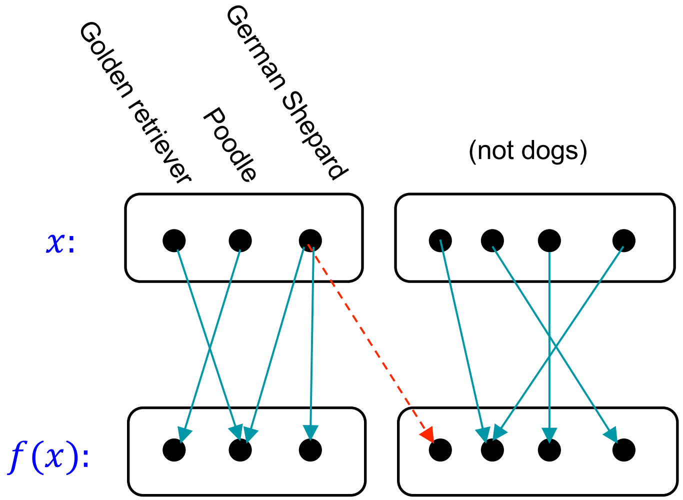
Does it at least classify dogs as *some type* of dog?

- Yes! (98% acc). High accuracy when ”zoomed out”

- Predicted by our conjecture:

“IF AlexNet could learn to classify dogs vs. not-dogs
(when trained on this binary task),
THEN AlexNet will classify most dogs as dogs
(when trained on 1000-class ImageNet)”

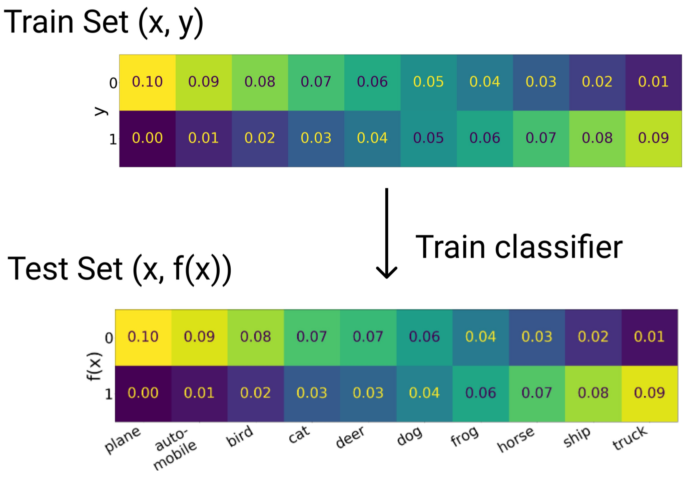
- Even “bad” classifiers (w.r.t. test error), can have “good” hidden structure



Theory Implications

1. “Overfitting is not always benign”

Fitting noise in train set → Noise at test time



2. “Interpolating neural networks are NOT consistent”

Early-stopped neural networks are consistent

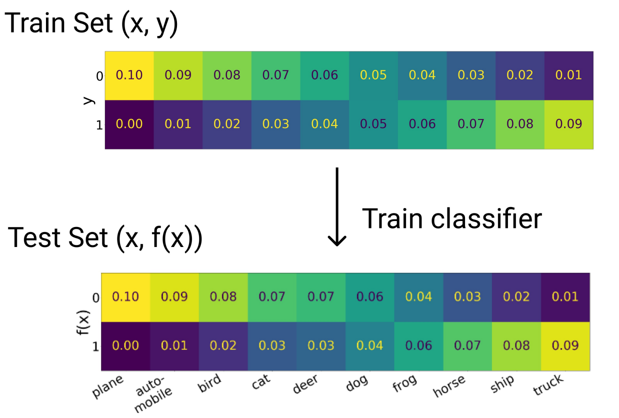
Ziwei Ji Justin D. Li Matus Telgarsky
<{ziweiji2,jdli3,mjt}@illinois.edu>
University of Illinois, Urbana-Champaign

Theory Implications

<u>Overparameterized Limit</u>: (data << model)

$$\lim_{N\to\infty}\lim_{S\to\infty} f_{N,S}(x) \sim p(y|x)$$

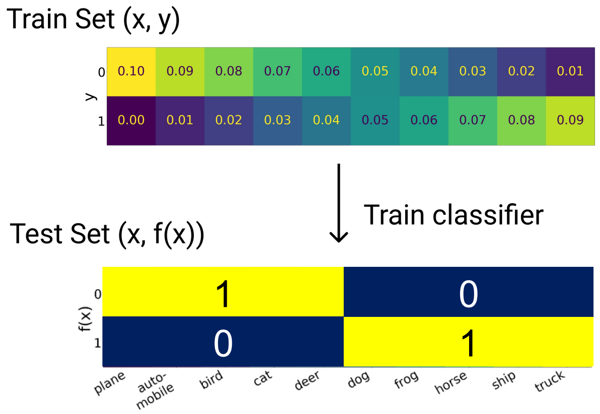
Fit the (noisy) train set.



<u>Underparameterized Limit:</u> (data >> model)

$$\lim_{S\to\infty}\lim_{N\to\infty} f_{N,S}(x) = \mathrm{argmax}_y p(y \mid x)$$

Train Set = Test Set.



Implications for Ensembling

Observation: Ensembling helps *a lot* for distributions with label noise.

"Explanation":
Single classifier approximates samples from conditional density:

$$f(x) \sim p(y|x)$$

⇒ Ensemble of classifiers approximate argmax:

$$\mathrm{plur}(f_1, f_2, \ldots, f_k)(x) \approx \operatorname*{argmax}_y p(y|x)$$

Non-Interpolating Models

Non-interpolating classifiers

Distributional generalization intuition:

$$(x, f(x))_{x \in \mathrm{TrainSet}} \approx (x, f(x))_{x \in \mathrm{TestSet}}$$

“behavior on train set ≈ behavior on test set”

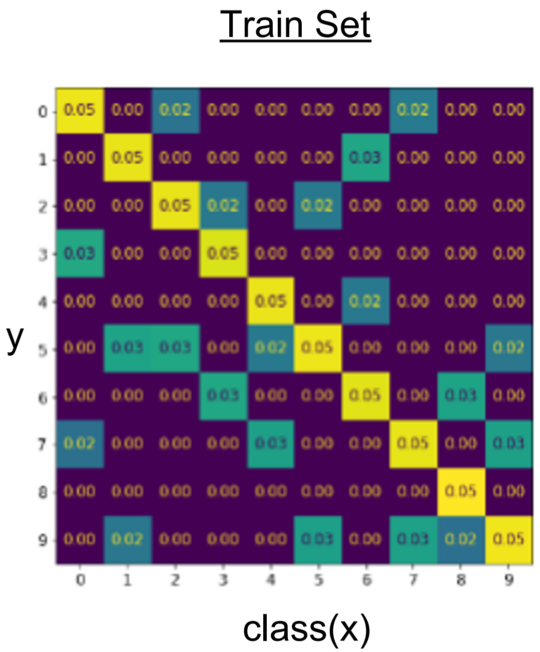
Non-interpolating classifiers

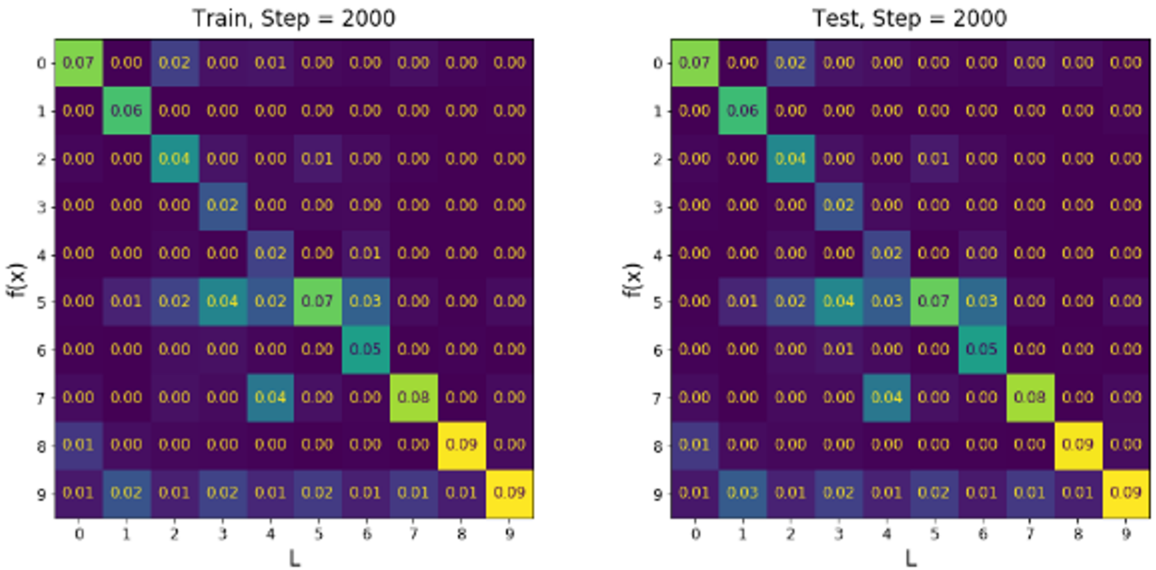
Distributional generalization intuition:

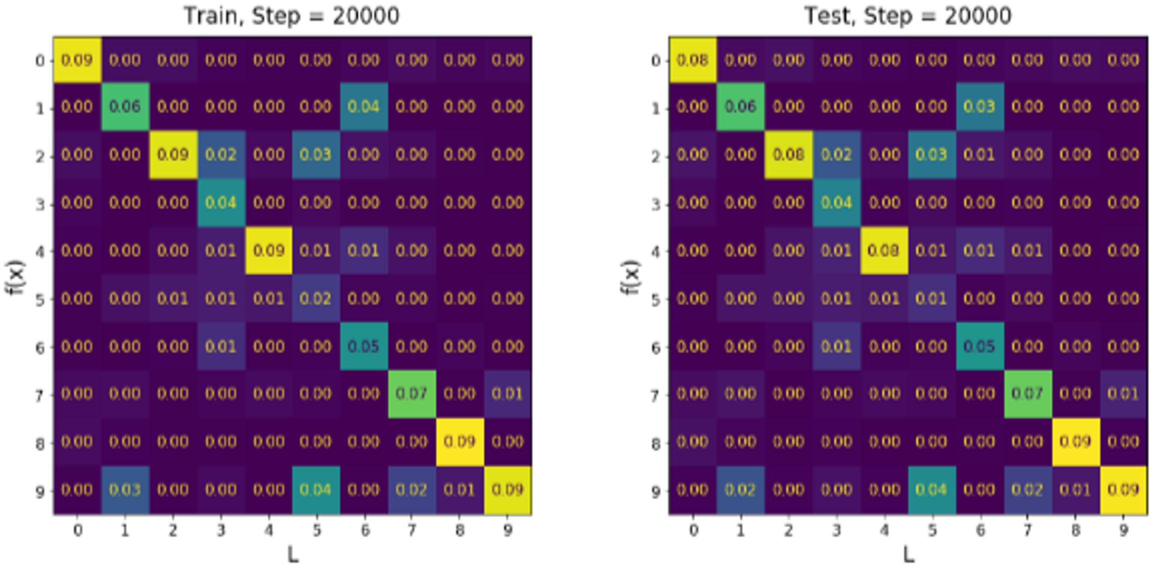
$$(x, f(x))_{x \in \text{TrainSet}} \approx (x, f(x))_{x \in \text{TestSet}}$$

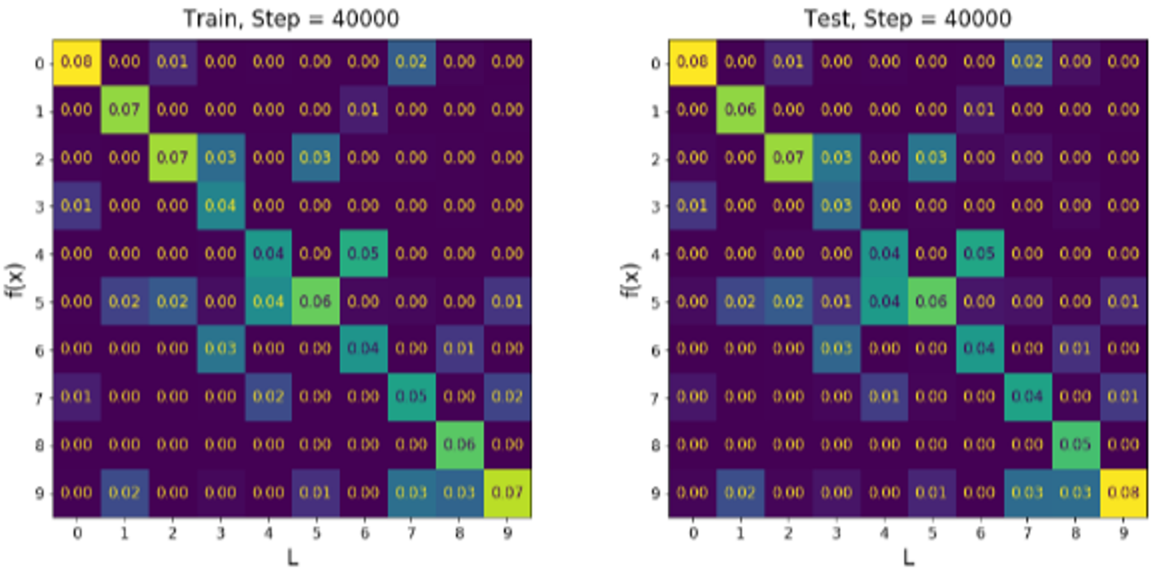
"behavior on train set ≈ behavior on test set"

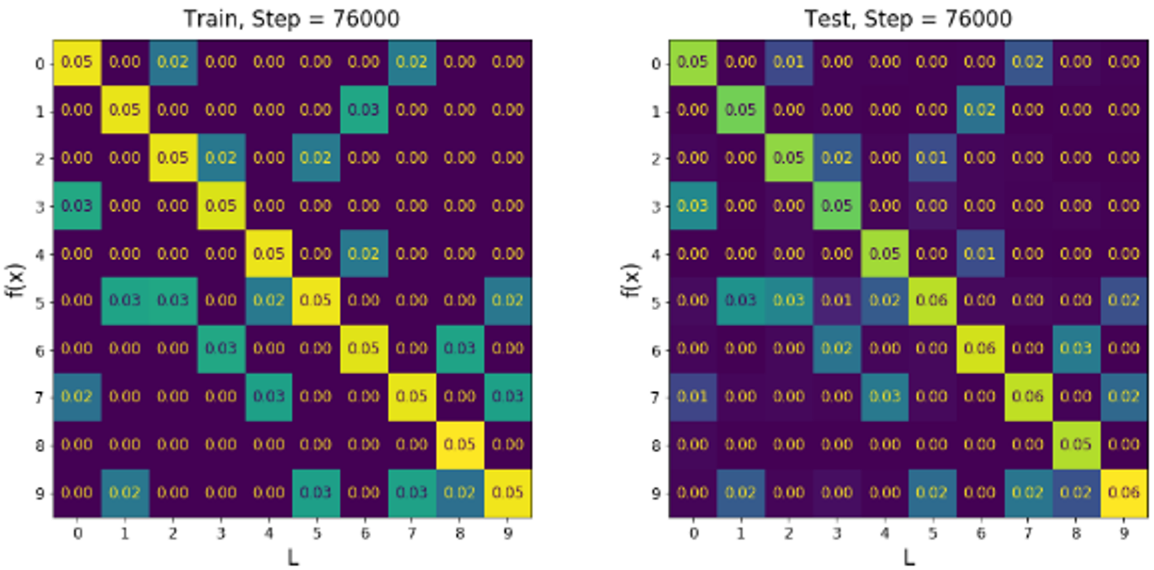
WideResNet-28-10 on CIFAR-10 w/ label noise




Train, Step = 2000
Test, Step = 2000
f(x)
L


Train, Step = 20000
0.09 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.06 0.00 0.00 0.00 0.00 0.04 0.00 0.00 0.00
0.00 0.00 0.09 0.02 0.00 0.03 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.01 0.09 0.01 0.01 0.00 0.00 0.00
0.00 0.00 0.01 0.01 0.01 0.02 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.01 0.00 0.00 0.05 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.07 0.00 0.01
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.09 0.00
0.00 0.03 0.00 0.00 0.00 0.04 0.00 0.02 0.01 0.09
f(x)
L
Test, Step = 20000
0.08 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.06 0.00 0.00 0.00 0.00 0.03 0.00 0.00 0.00
0.00 0.00 0.08 0.02 0.00 0.03 0.01 0.00 0.00 0.00
0.00 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.01 0.08 0.01 0.01 0.01 0.00 0.00
0.00 0.00 0.00 0.01 0.01 0.01 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.01 0.00 0.00 0.05 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.07 0.00 0.01
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.09 0.00
0.00 0.02 0.00 0.00 0.00 0.04 0.00 0.02 0.01 0.09
f(x)
L


Train, Step = 40000
Test, Step = 40000
f(x)
L
0.08 0.00 0.01 0.00 0.00 0.00 0.00 0.02 0.00 0.00
0.00 0.07 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00
0.00 0.00 0.07 0.03 0.00 0.03 0.00 0.00 0.00 0.00
0.01 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.04 0.00 0.05 0.00 0.00 0.00
0.00 0.02 0.02 0.00 0.04 0.06 0.00 0.00 0.00 0.01
0.00 0.00 0.00 0.03 0.00 0.00 0.04 0.00 0.01 0.00
0.01 0.00 0.00 0.00 0.02 0.00 0.00 0.05 0.00 0.02
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.06 0.00
0.00 0.02 0.00 0.00 0.00 0.01 0.00 0.03 0.03 0.07
0.08 0.00 0.01 0.00 0.00 0.00 0.00 0.02 0.00 0.00
0.00 0.06 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00
0.00 0.00 0.07 0.03 0.00 0.03 0.00 0.00 0.00 0.00
0.01 0.00 0.00 0.03 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.04 0.00 0.05 0.00 0.00 0.00
0.00 0.02 0.02 0.01 0.04 0.06 0.00 0.00 0.00 0.01
0.00 0.00 0.00 0.03 0.00 0.00 0.04 0.00 0.01 0.00
0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.04 0.00 0.01
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.05 0.00
0.00 0.02 0.00 0.00 0.00 0.01 0.00 0.03 0.03 0.08

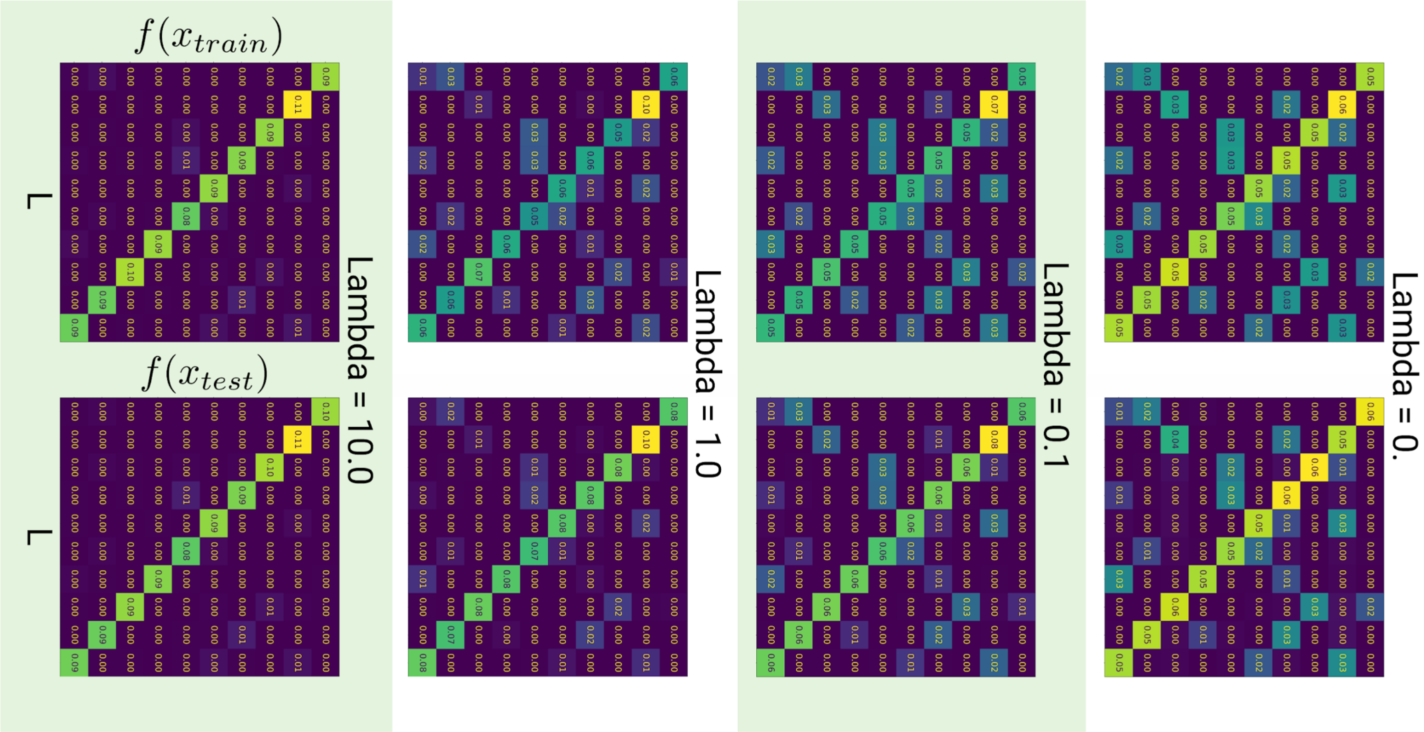


Classical generalization fails. DG holds.

MNIST+RBF, varying regularization

Train Set:

Test Set:



Summary of Connections/Significance

Implicit Bias: Many models with same train *and test* errors. Which one do we get?

This work: A “universal implicit bias” of interpolating models.

Structural constraints on the learnt classifier.

Benign Overfitting: Are interpolating models = “smooth part + benign interpolating part”?

This work: No, interpolation can hurt. Noise in train → noise in test.

Classical Generalization:

This work: DG can hold when classical generalization fails.

Even poorly-generalizing functions have predictable structure.

Fairness:

Fairness Implications

Setup: Distribution D on (X, Y).
Suppose there is a “protected attribute” R(x) e.g. Race

Assume that R(x) is independent of Y(x) on the distribution.

Train classifier F to predict Y from X.

Q: Will F(X) be independent of R(X)?

Confusion Matrix Implications

For standard **interpolating methods** on **balanced binary classification** tasks:

(False Positive Rate) $\approx$ (False Negative Rate)

is it true?

PART II: Agreement Property

Agreement Property

Experiment:

- Take two classifiers, trained on disjoint train sets, each with accuracy ~50% on a 10-class problem.
- What is the probability they agree with *each other* on test set?

$$\Pr_{\substack{f_1, f_2 \\ (x,y)\sim\mathcal{D}}} [f_1(x) = f_2(x)]$$

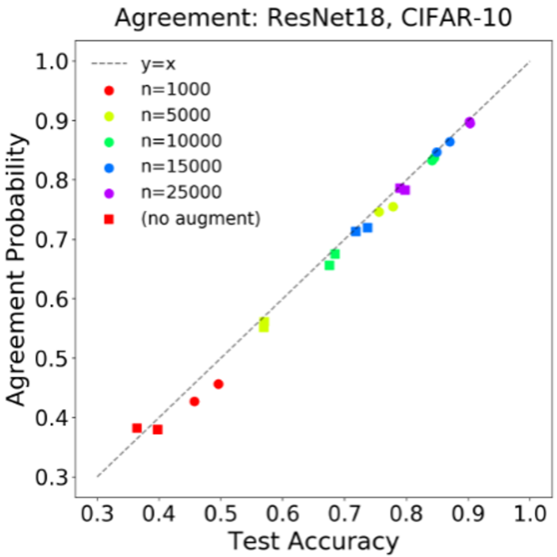
Agreement Property

Experiment:

- Take two classifiers, trained on disjoint train sets, each with accuracy ~50% on a 10-class problem.
- What is the probability they agree with *each other* on test set?

$$\Pr_{\substack{f_1 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = y] \approx \Pr_{\substack{f_1, f_2 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = f_2(x)]$$

Experiments



(a) ResNet18 on CIFAR-10.

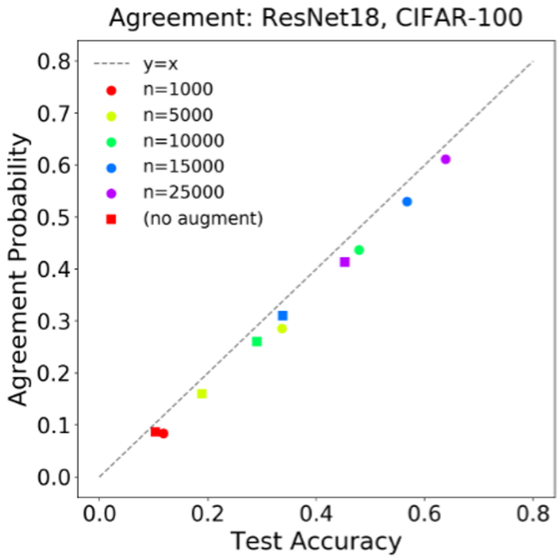
Agreement Property

<u>**Claim (informal):**</u> *For all natural classifiers and distributions, the **test accuracy** of a classifier f_1 is close to its **agreement probability** with an independent, identically-distributed classifier f_2*

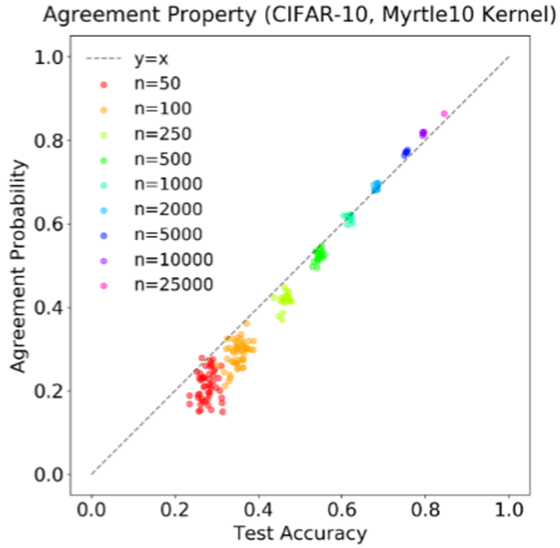
$$\Pr_{\substack{f_1 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = y] \approx \Pr_{\substack{f_1, f_2 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = f_2(x)]$$

* Special case of Distributional Generalization

Experiments

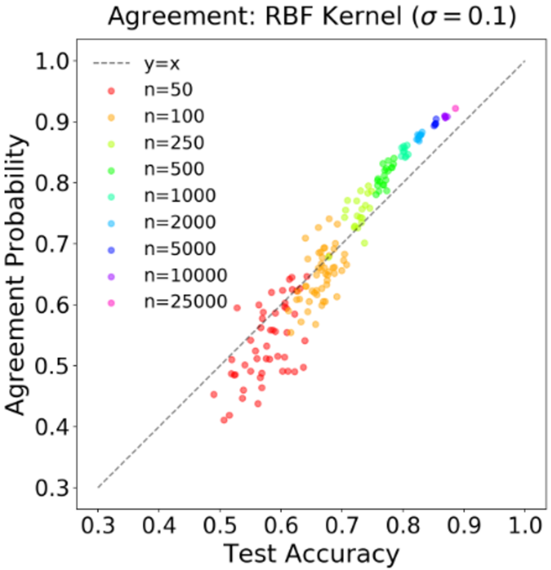


(b) ResNet18 on CIFAR-100.

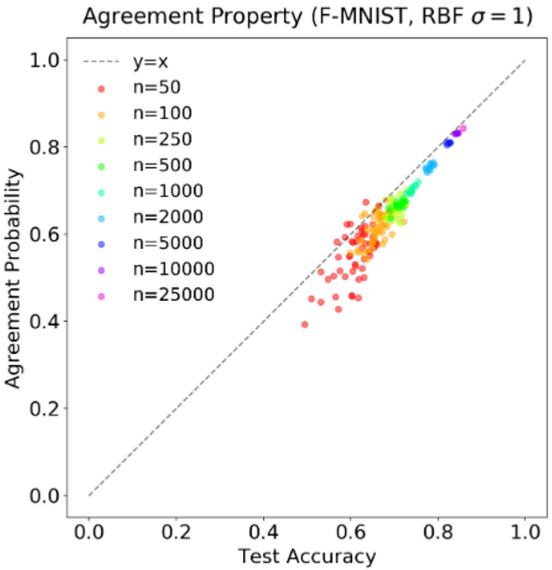


(c) Myrtle10 Kernel on CIFAR-10.

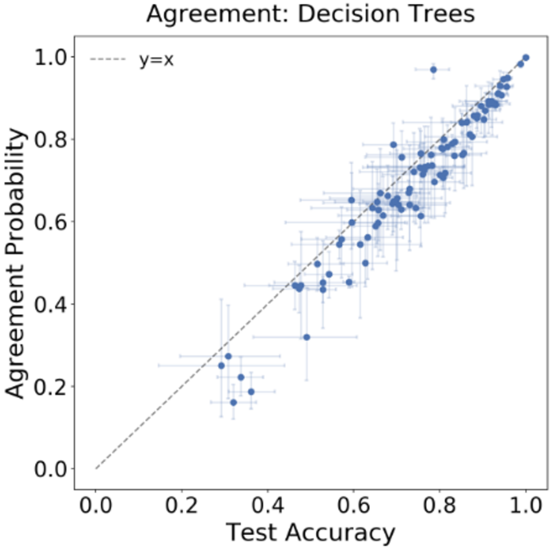
Experiments



(a) RBF on Fashion-MNIST



(b) RBF on Fashion-MNIST



(c) Decision Trees on UCI

Assessing Generalization of SGD via Disagreement

Yiding Jiang*
Carnegie Mellon University
ydjiang@cmu.edu

Vaishnavh Nagarajan*
Carnegie Mellon University
vaishnavh@cs.cmu.edu

Christina Baek
Carnegie Mellon University
kbaek@cs.cmu.edu

J. Zico Kolter
Carnegie Mellon University
Bosch Center for AI, Pittsburgh
zkolter@cs.cmu.edu

Structure of Confusion Matrices

<u>Claim (*speculation):</u> For interpolating, independent and identically-trained classifiers f_1, f_2 (iid), the joint densities:

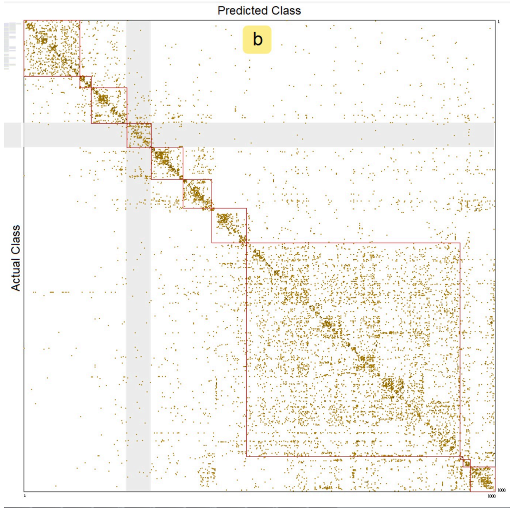
$$(f_1(x), y) \approx \big(f_1(x), f_2(x)\big)$$

as joint distributions over $\mathcal{Y} \times \mathcal{Y}$.

<u>Implies:</u>

- Agreement Property (trace): $\Pr_{\substack{f_1 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = y] \approx \Pr_{\substack{f_1, f_2 \\ (x,y)\sim\mathcal{D}}}[f_1(x) = f_2(x)]$

- Confusion matrix is **symmetric**
- Confusion matrix is **PSD**



https://arxiv.org/pdf/1710.06501.pdf

Limitations

- Conjectures not fully-specified: for which classifier families & distributions do they hold?
  - Tested various “natural” choices, but lacking formal conditions
  - Ensembles fail conjecture: Deep Ensembles, Random Forests, K-NN
- Sometimes slight deviations from predicted behavior
- No theoretical understanding beyond 1-NN

This work: First step in the study of Distributional Generalization

Conclusion

- Introduced "**distributional generalization**": fine-grained characterization of the outputs of classifiers (beyond just test error)

$$(x, f(x))_{x \in \mathrm{TrainSet}} \approx (x, f(x))_{x \in \mathrm{TestSet}}$$

- Several concrete instantiations: Feature Calibration, Agreement Property

- Holds in a variety of domains {neural-nets, kernels, decision-trees}
  More robust than classical generalization

- Hope: Deeper understanding of interpolating methods.
  Open questions abound...

Thanks!

preetum@cs.harvard.edu

EXTRAS

1-Nearest-Neighbor obeys Conjecture

Theorem 1. *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and let $n \in \mathbb{N}$ be the number of train samples. Assume the following regularity condition holds: Sampling the nearest-neighbor train point to a random test point yields (close to) a uniformly random test point. That is,*

$$\{\mathrm{NN}_S(x)\}_{\substack{S \sim \mathcal{D}^n \\ x \sim \mathcal{D}}} \quad \approx_\delta \quad \{x\}_{x \sim \mathcal{D}} \tag{4}$$

Then, Conjecture 1 holds. For all $(\varepsilon, \mathrm{NN}, \mathcal{D}, n)$-distinguishable partitions L, the following distributions are statistically close:

$$\{(y, L(x))\}_{x,y \sim \mathcal{D}} \quad \approx_{\varepsilon+\delta} \quad \{(\mathrm{NN}_S^{(y)}(x), L(x)\}_{\substack{S \sim \mathcal{D}^n \\ x,y \sim \mathcal{D}}} \tag{5}$$

Formal Definition

Source $\mathcal{D}$: (x, y) where $x, y \sim \mathcal{D}$	**Train $\mathcal{D}_{\text{tr}}$:** $(x, f(x))$ $S \sim \mathcal{D}^n$, $f \leftarrow \text{Train}_{\mathcal{F}}(S)$, $x, y \sim S$	**Test $\mathcal{D}_{\text{te}}$** $(x, f(x))$ $S \sim \mathcal{D}^n$, $f \leftarrow \text{Train}_{\mathcal{F}}(S)$, $x, y \sim \mathcal{D}$

Distributional Generalization:

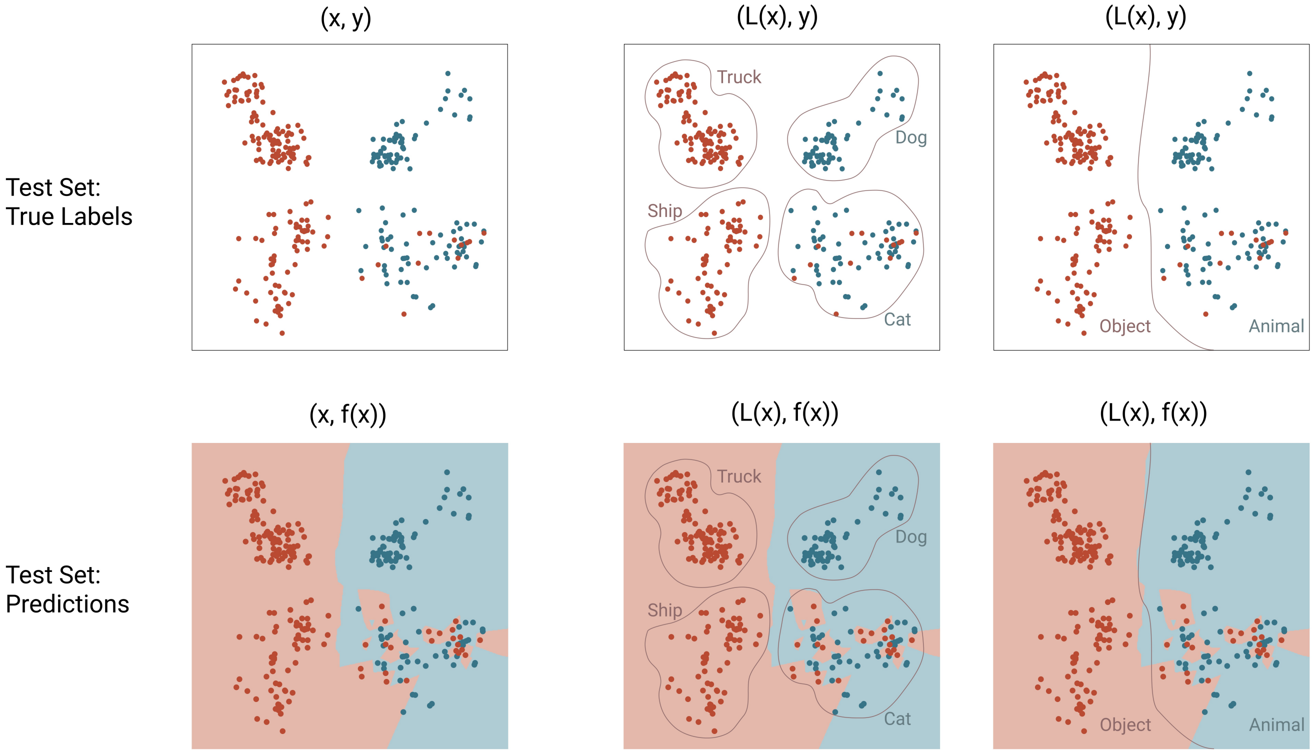
For interpolating classifiers: $\boxed{\mathcal{D} \equiv \mathcal{D}_{\text{tr}} \approx_{\varepsilon}^{\mathcal{C}} \mathcal{D}_{\text{te}}}$ $\quad \mathcal{C} \subseteq \{T : \mathcal{X} \times \mathcal{Y} \to [0,1]\}$

Defn: $$P \approx_{\varepsilon}^{\mathcal{C}} Q \iff \sup_{T \in \mathcal{C}} \left| \mathop{\mathbb{E}}_{(x,y)\sim P}[T(x,y)] - \mathop{\mathbb{E}}_{(x,y)\sim Q}[T(x,y)] \right| \leq \varepsilon$$

Equivalently: *"Marginals f(x), y close within each part of partition"*

$$(L(x), f(x)) \approx_\epsilon (L(x), y)$$

$$\Leftrightarrow \textit{For typical } \ell: \; p(f(x) \mid L(x) = \ell) \approx p(y \mid L(x) = \ell)$$



Connection/Significance

$$T_{\text{agree}} : (x, \widehat{y}) \mapsto \mathbb{1}\{f_1(x) = \widehat{y}\}$$

Special case of Indistinguishability:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[T(x,y)] \approx \mathop{\mathbb{E}}_{\substack{x\sim\mathrm{TestSet} \\ \hat{y}\leftarrow f_2(x)}}[T(x,\hat{y})]$$

Connection/Significance

$$T_{\text{agree}} : (x, \widehat{y}) \mapsto \mathbb{1}\{f_1(x) = \widehat{y}\}$$

Special case of Indistinguishability:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[T(x,y)] \approx \mathop{\mathbb{E}}_{\substack{x\sim\text{TestSet}\\ \hat{y}\leftarrow f_2(x)}}[T(x,\hat{y})]$$

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbb{1}\{f_1(x) = y\}] \approx \mathop{\mathbb{E}}_{\substack{x\sim\text{TestSet}\\ \hat{y}\leftarrow f_2(x)}}[\mathbb{1}\{f_1(x) = \hat{y}\}]$$

Connection/Significance

$$T_{\text{agree}} : (x, \widehat{y}) \mapsto \mathbb{1}\{f_1(x) = \widehat{y}\}$$

Special case of Indistinguishability:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[T(x,y)] \approx \mathop{\mathbb{E}}_{\substack{x\sim\text{TestSet}\\ \hat{y}\leftarrow f_2(x)}}[T(x,\hat{y})]$$

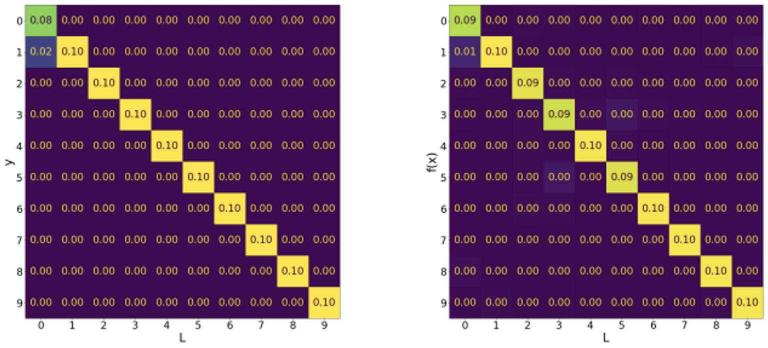
$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbb{1}\{f_1(x) = y\}] \approx \mathop{\mathbb{E}}_{\substack{x\sim\text{TestSet}\\ \hat{y}\leftarrow f_2(x)}}[\mathbb{1}\{f_1(x) = \hat{y}\}]$$

$$\Pr_{(x,y)\sim\mathcal{D}}[f_1(x) = y] \approx \Pr_{(x,y)\sim\mathcal{D}}[f_1(x) = f_2(x)]$$

ResNets on CIFAR-10

Label noise: 20% of plane → car on train set

⇒ roughly 20% plane → car on test set.



(a) Actual confusion matrix for $p = 0.2$ (b) Test Confusion matrix for $p = 0.2$

And for varying values of "20%":