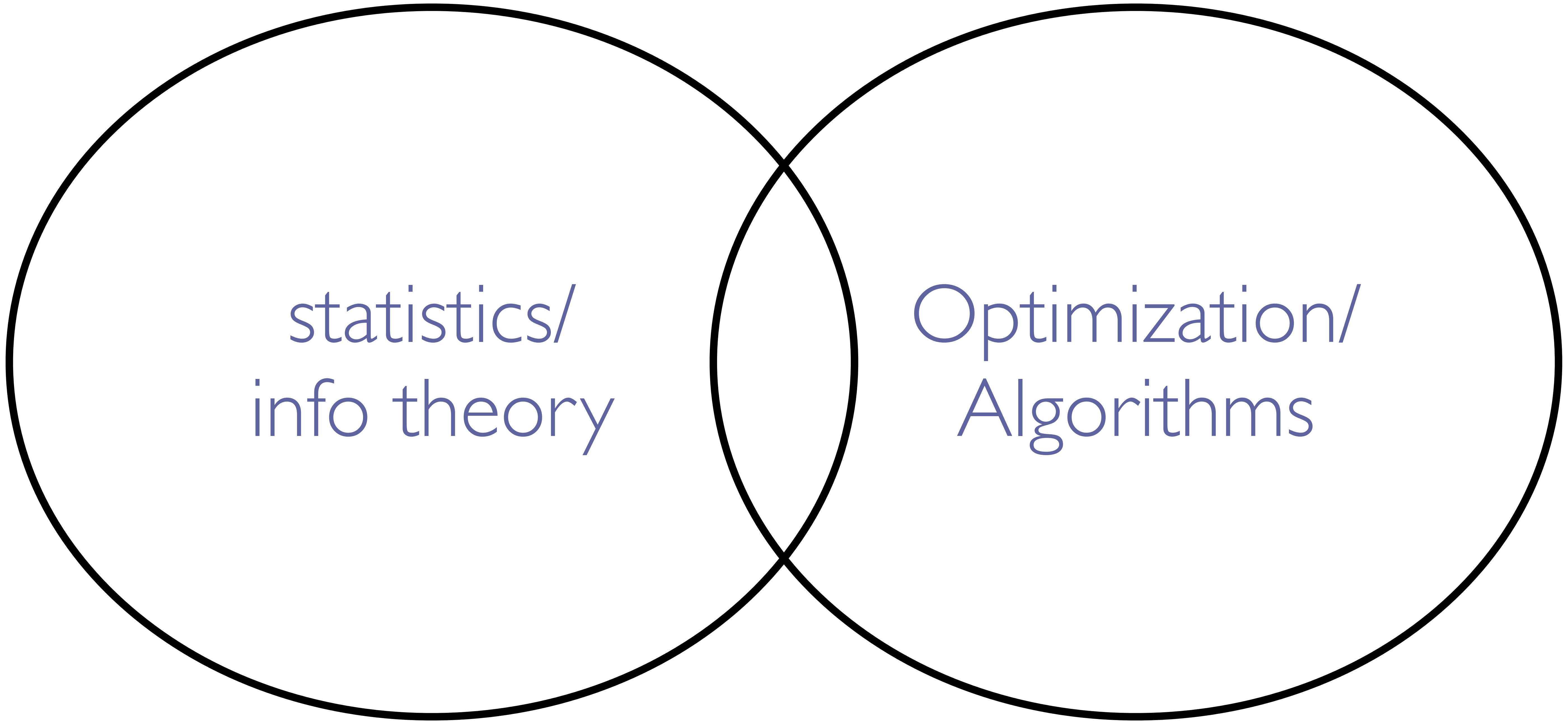
ECE826 Lecture 2:

Concentration of the Empirical Risk

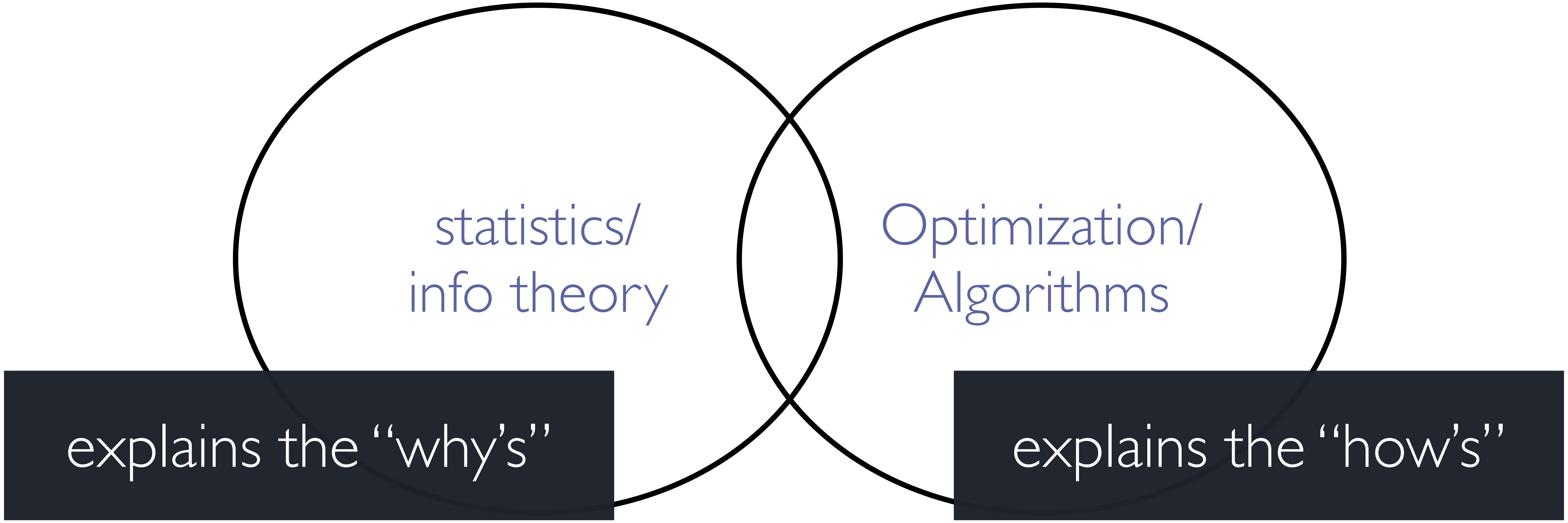
From This

To this

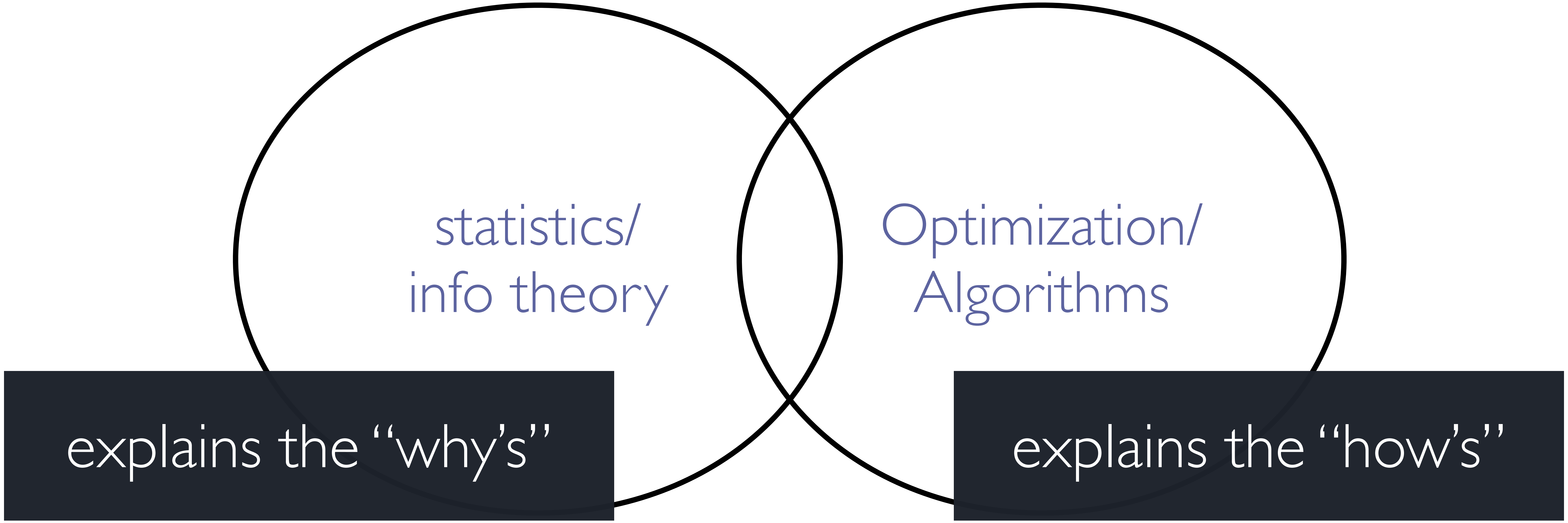
ML Research



ML Research


statistics/
info theory
Optimization/
Algorithms
explains the "why's"
explains the "how's"

ML Research



Today: Why/when does ERM work

Contents

- How to show concentration for ERM
- Parameter count bounds
- VC dim and Rademacher Complexity
- Do these bounds explain generalization in modern ML?

Reminder

- What we have: Labeled examples presented as (features, label)

$$(x_i, y_i) \sim \mathcal{D}$$

- A fixed hypothesis class (aka type of predictor) $\mathcal{H}$ (linear classifier, SVM, neural network, decision tree, etc)

Reminder

- <u>What we have:</u> Labeled examples presented as (features, label)

$$(x_i, y_i) \sim \mathscr{D}$$

- A fixed hypothesis class (aka type of predictor) $\mathscr{H}$ (linear classifier, SVM, neural network, decision tree, etc)

- <u>Goal:</u> We want to find the best $h \in \mathscr{H}$ for a given distribution $\mathscr{D}$ and loss function. How? ERM

Empirical Risk Minimization (ERM)

$$\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(h(x_i); y_i)}_{\text{performance of model } h \in \mathscr{H} \text{ on data point } x_i}$$

Empirical Risk Minimization (ERM)

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(h(x_i); y_i)}_{\text{performance of model } h \in \mathcal{H} \text{ on data point } x_i}$$

- Sidenote: Typically data set is split in three parts, [train|validation|test]
- 1) We use trainset to find models; 2) Performance evaluated on val set. 3) We pick one and report its performance on the test set.

- Please google: cross validation/hold out set/check literature on intro to stat. learning theory

Main Question for today

- When is the empirical risk a good estimator for the true risk

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h(x);y)\right]$$

- i.e., when does the loss of the ERMinimizer concentrate

Main Question for today

- When is the empirical risk a good estimator for the true risk

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h(x);y)\right]$$

- i.e., when does the loss of the ERMinimizer concentrate

- Today: How does the choice of the model affect the “worst case” concentration of the loss of the empirical risk?

Some Definitions

- There is an unknown distribution $\mathscr{D}$ over labeled examples from $\mathscr{X} \times \mathscr{Y}$ (i.e., feature x label space)

- We receive a "sample" data set of n i.i.d. examples

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

- For notation simplicity we may sometime use

$$z_i = (x, y)$$

Some Definitions

- Our goal is to find a hypothesis (classifier) h_S with small expected risk

$$R[h_S] = \mathbb{E}_{(x,y)\sim\mathscr{D}}\left[\ell(h_S(x); y)\right]$$

- The loss measures the disagreement between predictions and reality

Some Definitions

- Our goal is to find a hypothesis (classifier) h_S with small expected risk

$$R[h_S] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h_S(x); y)\right]$$

- The loss measures the disagreement between predictions and reality

- Since we can't directly measure $R[h_S]$ (our true cost function), we can consider optimizing its sample-average proxy, i.e., the empirical risk

$$\hat{R}[h_S] = \frac{1}{n}\sum_{i=1}^{n}\ell(h_S(x_i); y_i)$$

- Our hope is that $\hat{R}[h_S]$ is close to $R[h_S]$

The generalization gap

- The gap of the true cost function from the one we have access to

$$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]|$$

- <u>Question:</u> When is it possible to bound ϵ_{gen} by a small constant?

The generalization gap

- The gap of the true cost function from the one we have access to

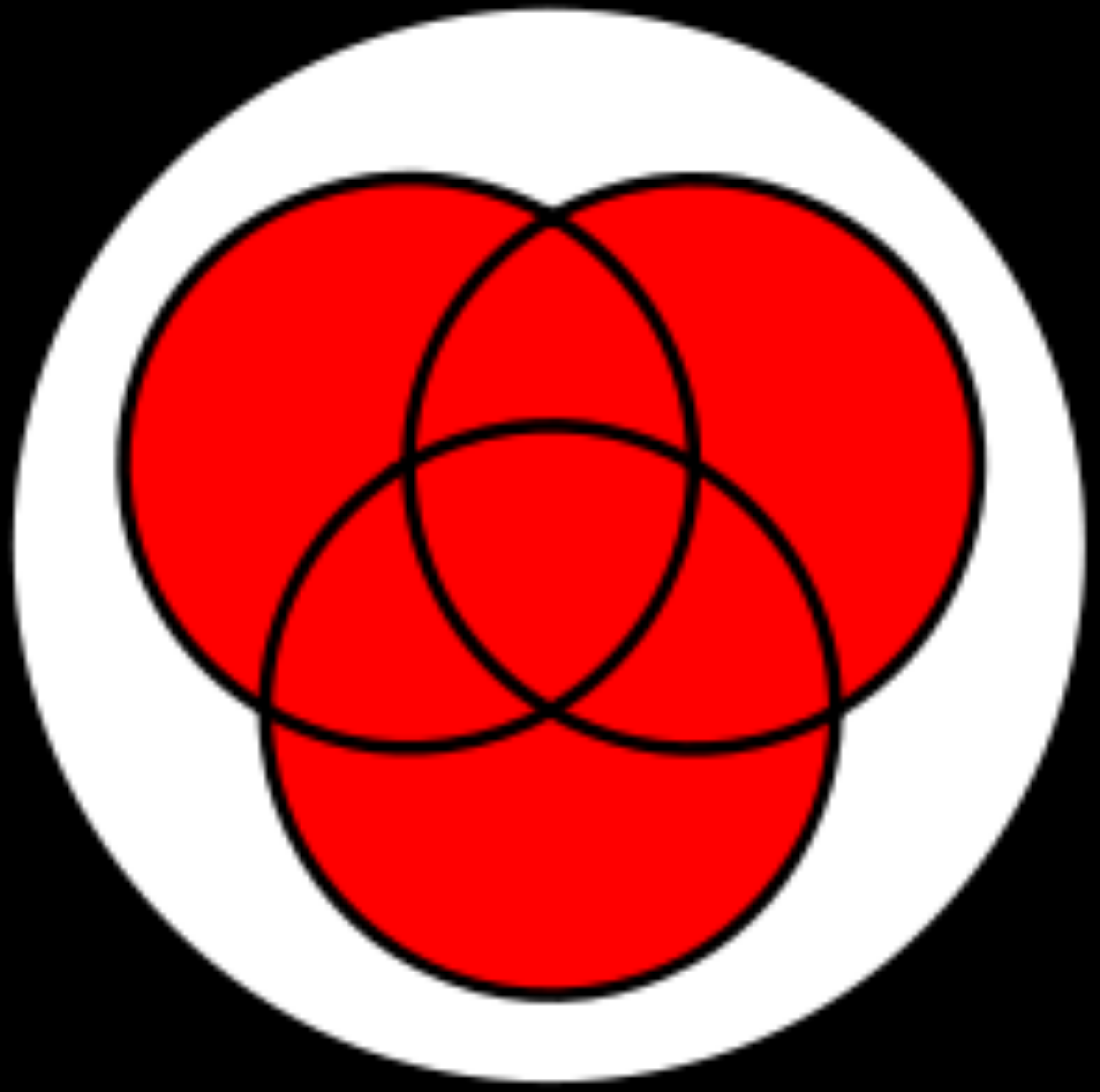
$$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]|$$

- Question: When is it possible to bound ϵ_{gen} by a small constant?

- The answer must depend on:

 1) n, the sample size
 2) $\mathscr{H}$, the sample size
 3) $\mathscr{D}$, the data distribution
 [4) the optimization algorithm that outputs our classifier]

Vanilla Union Bound Results



A first step towards concentration

- Assumption: Let the loss be bounded

$$0 \leq \ell(h(x); y) \leq 1$$

- Lets use Hoeffing's Inequality (H.I.) to prove concentration

A first step towards concentration

- Assumption: Let the loss be bounded

$$0 \leq \ell(h(x); y) \leq 1$$

- Lets use Hoeffing's Inequality (H.I.) to prove concentration

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \leq X_i \leq 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \geq 0$

$$\Pr\left(\left|\hat{X}_n - \mathbb{E}\{\hat{X}_n\}\right| \geq \epsilon\right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$$

A first step towards concentration

- Assumption: Let the loss be bounded

$$0 \leq \ell(h(x); y) \leq 1$$

- Lets use Hoeffing's Inequality (H.I.) to prove concentration

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \leq X_i \leq 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \geq 0$

$$\Pr\left(\left|\hat{X}_n - \mathbb{E}\{\hat{X}_n\}\right| \geq \epsilon\right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$$

Concentration: a random variable is behaves almost like a constant

A first step towards concentration

- <u>Assumption</u>: Let the loss be bounded

$$0 \le \ell(h(x); y) \le 1$$

- Lets use Hoeffing's Inequality (H.I.) to prove concentration

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \le X_i \le 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \ge 0$

$$\mathbf{Pr}\left(\left| \hat{X}_n - \mathbb{E}\{\hat{X}_n\} \right| \ge \epsilon \right) \le 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$$

- The above is true irrespective of the distribution of the RVs

Simple application of H.I

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \leq X_i \leq 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \geq 0$

$$\Pr\left(\left|\hat{X}_n - \mathbb{E}\{\hat{X}_n\}\right| \geq \epsilon\right) \leq 2 \cdot e^{-2\cdot n\cdot \epsilon^2}$$

- Q: How many samples n do we need to guarantee $\hat{X}_n = \mathbb{E}\hat{X}_n \pm \epsilon$ with probability $1 - \delta$?

Simple application of H.I

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \leq X_i \leq 1$. Also let, $\hat{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \geq 0$

$$\Pr\left(\left| \hat{X}_n - \mathbb{E}\{\hat{X}_n\} \right| \geq \epsilon \right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$$

- Q: How many samples n do we need to guarantee $\hat{X}_n = \mathbb{E}\hat{X}_n \pm \epsilon$ with probability $1 - \delta$?

$$\delta = 2e^{-2n\epsilon^2} \Rightarrow \log\left(\frac{\delta}{2}\right) = -2n\epsilon^2$$

$$\Rightarrow n = -\frac{\log\left(\frac{\delta}{2}\right)}{\epsilon^2} = C \cdot \frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

Simple application of H.I

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \le X_i \le 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \ge 0$

$$\mathbf{Pr}\left(\left|\hat{X}_n - \mathbb{E}\{\hat{X}_n\}\right| \ge \epsilon\right) \le 2 \cdot e^{-2\cdot n\cdot \epsilon^2}$$

- Q: How many samples n do we need to guarantee $\hat{X}_n = \mathbb{E}\hat{X}_n \pm \epsilon$ with probability $1 - \delta$?

$$\delta = 2e^{-2n\epsilon^2} \Rightarrow \log\left(\frac{\delta}{2}\right) = -2n\epsilon^2$$

$$\Rightarrow n = -\frac{\log\left(\frac{\delta}{2}\right)}{\epsilon^2} = C \cdot \frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

"Error" scales like $\sqrt{1/n}$

Simple application of H.I

Theorem: Let $X_1, \ldots, X_n \in \mathbb{R}$ be independent RVs, such that $0 \leq X_i \leq 1$. Also let, $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then, for all $\epsilon \geq 0$

$$\Pr\left(\left|\hat{X}_n - \mathbb{E}\{\hat{X}_n\}\right| \geq \epsilon\right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$$

- Q: How many samples n do we need to guarantee $\hat{X}_n = \mathbb{E}\hat{X}_n \pm \epsilon$ with probability $1 - \delta$?

Warning!

Powerful statements like this tend to be very restrictive!
H.I. is after all is oblivious to the distribution of RVs

Let's try H.I on the empirical risk

- Assume that our predictor $h(;)$ is fixed, and does not depend on the training data (what?!)

- Let $X_i = \ell(h(x_i); y_i)$. (observe that X_is are independent)

Let's try H.I on the empirical risk

- Assume that our predictor $h(;)$ is fixed, and does not depend on the training data (what?!)

- Let $X_i = \ell(h(x_i); y_i)$. (observe that X_is are independent)

- Due to the iid assumption $\mathbb{E}X_i = \mathbb{E}_{x_i \sim \mathcal{D}}[\ell(h(x_i); y_i)]$ = true risk of h

Let’s try H.I on the empirical risk

- Assume that our predictor $h(;)$ is fixed, and does not depend on the training data (what?!)

- Let $X_i = \ell(h(x_i); y_i)$. (observe that X_is are independent)

- Due to the iid assumption $\mathbb{E}X_i = \mathbb{E}_{x_i \sim \mathcal{D}}[\ell(h(x_i); y_i)]$ = true risk of h
- Then, by H.I we have $\mathbf{Pr}\left(\epsilon_{gen}[h] \geq \epsilon\right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$

Let's try H.I on the empirical risk

- Assume that our predictor $h(;)$ is fixed, and does not depend on the training data (what?!)

- Let $X_i = \ell(h(x_i); y_i)$. (observe that X_is are independent)

- Due to the iid assumption $\mathbb{E}X_i = \mathbb{E}_{x_i \sim \mathcal{D}}[\ell(h(x_i); y_i)]$ = true risk of h
- Then, by H.I we have $\mathbf{Pr}\left(\epsilon_{gen}[h] \geq \epsilon\right) \leq 2 \cdot e^{-2 \cdot n \cdot \epsilon^2}$

Corollary:

For any given (fixed) classifier h the empirical risk "converges" to the true risk with rate $\sim \dfrac{1}{\sqrt{n}}$

Let's try H.I on the empirical risk

- Assume that our predictor $h(;)$ is fixed, and does not depend on the training data (what?!)

[redacted]

Q: Is this result sufficient for ERM concentration?

No!! The result only applies to a single h

Corollary:

For any given (fixed) classifier h the empirical risk "converges" to the true risk with rate $\sim \frac{1}{\sqrt{n}}$

H.I on an entire family of classifiers

- What we need: Results for at least a family $\mathscr{H}$ of predictors

H.I on an entire family of classifiers

- <u>What we need</u>: Results for at least a family $\mathscr{H}$ of predictors

Say we are given a finite set of predictors $\mathscr{H}$ (think of a large bag that contains a lot of models). Then, we can bound the "worst-case" generalization gap for this collection of models, using the union bound and H.I.

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

H.I on an entire family of classifiers

- What we need: Results for at least a family $\mathscr{H}$ of predictors

Say we are given a finite set of predictors $\mathscr{H}$ (think of a large bag that contains a lot of models). Then, we can bound the "worst-case" generalization gap for this collection of models, using the union bound and H.I.

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

Let E_h be the event that h has generalization error more than $\epsilon_{gen}[h] = |R[h] - \hat{R}[h]| \geq \epsilon$.

H.I on an entire family of classifiers

- <u>What we need</u>: Results for at least a family $\mathscr{H}$ of predictors

Say we are given a finite set of predictors $\mathscr{H}$ (think of a large bag that contains a lot of models). Then, we can bound the "worst-case" generalization gap for this collection of models, using the union bound and H.I.

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

Let E_h be the event that h has generalization error more than $\epsilon_{gen}[h] = |R[h] - \hat{R}[h]| \geq \epsilon$.
Then,

$$\Pr\left(\max_{h\in\mathscr{H}} \epsilon_{gen}[h]\right) \leq \Pr\left(\bigcup_{h\in\mathscr{H}} \left\{\epsilon_{gen}[h] \leq \epsilon\right\}\right) \leq |\mathscr{H}| \cdot \max_{h\in\mathscr{H}} \Pr\left(\epsilon_{gen}[h]\right)$$
$$\leq |\mathscr{H}| \cdot 2e^{-2n\epsilon^2}$$

H.I on an entire family of classifiers

- <u>What we need</u>: Results for at least a family $\mathscr{H}$ of predictors

Say we are given a finite set of predictors $\mathscr{H}$ (think of a large bag that contains a lot of models). Then, we can bound the "worst-case" generalization gap for this collection of models, using the union bound and H.I.

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

Let E_h be the event that h has generalization error more than $\epsilon_{gen}[h] = |R[h] - \hat{R}[h]| \geq \epsilon$.
Then,

$$\Pr\left(\max_{h\in\mathscr{H}} \epsilon_{gen}[h]\right) \leq \Pr\left(\bigcup_{h\in\mathscr{H}} \left\{\epsilon_{gen}[h] \leq \epsilon\right\}\right) \leq |\mathscr{H}| \cdot \max_{h\in\mathscr{H}} \Pr\left(\epsilon_{gen}[h]\right)$$
$$\leq |\mathscr{H}| \cdot 2e^{-2n\epsilon^2}$$

- The above says. EVERYTHING in the $\mathscr{H}$ bag generalizes well. How big can this bag be?

H.I on an entire family of classifiers

- HI + UB can handle families of up to size $|\mathscr{H}| = O\left(2^{n\epsilon^2 \cdot \delta}\right)$
- That doesn't sound too bad!
- What about hypothesis classes that actually "learn" stuff? (e.g., linear classifiers, NNs, etc?)

Example 0: Linear Classifiers

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- $\mathscr{H}$ is the set of all hyper planes

$$|\mathscr{H}| =$$

Example 0: Linear Classifiers

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- $\mathscr{H}$ is the set of all hyper planes

$$|\mathscr{H}| = \infty$$

- Vanilla U.B. can't help us much here

Example 0: Linear Classifiers

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- $\mathscr{H}$ is the set of all hyper planes

$$|\mathscr{H}| = \infty$$

- Vanilla U.B. can't help us much here

we'll handle infinite families soon

Example 1: Linear Classifiers with finite precision

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- Let us also consider that w, b are floats (32 bits/variable)

$$|\mathcal{H}| =$$

Example 1: Linear Classifiers with finite precision

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- Let us also consider that w, b are floats (32 bits/variable)

Corollary:

For the set of all linear classifiers we have $\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

Example 1: Linear Classifiers with finite precision

- Let us consider the following binary classifier $y = sign(w^T x - b)$, where $x \in \mathbb{R}^d$.
- Let us also consider that w, b are floats (32 bits/variable)

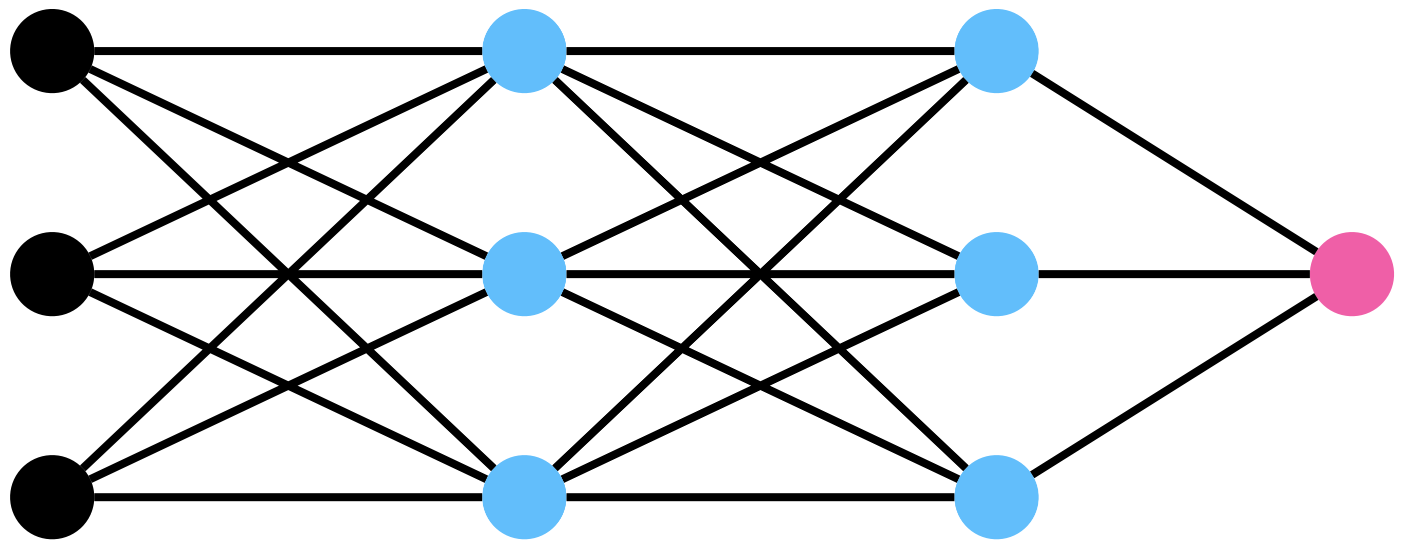
Corollary:

For the set of all linear classifiers we have $\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

When assuming floating point H.I. can be useful

Example 2: Fully Connected ReLU network with floats

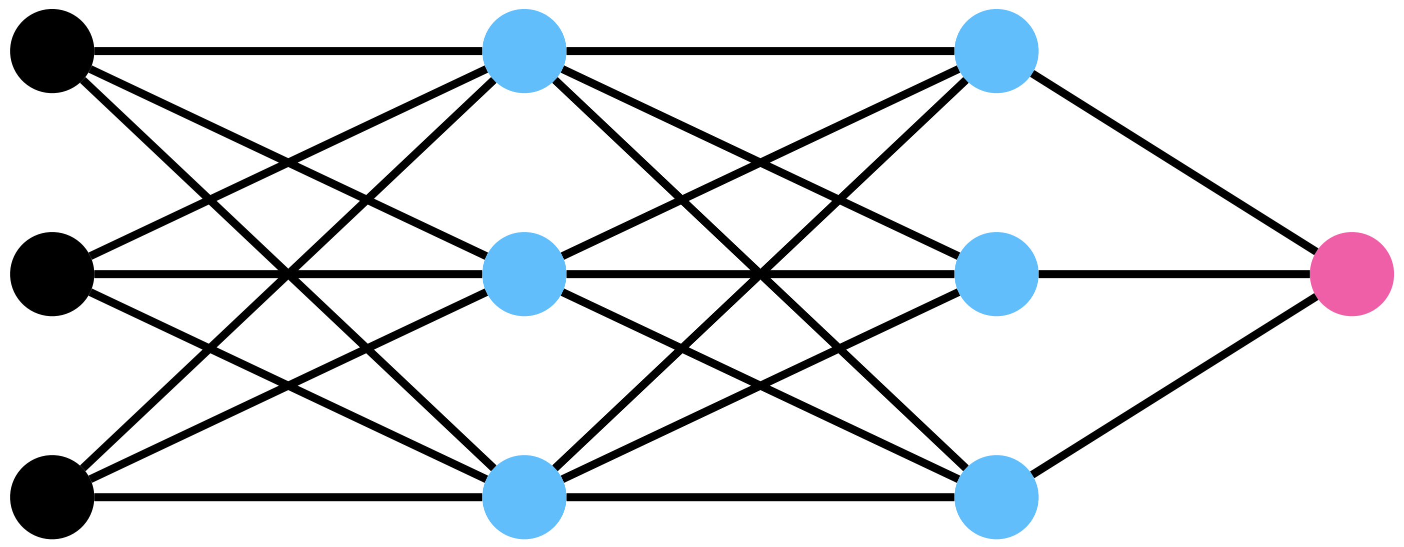
- Let us consider the following binary classifier $y = sign(h(w; x))$, where $x \in \mathbb{R}^d$, where w is the set of all weights



$$y = sign(h(w; x))$$

Example 2: Fully Connected ReLU network with floats

- Let us consider the following binary classifier $y = sign(h(w; x))$, where $x \in \mathbb{R}^d$, where w is the set of all weights
- assume they are floats (32 bit each)



$$y = sign(h(w; x))$$

$$|\mathcal{H}| =$$

Example 2: Fully Connected ReLU network with floats

- Let us consider the following binary classifier $y = sign(h(w; x))$, where $x \in \mathbb{R}^d$, where w is the set of all weights
- assume they are floats (32 bit each)

Corollary:

For the set of all finite precision NN classifiers with d weights, we have

$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

note that $d \cdot \log(32)$ is the size of the bit description of the model

Example 2: Fully Connected ReLU network with floats

- Let us consider the following binary classifier $y = sign(h(w; x))$, where $x \in \mathbb{R}^d$, where w is the set of all weights
- assume they are floats (32 bit each)

Corollary:

For the set of all finite precision NN classifiers with d weights, we have

$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

if n > #params, then <u>all</u> FCs (accurate or not) generalize.

Q: does this lead to non-vacuous bounds in practice?

Example 2.1: LeNet5 on ImageNet

Reminder:

Corollary:

For any parametric model with d parameters of finite precision, we have

$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

- LeNet5 has ~60K parameters
- ImageNet has ~1.2 million images

$$\sqrt{d/n} \approx 0.22$$

*assumes imagenet samples are iid (they are not)

Example 2.1: LeNet5 on ImageNet

Reminder:

Corollary:

For any parametric model with d parameters of finite precision, we have

$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

- LeNet5 has ~60K parameters
- ImageNet has ~1.2 million images

$$\sqrt{d/n} \approx 0.22$$

Nice!

*assumes imagenet samples are iid (they are not)

Example 2.2: ResNet50 on ImageNet

Reminder:

Corollary:

For any parametric model with d parameters of finite precision, we have

$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right)$, with probability $1 - \delta$, and any $0 < \delta < 1$

- ResNet 50 has ~23 million parameters
- ImageNet has ~1.2 million images

$$\sqrt{d/n} >> 1$$

Example 2.2: ResNet50 on ImageNet

Reminder:

Corollary:

For any parametric model with d parameters of finite precision, we have

$$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]| = O\left(\sqrt{d/n}\right), \text{ with probability } 1-\delta, \text{ and any } 0 < \delta < 1$$

- ResNet 50 has ~23 million parameters
- ImageNet has ~1.2 million images

$$\sqrt{d/n} >> 1$$

U.B. style results yield vacuous generalization error bounds

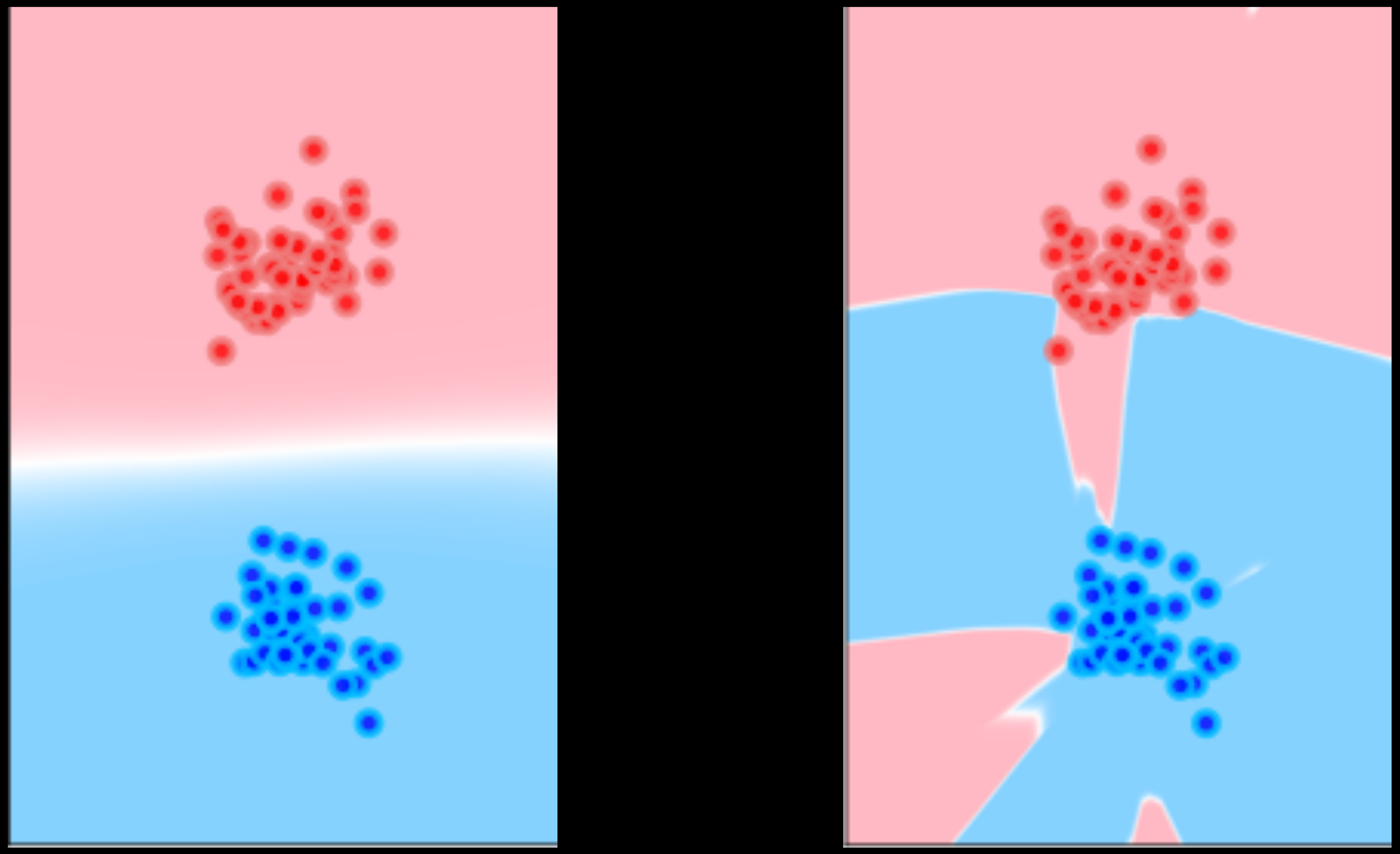
So far, only finite classes

- If Floats+parametric model => n > #params for generalization

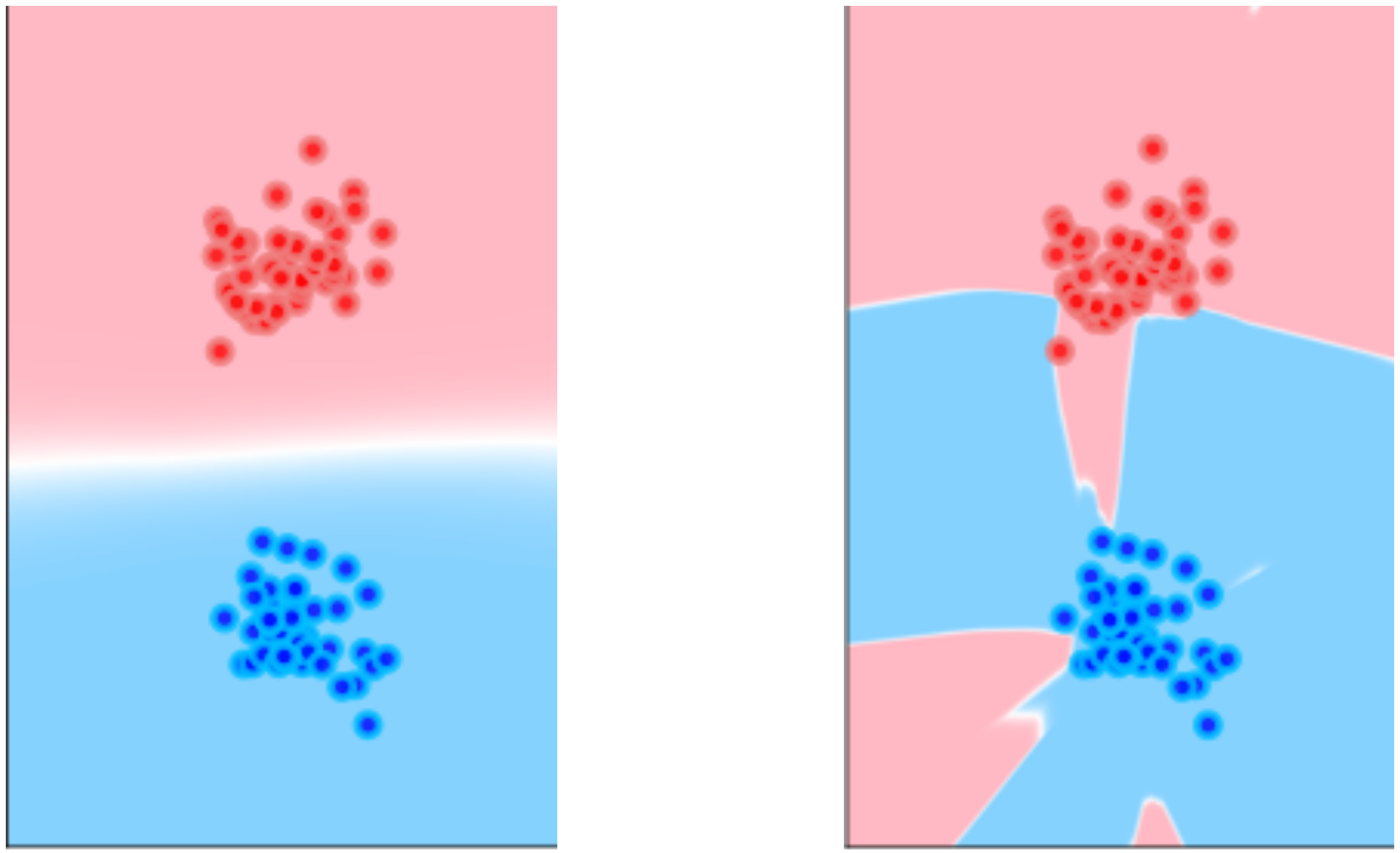
So far, only finite classes

- If Floats+parametric model => n > #params for generalization
- Traditional theory for generalization bounds tries to handle infinite classes.
- VC-dimension, fat-shattering dimension, rademacher complexity, etc
- Can these more elaborate approaches result in interesting gen bounds for real models/data?

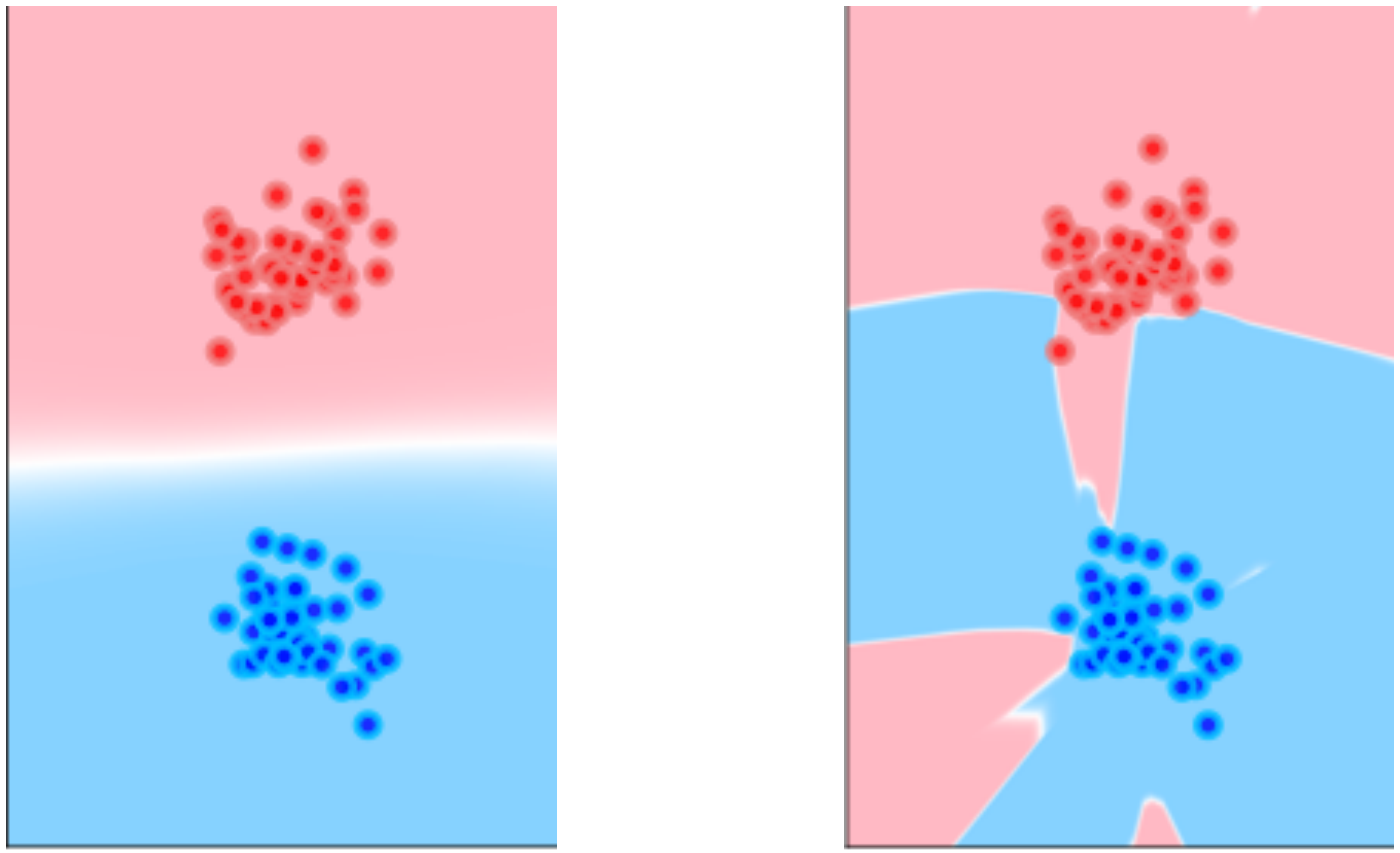
Measuring Complexity



Which one is more complex?



Which one is more complex?



"complexity" not captured by "raw" bit complexity/param count of a model

Bounding generalization via complexity measure

- General idea:

Bounding the expressiveness of a model => bounding the number of bits needed to describe it
=> bounding the generalization gap.

In other words, the less expressive/complex a class, the less surprises we'll have at test time.

Bounding generalization via complexity measure

- General idea:

Bounding the expressiveness of a model => bounding the number of bits needed to describe it => bounding the generalization gap.

In other words, the less expressive/complex a class, the less surprises we'll have at test time.

- Standard techniques: VC dimension and Rademacher Complexity

- Q: How do they work, what types of bounds do they imply?

VC dimension

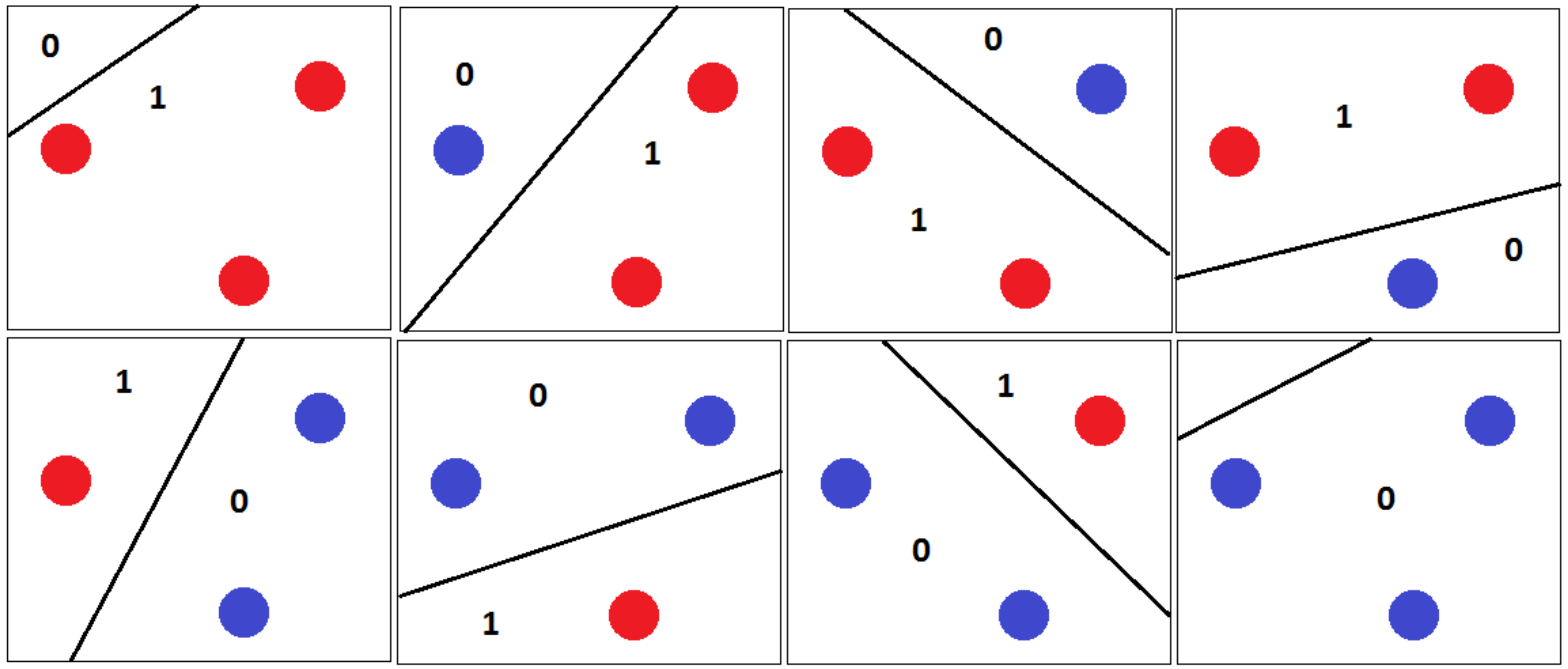
- VC dimension = measures expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., for any labels $y_1, \ldots, y_n$ of S, $h(x_i) = y_i$ for all $x_i \in S$

VC dimension

- VC dimension = measures expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., for any labels $y_1, \ldots, y_n$ of S, $h(x_i) = y_i$ for all $x_i \in S$



VC dimension

- VC dimension = measures expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathcal{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathcal{H}$, i.e., for any labels $y_1, \ldots, y_n$ of S, $h(x_i) = y_i$ for all $x_i \in S$

- E.g., largest set of images that a classifier can give any set of labels.

VC dimension

- VC dimension = measures expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., for any labels $y_1, \ldots, y_n$ of S, $h(x_i) = y_i$ for all $x_i \in S$

- E.g., largest set of images that a classifier can give any set of labels.

- Similar to memorization, but not quite.

VC dimension

- VC dimension = measures expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., for any labels $y_1, \ldots, y_n$ of S, $h(x_i) = y_i$ for all $x_i \in S$

- E.g., largest set of images that a classifier can give any set of labels.

- Similar to memorization, but not quite.

- Q: how does VC connect with generalization error?

VC dimension

- VC dimension can handle infinite classes

Theorem:

For any $\epsilon, \delta > 0$, suppose that $VCdim(\mathscr{H}) = d$, and we draw a sample S of size

$$n \geq \frac{C}{\epsilon^2}\left(d \log(1/\epsilon) + \log(1/\delta)\right)$$

then with probability at least $1 - \delta$, we have that $\max_{h \in \mathscr{H}} \epsilon_{gen}[h] \leq \epsilon$

VC dimension

- VC dimension can handle infinite classes

Theorem:

For any $\epsilon, \delta > 0$, suppose that $VCdim(\mathscr{H}) = d$, and we draw a sample S of size

$$n \geq \frac{C}{\epsilon^2}\left(d\log(1/\epsilon) + \log(1/\delta)\right)$$

then with probability at least $1 - \delta$, we have that $\max_{h \in \mathscr{H}} \epsilon_{gen}[h] \leq \epsilon$

We need again $n > VC(\mathscr{H})$, for good generalization

Q: does this lead to non-vacuous bounds in practice?

VC dimension

- VC dimension = measure of expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., if $y_1, \ldots, y_n$ are the labels of S, then $h(x_i) = y_i$ for all $(x_i, y_i) \in S$

VC dimension

- VC dimension = measure of expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., if $y_1, \ldots, y_n$ are the labels of S, then $h(x_i) = y_i$ for all $(x_i, y_i) \in S$

Examples:

- $\mathscr{H} = \left\{h \,|\, h(x) = sign(w^T x - b)\right\}$, $VC(\mathscr{H}) = d + 1$
- $\mathscr{H}$ = neural nets with thresholds and d parameters, $VC(\mathscr{H}) = O(d \log d)$
- $\mathscr{H}$ = ReLU NNs with d parameters and depth D $VC(\mathscr{H}) = O(dD \log d)$

VC dimension

- VC dimension = measure of expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathscr{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathscr{H}$, i.e., if $y_1, \ldots, y_n$ are the labels of S, then $h(x_i) = y_i$ for all $(x_i, y_i) \in S$

Examples:
- $\mathscr{H} = \left\{h \,|\, h(x) = sign(w^T x - b)\right\}, VC(\mathscr{H}) = d + 1$
- $\mathscr{H}$ = neural nets with thresholds and d parameters, $VC(\mathscr{H}) = O(d \log d)$
- $\mathscr{H}$ = ReLU NNs with d parameters and depth D $VC(\mathscr{H}) = O(dD \log d)$

- For NNs it seems that VC dimension > #params.. Worse generalization than parameter count on FP networks…

VC dimension

- VC dimension = measure of expressiveness of a hypothesis class

Definition:
The VC-dimension of $\mathcal{H}$ is the largest number d such that there exist a set S of d samples that is shattered by a classifier $h \in \mathcal{H}$, i.e., if $y_1, \ldots, y_n$ are the labels of S, then $h(x_i) = y_i$ for all $(x_i, y_i) \in S$

Examples:
- $\mathcal{H} = \{h \mid h(x) = sign(w^T x - b)\}, VC(\mathcal{H}) = d + 1$

[redacted]

For finite Precision VC doesn't lead to anything better than the simple UB technique from earlier…

[redacted]

Conclusion

- Concentration of the ERM implies generalization
- Algorithm/Data agnostic generalization bounds are… tricky
- Next: Can we refine these bounds?