

Lecture 6. PAC Learning Theory

COMP90051 Statistical Machine Learning

Lecturer: Feng Liu



THE UNIVERSITY OF
MELBOURNE

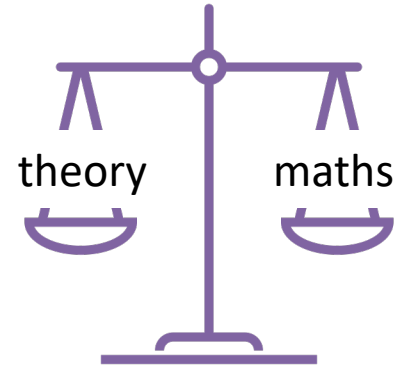
This lecture

- Excess risk
 - * Decomposition: Estimation vs approximation
 - * Bayes risk – irreducible error
- Probably approximation correct learning
- Bounding generalisation error with high probability
 - * Single model: Hoeffding's inequality
 - * Finite model class: Also use the union bound
- Importance & limitations of uniform deviation bounds



Generalisation and Model Complexity

- Theory we've seen so far (mostly statistics)
 - * Asymptotic notions (consistency, efficiency)
 - * Convergence could be really slow
 - * Model complexity undefined
- Want: finite sample theory; convergence *rates*, trade-offs
- Want: define model complexity and relate it to test error
 - * Test error can't be measured in real life, but it can be provably bounded!
 - * Growth function, VC dimension
- Want: distribution-independent, learner-independent theory
 - * A fundamental theory applicable *throughout ML*
 - * Unlike bias-variance: distribution dependent, no model complexity,



Probably Approximately Correct Learning

*The bedrock of machine learning theory in
computer science.*

Standard setup

Problem we consider here: Supervised binary classification of

- data in \mathcal{X} into label set $\mathcal{Y} = \{-1, 1\}$

What we have:

- iid data $D^{\text{train}} = \{(x_i, y_i)\}_{i=1}^m \sim D$ some fixed unknown distribution. The D^{train} is called training data.
- **Training error** of a function f on D^{train} can be expressed by
$$\hat{R}[f] = \frac{1}{m} \sum_{i=1}^m \ell(y_i, f(x_i)).$$

What we will do in supervised binary classification:

- Learn a function f_m from a class of function \mathcal{F} mapping (classifying) \mathcal{X} into \mathcal{Y} such that $f_m = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f]$.

Standard setup

Now, we have

$$\triangleright f_m = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f] = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(y_i, f(x_i))$$

and **want to**

analyse the performance of f_m on **new data** from the fixed distribution D .

Can you write down the test error based on f_m and D ?

Standard setup


Now, we have

$$\triangleright f_m = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f] = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(y_i, f(x_i))$$

and **want to**

analyse the performance of f_m on **new data** from the fixed distribution D .

Can you write down the test error based on f_m and D ?

A lower $R[f_m]$ is better.  $R[f_m] = \mathbb{E}_{(X,Y) \sim D} [\ell(Y, f_m(X))]$ to represent the risk (or test error) of f_m on D .

Standard setup

Now, we have

$$\triangleright f_m = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f] = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(y_i, f(x_i))$$

and **want to (The Theoretical AIM)**

$$\text{analyse } R[f_m] = \mathbb{E}_{(X,Y) \sim D} [\ell(Y, f_m(X))]$$

- What parts depend on the sample of data
 - Empirical risk $\hat{R}[f]$ that averages loss over the sample
 - $f_m \in \mathcal{F}$ the learned model (it could be same sample or different; theory is actually fully general here)

The Bayes Risk: One thing we cannot ignore

- We usually cannot even hope for perfection!
 - * $R^* \in \inf_f R[f]$ called the **Bayes risk**;
 - * **cannot** expect zero $R[f]$ and a clear decision boundary.
- Thus, we care about the following risk more:

$$R[f_m] - R^*$$

Excess risk

Decomposed Risk: The good, bad and ugly

$$R[f_m] - R^* = (R[f_m] - R[f^*]) + (R[f^*] - R^*)$$

- **Good:** what we'd aim for in our class, with infinite data
 - * $R[f^*]$ true risk of **best in class** $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R[f]$
- **Bad:** we get what we get and don't get upset
 - * $R[f_m]$ true risk of **learned** $f_m \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f] + C\|f\|^2$ (e.g.)
- **Ugly:** we usually cannot even hope for perfection!
 - * $R^* \in \inf_f R[f]$ called the **Bayes risk**;
 - * **cannot** expect zero $R[f]$ and a clear decision boundary.

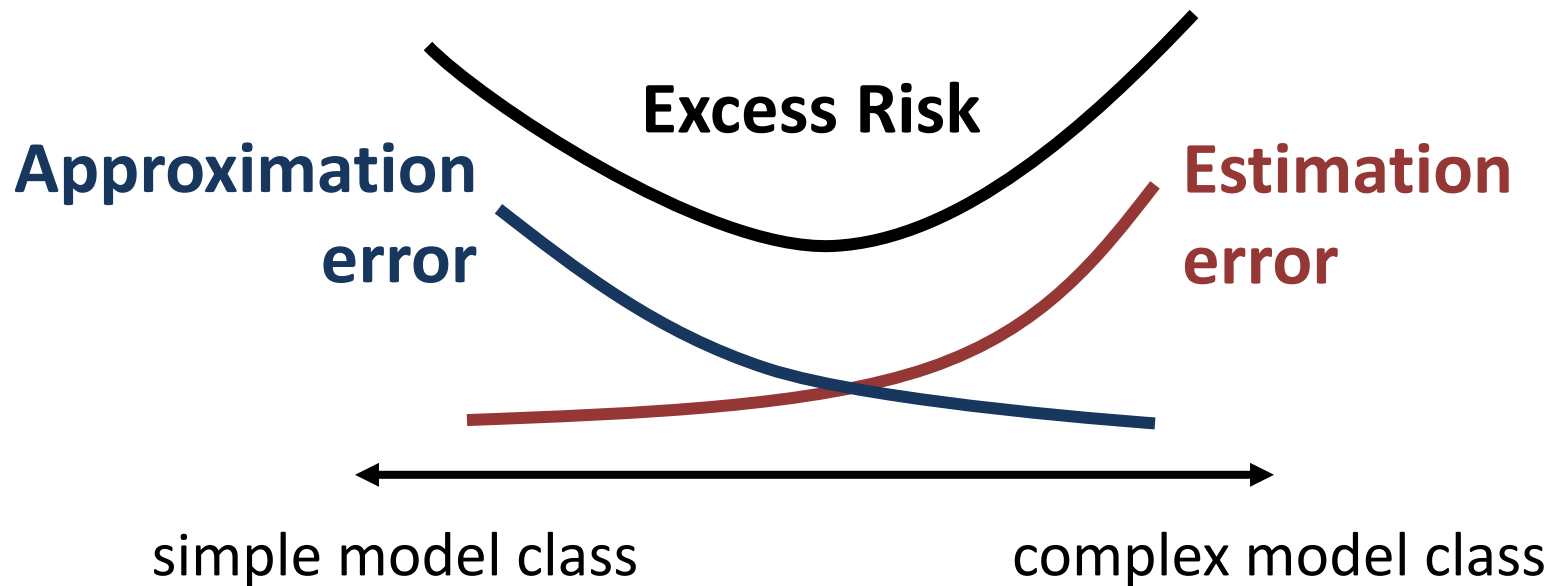
Decomposed Risk: The good, bad and ugly

$$\underbrace{R[f_m] - R^*}_{\text{Excess risk}} = \underbrace{(R[f_m] - R[f^*])}_{\text{Estimation error}} + \underbrace{(R[f^*] - R^*)}_{\text{Approximation error}}$$

- **Good**: what we'd aim for in our class, with infinite data
 - * $R[f^*]$ true risk of **best in class** $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R[f]$
- **Bad**: we get what we get and don't get upset
 - * $R[f_m]$ true risk of **learned** $f_m \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}[f] + C\|f\|^2$ (e.g.)
- **Ugly**: we usually cannot even hope for perfection!
 - * $R^* \in \inf_f R[f]$ called the **Bayes risk**;
 - * **cannot** expect zero $R[f]$ and a clear decision boundary.

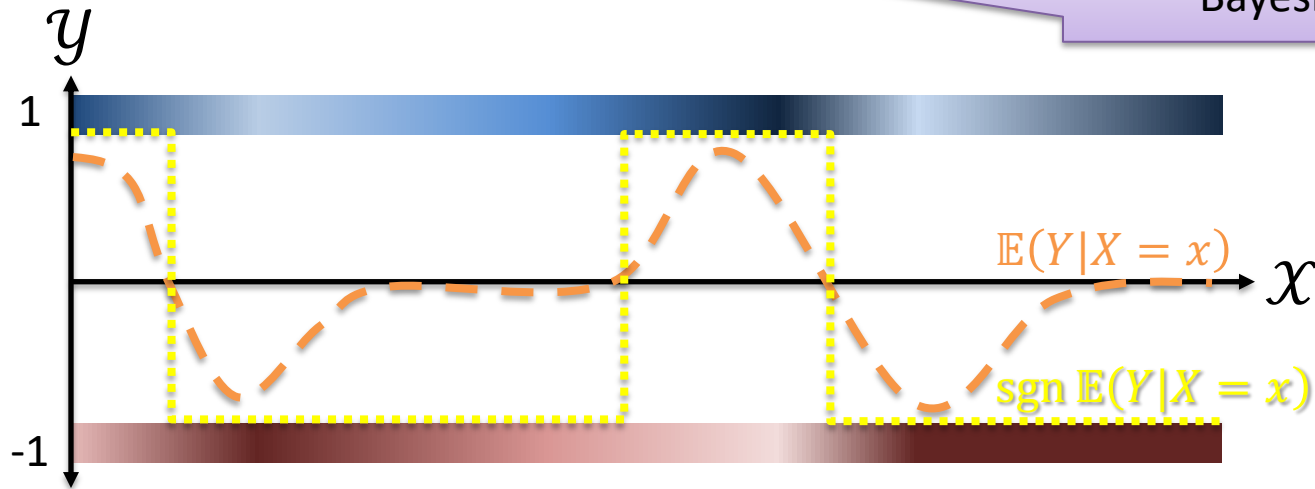
A familiar trade-off: More intuition

- simple family \rightarrow may underfit due to approximation error
- complex family \rightarrow may overfit due to estimation error



About Bayes risk

Named after Bayes. Not Bayesian ML.



- **Bayes risk** $R^* \in \inf_f R[f]$
 - * Best risk possible, ever; but can be large
 - * Depends on distribution and loss function
- **Bayes classifier** achieves Bayes risk
 - * $f_{\text{Bayes}}(x) = \text{sgn } \mathbb{E}(Y|X=x)$

Let's focus on $R[f_m]$



Leslie Valiant
CCA2.0 Renate Schmid

- Since we don't know data distribution, we need to bound generalisation to be small
 - * Bound by test error $\hat{R}[f_m] = \frac{1}{m} \sum_{i=1}^m f(X_i, Y_i)$
 - * Abusing notation: $f(X_i, Y_i) = l(Y_i, f(X_i))$
$$R[f_m] \leq \hat{R}[f_m] + \varepsilon(m, \mathcal{F})$$
- Unlucky training sets, no always-guarantees possible!
- With probability $\geq 1 - \delta$: $R[f_m] \leq \hat{R}[f_m] + \varepsilon(m, \mathcal{F}, \delta)$
- Called Probably Approximately Correct (**PAC**) learning
 - * \mathcal{F} called **PAC learnable** if $m = O(\text{poly}(1/\varepsilon, 1/\delta))$ to learn f_m for any ε, δ
 - * Won Leslie Valiant (Harvard) the 2010 **Turing Award**
- Later: Why this bounds estimation error.

Don't require
exponential growth
in training size m

Mini Summary

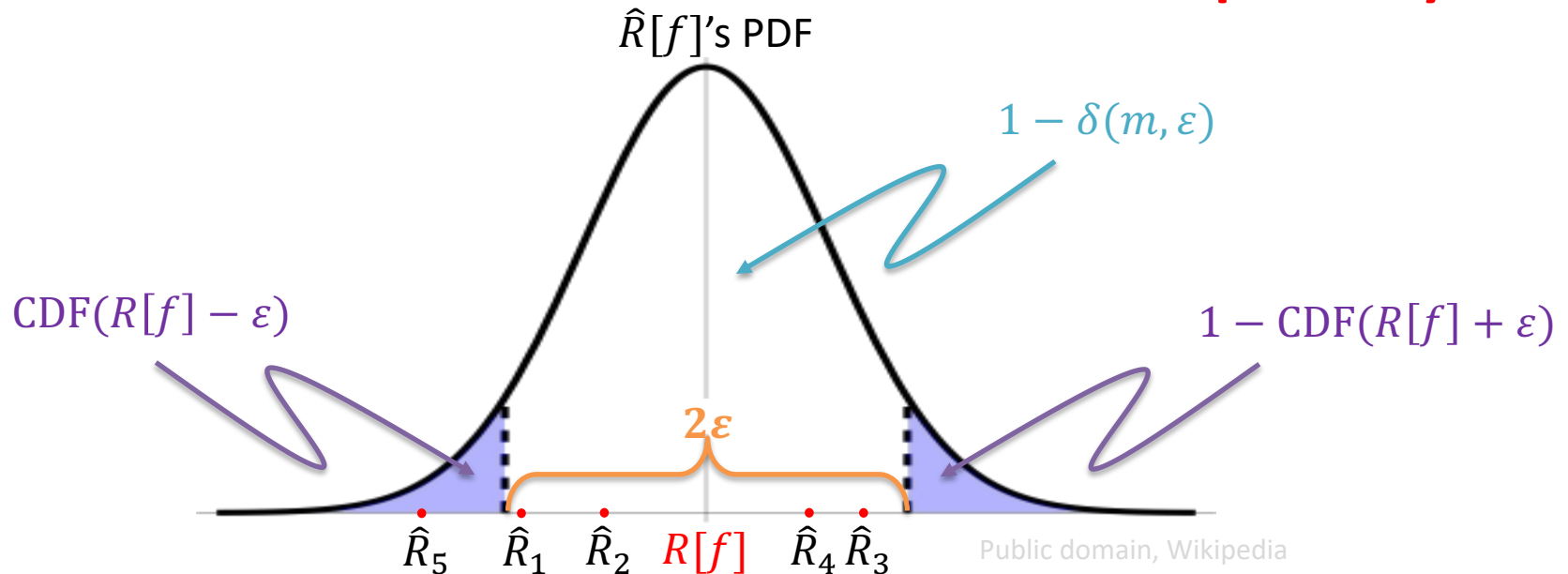
- Excess risk as the goal of ML
- Decomposition into approximation, estimation errors
- Probably Approximately Correct (PAC) learning
 - * Like asymptotic theory in stats, but for finite sample size
 - * Worst-case on distributions: We don't want to assume something unrealistic about where the data comes from
 - * Worst-case on models: We don't want a theory that applies to narrow set of learners, but to ML in general
 - * We want it to produce a useful measure of model complexity

Next: First step to PAC theory – bounding single model risk

Bounding true risk of one function

One step at a time

We need a concentration inequality



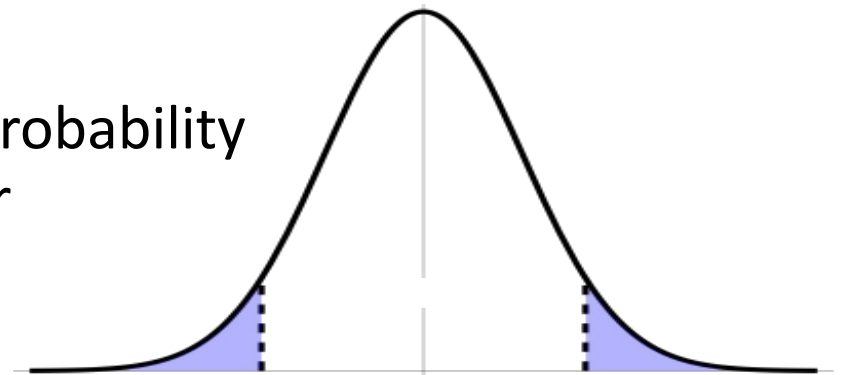
- $\hat{R}[f]$ is an unbiased estimate of $R[f]$ for any fixed f (*why?*)
- That means on average $\hat{R}[f]$ lands on $R[f]$
- What's the likelihood $1 - \delta$ that $\hat{R}[f]$ lands within ε of $R[f]$? Or more precisely, what $1 - \delta(m, \varepsilon)$ achieves a given $\varepsilon > 0$?
- Intuition: Just bounding CDF of $\hat{R}[f]$, independent of distribution!!

Hoeffding's inequality

- Many such concentration inequalities; a simplest one...
- **Theorem:** Let Z_1, \dots, Z_m, Z be iid random variables and $h(z) \in [a, b]$ be a bounded function. For all $\varepsilon > 0$

$$\Pr\left(\left|\mathbb{E}[h(Z)] - \frac{1}{m} \sum_{i=1}^m h(Z_i)\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right)$$
$$\Pr\left(\mathbb{E}[h(Z)] - \frac{1}{m} \sum_{i=1}^m h(Z_i) \geq \varepsilon\right) \leq \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right)$$

- Two-sided case in words: The probability that the empirical average is far from the expectation is **small**.



Public domain, Wikipedia

Et voila: A bound on true risk!

Result! $R[f] \leq \hat{R}[f] + \sqrt{\frac{\log(1/\delta)}{2m}}$ with high probability (**w.h.p.**) $\geq 1 - \delta$

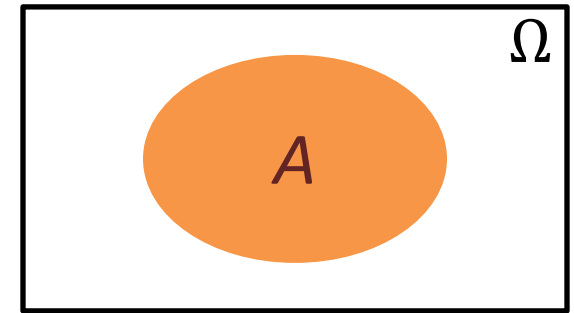
Proof

- Take the Z_i as labelled examples (X_i, Y_i)
- Take $h(X, Y) = l(Y, f(X))$ zero-one loss for some fixed $f \in \mathcal{F}$
then $h(X, Y) \in [0, 1]$
- Apply one-sided Hoeffding: $\Pr(R[f] - \hat{R}[f] \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$
- Then, substitute $\varepsilon = \sqrt{\frac{\log(1/\delta)}{2m}}$ into the above inequality, we have
- $\Pr\left(R[f] - \hat{R}[f] \geq \sqrt{\frac{\log(1/\delta)}{2m}}\right) \leq \delta$, i.e., $\Pr\left(R[f] - \hat{R}[f] \leq \sqrt{\frac{\log(1/\delta)}{2m}}\right) \geq 1 - \delta$

Common probability 'tricks'

- Inversion:

- * For any event A , $\Pr(\bar{A}) = 1 - \Pr(A)$
- * Application: $\Pr(X > \varepsilon) \leq \delta$
implies $\Pr(X \leq \varepsilon) \geq 1 - \delta$



- Solving for, in high-probability bounds:

- * For given ε with $\delta(\varepsilon)$ function ε : $\Pr(X > \varepsilon) \leq \delta(\varepsilon)$
- * Given δ' can write $\varepsilon = \delta^{-1}(\delta')$: $\Pr(X > \delta^{-1}(\delta')) \leq \delta'$
- * Let's you specify either parameter
- * Sometimes sample size m a variable we can solve for too

Try to derive the bound on your own!

Mini Summary

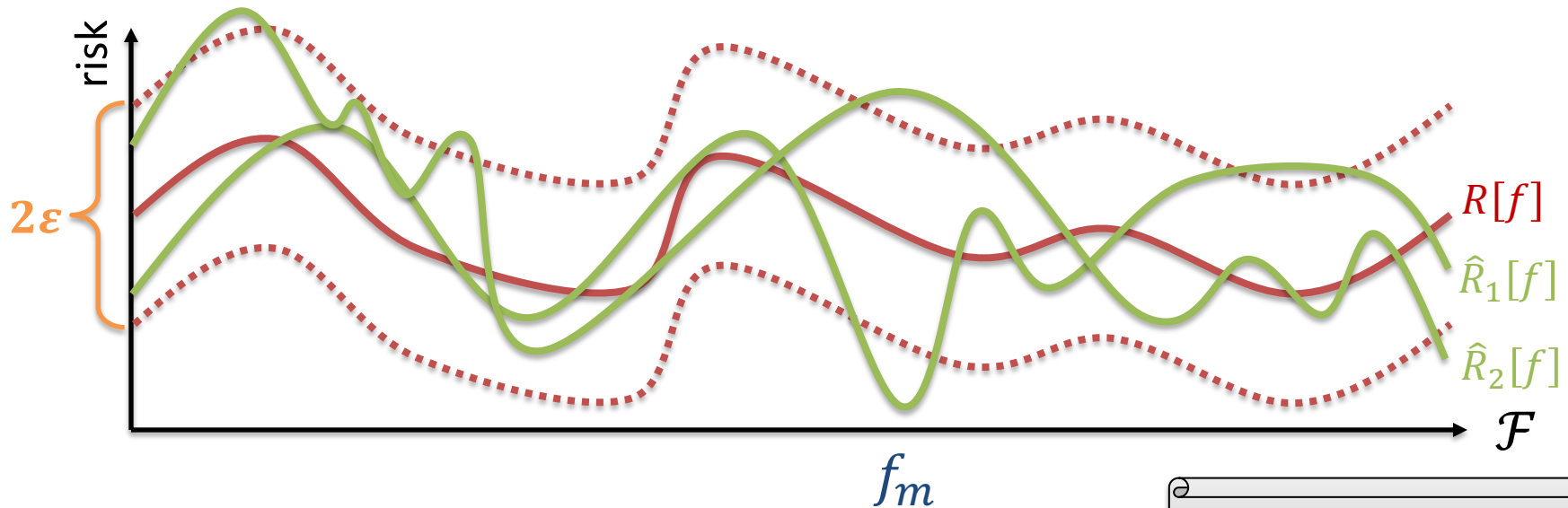
- Goal: Bound true risk of a classifier based on its empirical risk plus “stuff”
- Caveat: Bound is “with high probability” since we could be unlucky with the data
- Approach: Hoeffding’s inequality which bounds how far a mean is likely to be from an expectation

Next: PAC learning as uniform deviation bounds

Uniform deviation bounds

*Why we need our bound to **simultaneously** (or uniformly) hold over a family of functions.*

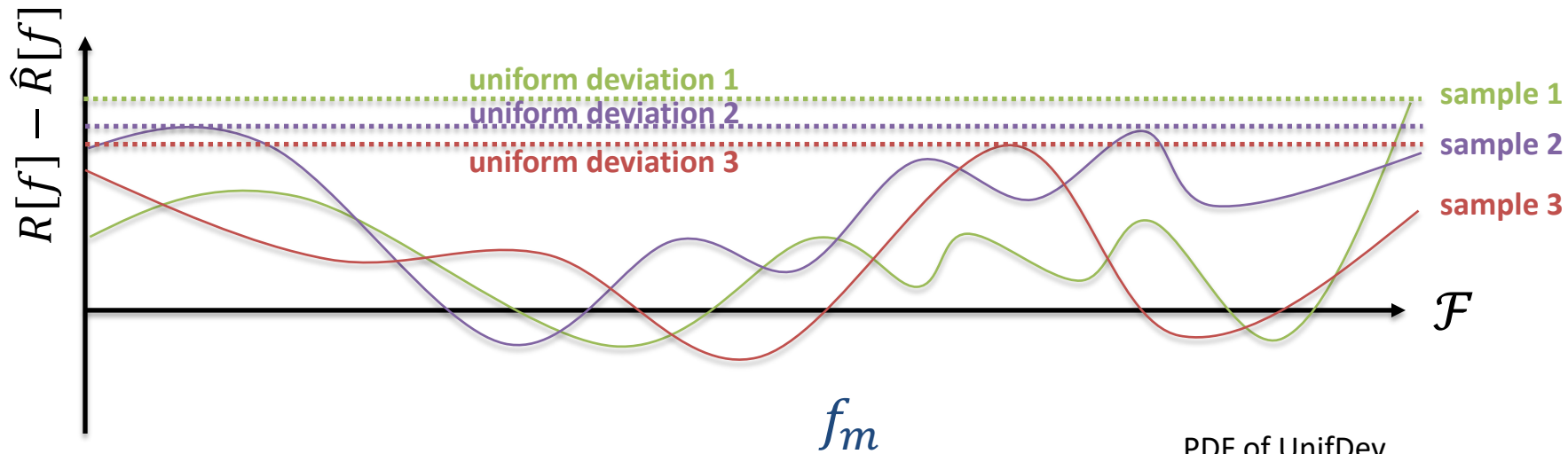
Our bound doesn't hold for $f = f_m$



Problematic that f_m depends on data

- Result says there's set S of good samples for which $R[f] \leq \hat{R}[f] + \sqrt{\frac{\log(1/\delta)}{2m}}$ and $\Pr(\mathbf{Z} \in S) \geq 1 - \delta$
- But for different functions f_1, f_2, \dots we might get very different sets S_1, S_2, \dots
- S observed may be bad for f_m . Learning minimises $\hat{R}[f_m]$, exacerbating this

Uniform deviation bounds

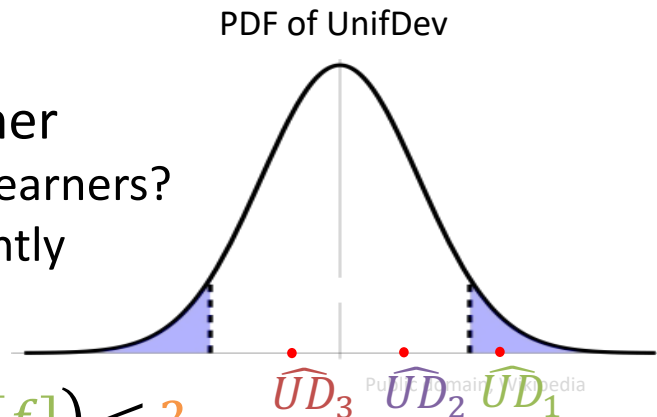


- We could analyse risks of f_m from specific learner
 - * But repeating for new learners? How to compare learners?
 - * Note there are ways to do this, and data-dependently

- Bound uniform deviations across whole class \mathcal{F}

$$R[f_m] - \hat{R}[f_m] \leq \sup_{f \in \mathcal{F}} (R[f] - \hat{R}[f]) \leq ?$$

- * Worst deviation over an entire class bounds learned risk!
- * Convenient, but could be much worse than the actual gap for f_m



Relation to estimation error?

- Recall **estimation error**? *Learning* part of excess risk!

$$R[f_m] - R^* = (\textcolor{red}{R[f_m]} - \textcolor{red}{R[f^*]}) + (R[f^*] - R^*)$$

Theorem: ERM's estimation error is at most twice the uniform divergence



- * Proof: a bunch of algebra!

$$\begin{aligned} R[f_m] &\leq (\hat{R}[f^*] - \hat{R}[f_m]) + R[f_m] - R[f^*] + R[f^*] \\ &= \hat{R}[f^*] - R[f^*] + R[f_m] - \hat{R}[f_m] + R[f^*] \\ &\leq |R[f^*] - \hat{R}[f^*]| + |R[f_m] - \hat{R}[f_m]| + R[f^*] \\ &\leq 2 \sup_{f \in \mathcal{F}} |R[f] - \hat{R}[f]| + R[f^*] \end{aligned}$$

Mini Summary

- Why Hoeffding doesn't cover a model f_m learned from data, only a fixed data-independent f
- Uniform deviation idea: Cover the worst case deviation between risk and empirical risk, across \mathcal{F}
- Advantages: works for any learner, data distribution
- Connection back to bounding estimation error

Next: Next step for PAC learning – finite classes

Error bound for finite function classes

Our first uniform deviation bound

The Union Bound

- If each model f having large risk deviation is a “bad event”, we need a tool to bound the probability that any bad event happens. I.e. the union of bad events!
- **Union bound:** for a sequence of events $A_1, A_2 \dots$

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

Proof:

Define $B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j$ with $B_1 = A_1$.

1. We know: $\bigcup_i B_i = \bigcup_i A_i$ (could prove by induction)
2. The B_i are disjoint (empty intersections)
3. We know: $B_i \subseteq A_i$ so $\Pr(B_i) \leq \Pr(A_i)$ by monotonicity
4. $\Pr(\bigcup_i A_i) = \Pr(\bigcup_i B_i) = \sum_i \Pr(B_i) \leq \sum_i \Pr(A_i)$

Bound for finite classes \mathcal{F}

- A uniform deviation bound over *any* finite class or distribution

Theorem: Consider any $\delta > 0$ and *finite* class \mathcal{F} . Then w.h.p

at least $1 - \delta$: For all $f \in \mathcal{F}$, $R[f] \leq \hat{R}[f] + \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{2m}}$

Proof:

- If each model f having large risk deviation is a “bad event”, we bound the probability that any bad event happens.
- $\Pr(\exists f \in \mathcal{F}, R[f] - \hat{R}[f] \geq \varepsilon) \leq \sum_{f \in \mathcal{F}} \Pr(R[f] - \hat{R}[f] \geq \varepsilon)$
- $\leq |\mathcal{F}| \exp(-2m\varepsilon^2)$ by the union bound
- Followed by inversion, setting $\delta = |\mathcal{F}| \exp(-2m\varepsilon^2)$

Discussion

- Hoeffding's inequality only uses boundedness of the loss, not the variance of the loss random variables
 - * Fancier concentration inequalities leverage variance
- Uniform deviation is worst-case, ERM on a very large over-parametrised \mathcal{F} may approach the worst-case, but learners generally may not
 - * Custom analysis, data-dependent bounds, PAC-Bayes, etc.
- Dependent data?
 - * Martingale theory
- Union bound is in general loose, as bad is if all the bad events were independent (not necessarily the case even though underlying data modelled as independent); and **finite** \mathcal{F}
 - * VC theory coming up next!

Mini Summary

- More on uniform deviation bounds
- The union bound (generic tool in probability theory)
- Finite classes: Bounding uniform deviation with union+Hoeffding

Next time: PAC learning with infinite function classes!