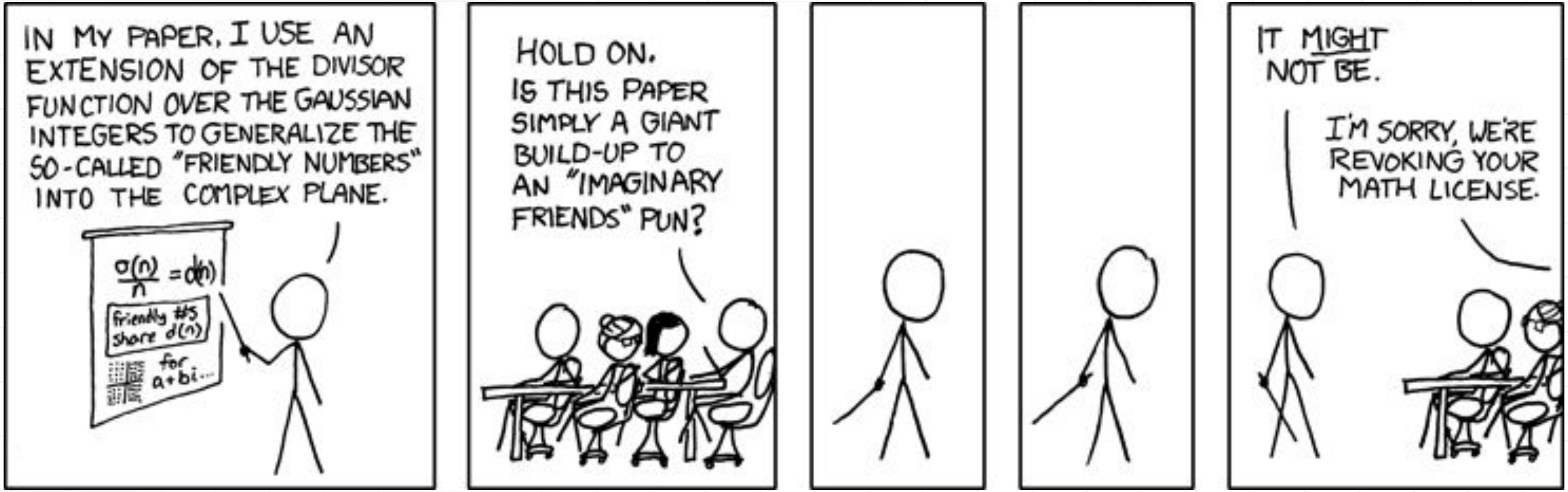
https://xkcd.com/410/


IN MY PAPER, I USE AN EXTENSION OF THE DIVISOR FUNCTION OVER THE GAUSSIAN INTEGERS TO GENERALIZE THE SO-CALLED "FRIENDLY NUMBERS" INTO THE COMPLEX PLANE.
$\frac{\sigma(n)}{n} = d(n)$
friendly #s share d(n)
for a+bi...
HOLD ON. IS THIS PAPER SIMPLY A GIANT BUILD-UP TO AN "IMAGINARY FRIENDS" PUN?
IT MIGHT NOT BE.
I'M SORRY, WE'RE REVOKING YOUR MATH LICENSE.

announcements

Quiz 1 – open until Fri 10am

Assignment 1 – due in < 2 weeks (Mon noon week 6)

(a bite-sized intro to) Generalisation

The MLStory book
https://mlstory.org/generalization.html

High level questions for today:

Why learning works? Why over-parameterisation works?

The notion of generalization gap

Overparameterization: empirical phenomena

Prelude: three inequalities

Theories of generalization

- Algorithmic stability
- Model complexity and uniform convergence
- Generalization from algorithms

Moritz Hardt

Director
Social Foundations of Computation
Max Planck Institute for Intelligent Systems, Tübingen

Associate Professor, on leave
Electrical Engineering and Computer Sciences
University of California, Berkeley

Benjamin Recht
Professor, UC Berkeley

Peter Bartlett

Bob Williamson

Notations - Loss function and risks

predictor

$$R[f] = \mathbb{E}\left[loss(f(X), Y)\right]$$

Risk

E - over true data distribution (X, Y)
– we don’t have access to this

target

Stretched notation, loss on one data point (x,y)

$$loss(f, (x, y))$$

$$loss(w, (x, y))$$

$$S = ((x_1, y_1), \ldots\ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n \,.$$

A sample as ordered tuples

$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} loss(f(x_i), y_i) \,.$$

Empirical risk for this sample

Empirical risk minimisation (ERM)

seeks to find a predictor f* in a in a specified class Ƒ that minimizes the empirical risk

$$R_S[f^*] = \min_{f \in \mathcal{F}} R_S[f]$$

min. training error, training loss

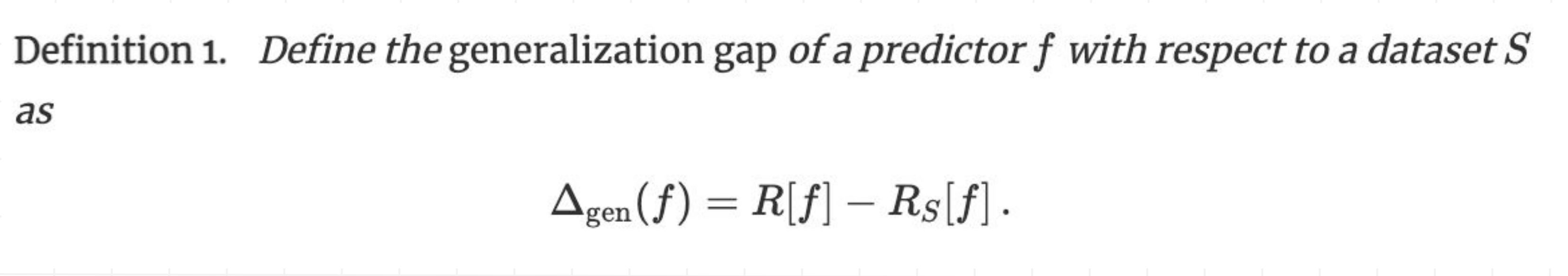
Ideally $$R_S[f] \approx R[f].$$

loss on seen examples

loss on unseen (and seen) examples

We expect this to be worse (larger loss/risk)

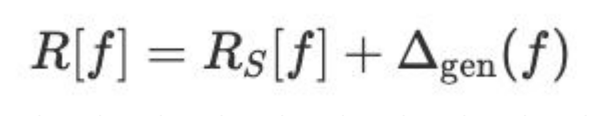
Generalisation gap



Definition 1. *Define the* generalization gap *of a predictor f with respect to a dataset S as*

$$\Delta_{\text{gen}}(f) = R[f] - R_S[f]\,.$$

Aka *generalisation error*, or *excess risk*



$$R[f] = R_S[f] + \Delta_{\text{gen}}(f)$$

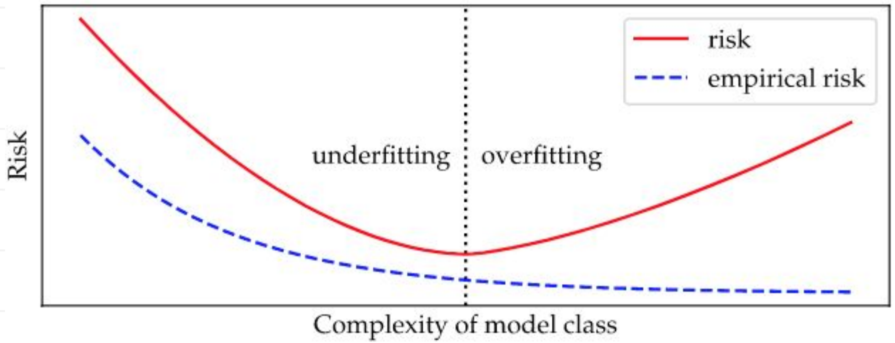
"If we manage to make the empirical risk small through optimization (most of this class and other ML classes), then all that remains to worry about is generalization gap."

But how?

Evidence from ML practice: overparameterization

Model size / complexity (informally): number of trainable parameters, for a given model family.

Traditional view of generalization



old theory: over-parameterisation is bad

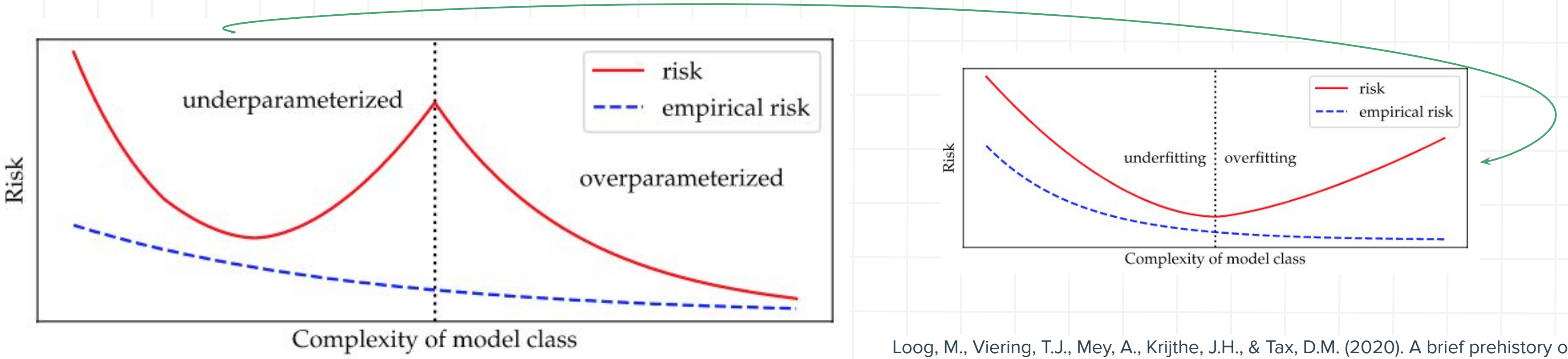
$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.73}$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2} M \ln N \tag{4.139}$$

M - number of parameters; N - number of points

Evidence from ML practice: double descent

- Complex models also can simultaneously achieve close to zero training loss and still generalize well
- Risk continues to decreases as model complexity grows and training data are interpolated exactly down to (nearly) zero training loss=
- Empirical relationship between overparameterization and risk appears to be robust and manifests in numerous model classes, including overparameterized linear models, ensemble methods, and neural networks.
- Increasing model complexity in the overparameterized regime continues to decrease risk indefinitely, albeit at decreasing marginal returns, toward some convergence point.



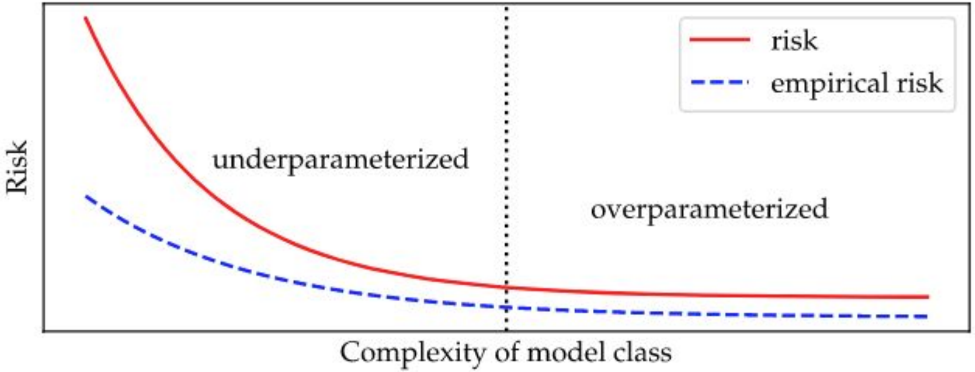
M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15849–15854 (2019).

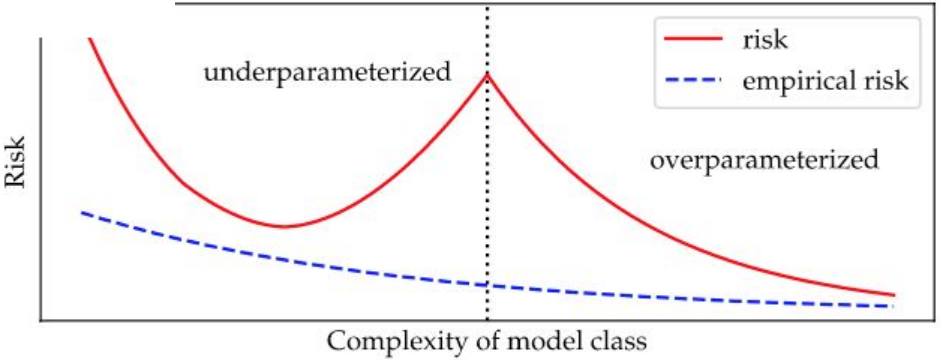
Loog, M., Viering, T.J., Mey, A., Krijthe, J.H., & Tax, D.M. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences, 117*, 10625 - 10626.

Dar, Y., Muthukumar, V., & Baraniuk, R. (2021). A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning. *ArXiv, abs/2109.02355*.

Single descent: larger models work better …

e.g. ResNet [He et al 2016] in computer vision





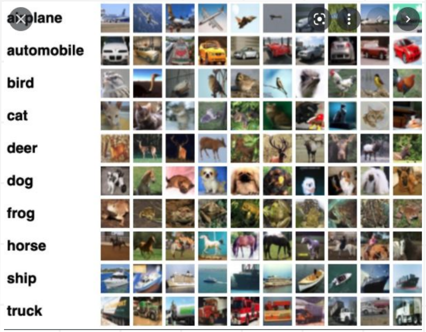
*Sometimes we see multiple bumps too …

Optimisation versus generalisation

$$R[f] = R_S[f] + \Delta_{\text{gen}}(f)$$

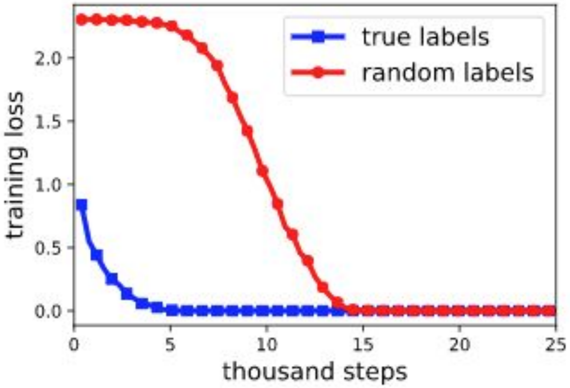
Training/optimisation is “easy” i.e. when # parameters > data points

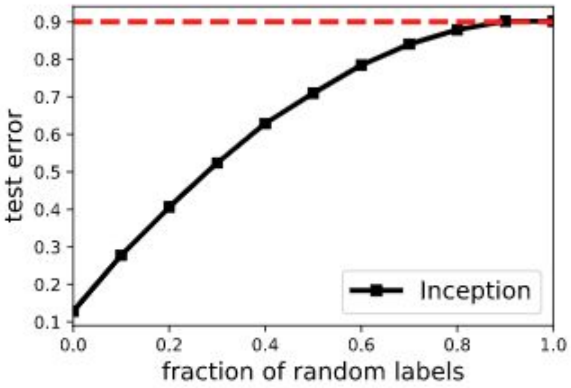
BUT, overparameterization puts burden on generalisation.



Experiment: Training with **random (!)** labels on CIFAR-10 (10 classes)
R[f] is known … test accuracy should be 1/10

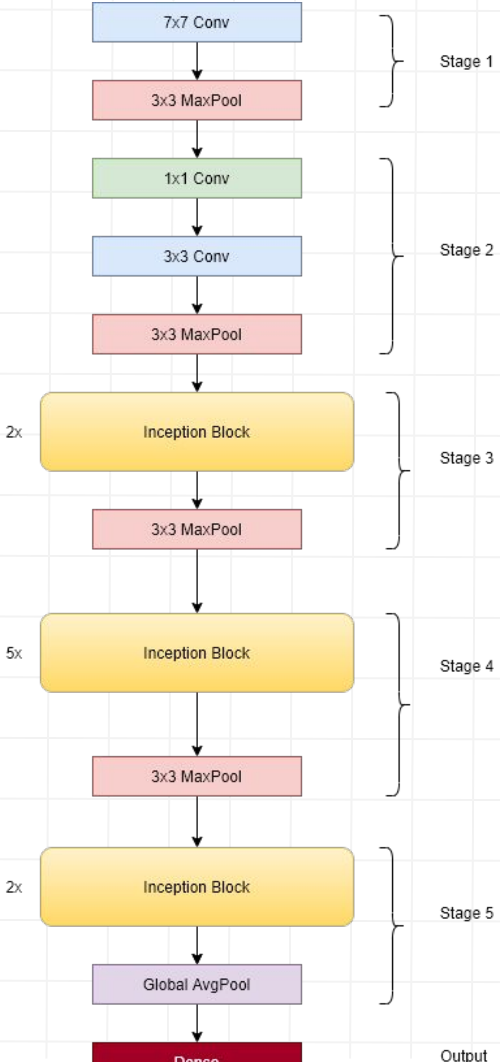
Training error is driven to zero by the optimisation algorithm → overfits
(similar observations hold for many overparameterized architectures in literature)

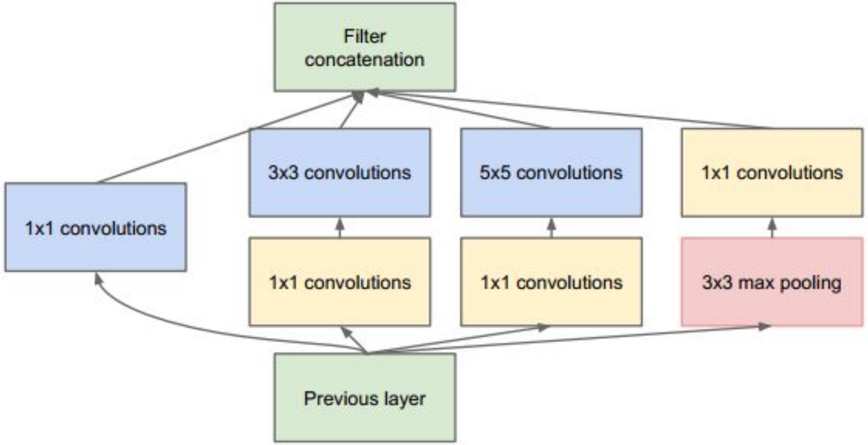




→ proof of convergence in optimisation may not reveal insights into the nature of generalisation.

Inception illustrated

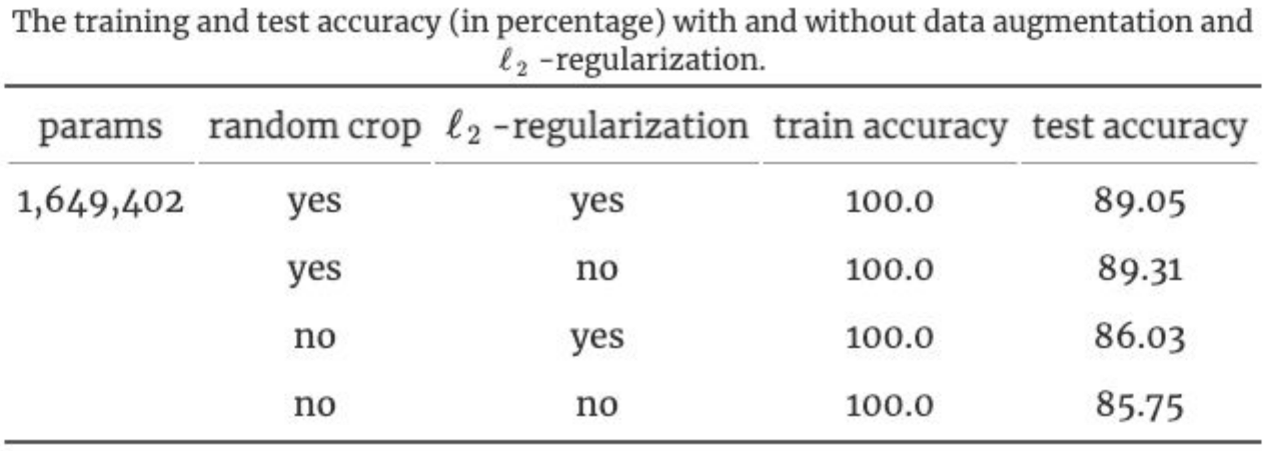




What about regularization?

- L2 regularisation (regression, large neural nets)
- Data augmentation (e.g. random cropping and rotation of training images)

Experiment on CIFAR-10 (50K training examples) + Inception (1.5M param)



The training and test accuracy (in percentage) with and without data augmentation and ℓ_2 -regularization.

params	random crop	ℓ_2 -regularization	train accuracy	test accuracy
1,649,402	yes	yes	100.0	89.05
	yes	no	100.0	89.31
	no	yes	100.0	86.03
	no	no	100.0	85.75

→ yes regularizations help, but is by no means necessary for strong generalisation.

Why learning works?

Four views presented in this chapter

- *Algorithmic stability*: generalization arises when ***models are insensitive to perturbations in the data*** on which they are trained.
- *VC dimension and Rademacher complexity*: how small generalization gaps can arise when we ***restrict the complexity of models*** we wish to fit to data.
- *Margin bounds*: whenever the ***data is easily separable***, good generalization will occur.
- *Optimization*: how ***choice of an algorithmic scheme*** itself can yield models with desired generalization properties

The four different views of generalization can all arrive at similar results – the UPPER bound on $\Delta_{gen}(f)$ depends on n (decreases as n increase) and the complexity of the ideal predictor (increases as complexity increase)

Generalization is multifaceted and multiple perspectives are useful when designing data-driven predictive systems.

How do we expect the gap to scale?

$$\boxed{R[f] = R_S[f] + \Delta_{\text{gen}}(f)}$$

For a fixed prediction function f, with infinite amount of data.
E_S[empirical risk] = population risk R[f].

Recall CLT (Central Limit Theorem)

If *Z* is a random variable with bounded variance then, then its sample mean converges in distribution to a Gaussian random variable with mean zero and variance on the order of 1/n.

Goal: Upper bound on Δ_{gen}(f), we want it to be small with high probability

$P [\mid R[f] - R_s[f] \mid] \geq \epsilon] \leq \delta$ OR $P [\mid R[f] - R_s[f] \mid] \leq \epsilon] \geq 1\text{-}\delta$

How fast does Δ_{gen}(f) shrink w.r.t. number of data points n?

E.g. k^n, n^k, O(n), log(n)

(a bite-sized intro to) Generalisation

The notion of generalization gap

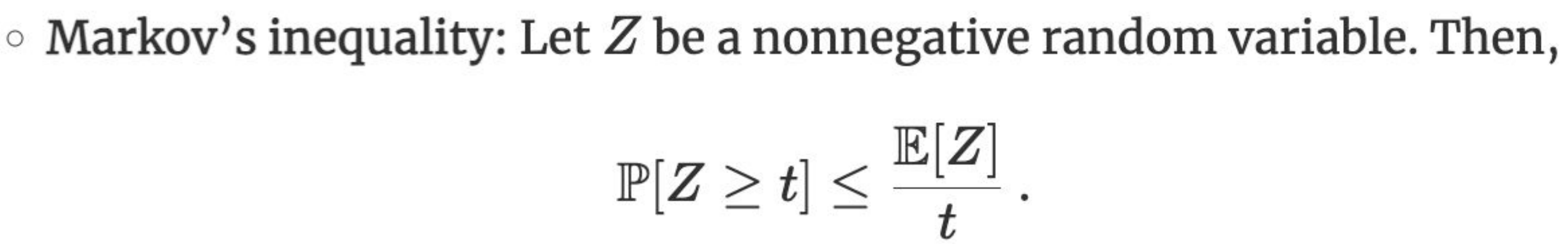
Overparameterization: empirical phenomena

Prelude: three inequalities – that we'll need later

Theories of generalization

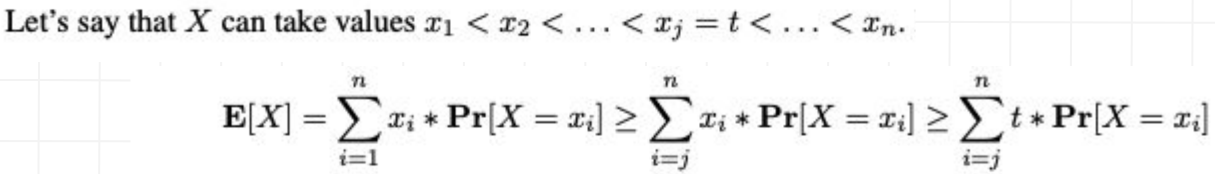
- Algorithmic stability
- Model complexity and uniform convergence
- Margin bounds
- Generalization from algorithms

Concentration inequalities



- Markov's inequality: Let Z be a nonnegative random variable. Then,

$$\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}\ .$$

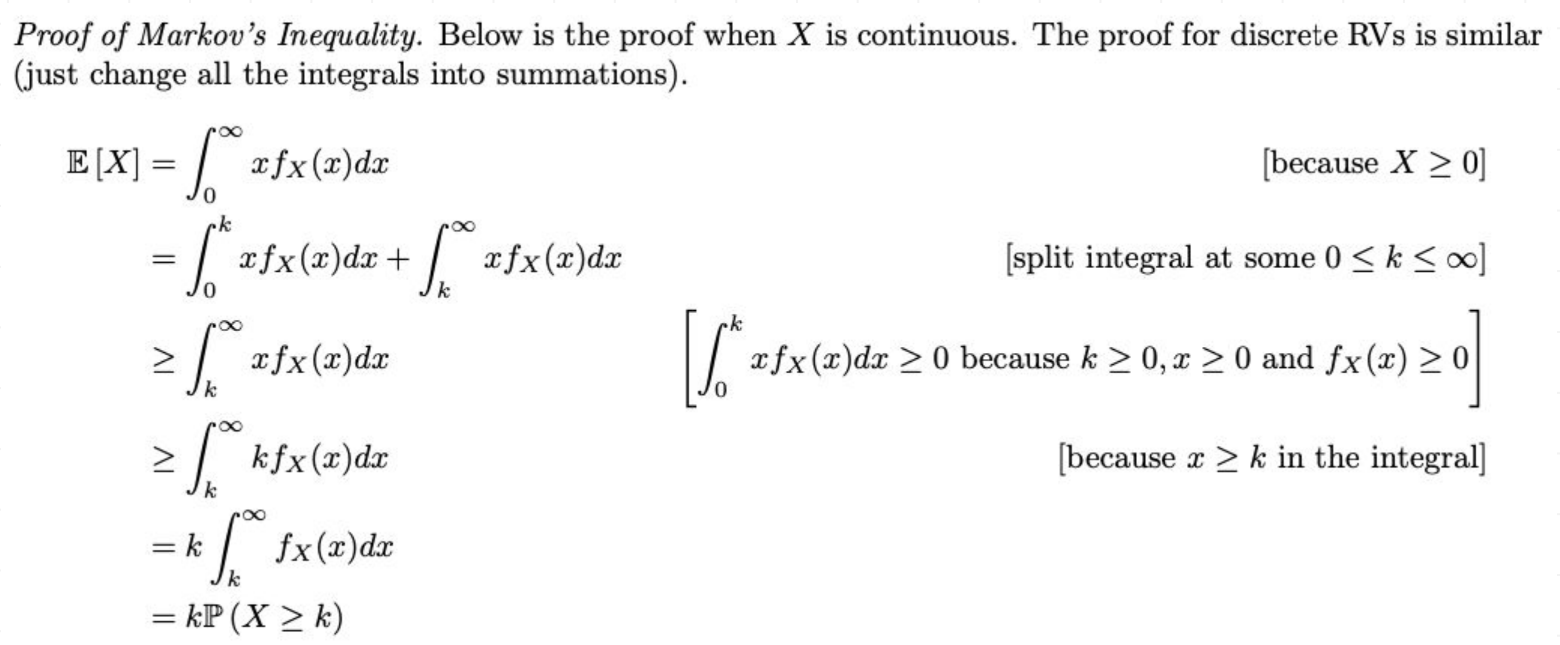


Let's say that X can take values $x_1 < x_2 < \ldots < x_j = t < \ldots < x_n$.

$$\mathbf{E}[X] = \sum_{i=1}^{n} x_i * \mathbf{Pr}[X = x_i] \geq \sum_{i=j}^{n} x_i * \mathbf{Pr}[X = x_i] \geq \sum_{i=j}^{n} t * \mathbf{Pr}[X = x_i]$$

[thanks: UW CS312]

Second form let $t = s\mathbf{E}[X]$, s>0 → $\mathbf{Pr}[X \geq s \cdot \mathbf{E}[X]\,] \leq \frac{1}{s}.$

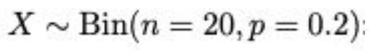


Proof of Markov's Inequality. Below is the proof when X is continuous. The proof for discrete RVs is similar (just change all the integrals into summations).

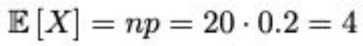
$$\begin{aligned}
\mathbb{E}\left[X\right] &= \int_0^\infty x f_X(x) dx && \text{[because } X \geq 0\text{]} \\
&= \int_0^k x f_X(x) dx + \int_k^\infty x f_X(x) dx && \text{[split integral at some } 0 \leq k \leq \infty\text{]} \\
&\geq \int_k^\infty x f_X(x) dx && \left[\int_0^k x f_X(x) dx \geq 0 \text{ because } k \geq 0, x \geq 0 \text{ and } f_X(x) \geq 0\right] \\
&\geq \int_k^\infty k f_X(x) dx && \text{[because } x \geq k \text{ in the integral]} \\
&= k \int_k^\infty f_X(x) dx \\
&= k \mathbb{P}\left(X \geq k\right)
\end{aligned}$$

Example: a weighted coin

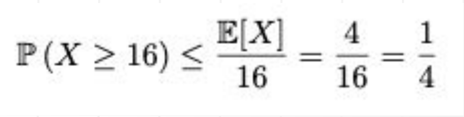
A coin is weighted so that its probability of landing on heads is 20%, independently of other flips.
Suppose the coin is flipped 20 times.
Use Markov’s inequality to bound the probability it lands on heads at least 16 times.
Also use Chebyshev’s inequality to upper bound the same probability.



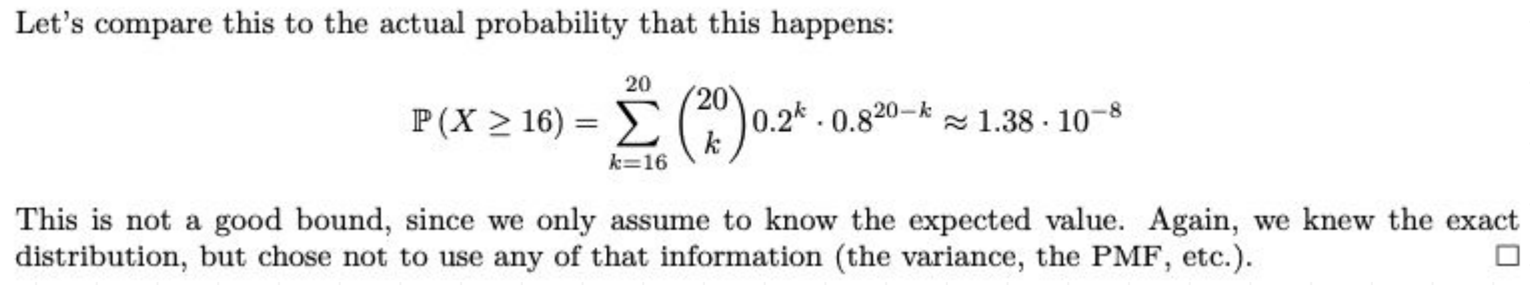
$X \sim \mathrm{Bin}(n = 20, p = 0.2)$:



$$\mathbb{E}[X] = np = 20 \cdot 0.2 = 4$$



$$\mathbb{P}(X \geq 16) \leq \frac{\mathbb{E}[X]}{16} = \frac{4}{16} = \frac{1}{4}$$



Let’s compare this to the actual probability that this happens:

$$\mathbb{P}(X \geq 16) = \sum_{k=16}^{20} \binom{20}{k} 0.2^k \cdot 0.8^{20-k} \approx 1.38 \cdot 10^{-8}$$

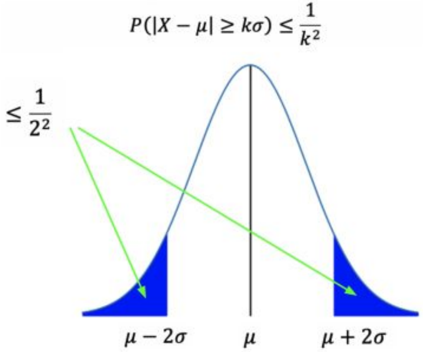
This is not a good bound, since we only assume to know the expected value. Again, we knew the exact distribution, but chose not to use any of that information (the variance, the PMF, etc.). □

[thanks: UW CS312]

- Chebyshev's inequality: Suppose Z is a random variable with mean μ_Z and variance σ_Z^2. Then,

$$\mathbb{P}[Z \geq t + \mu_Z] \leq \frac{\sigma_Z^2}{t^2}$$

Explains why sample averages are good estimates of the mean.



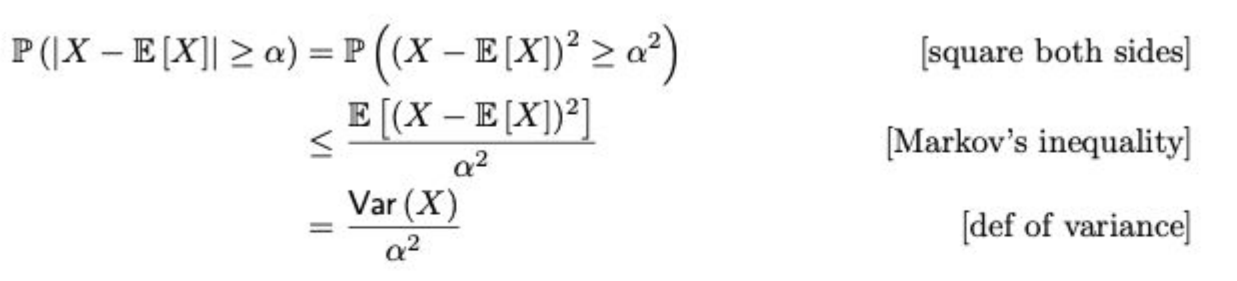
[thanks: UW CS312]

$$\mathbb{P}[\hat{\mu} \geq t + \mu_X] \leq \frac{\sigma_X^2}{nt^2}, \qquad \mathbb{P}[\hat{\mu} \geq 2\mu_X] \leq \frac{\sigma_X^2}{n\mu_X^2}.$$

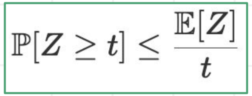
- Chebyshev's inequality: Suppose Z is a random variable with mean μ_Z and variance σ_Z^2. Then,

$$\mathbb{P}[Z \geq t + \mu_Z] \leq \frac{\sigma_Z^2}{t^2}$$

Proof sketch



$$\begin{aligned}\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \alpha\right) &= \mathbb{P}\left(\left(X - \mathbb{E}[X]\right)^2 \geq \alpha^2\right) && \text{[square both sides]}\\ &\leq \frac{\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]}{\alpha^2} && \text{[Markov's inequality]}\\ &= \frac{\mathsf{Var}(X)}{\alpha^2} && \text{[def of variance]}\end{aligned}$$



$$\boxed{\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}}$$

[thanks: UW CS312]

Example: a weighted coin (continued)

A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times.
Use Markov's inequality to bound the probability it lands on heads at least 16 times.
Also use Chebyshev's inequality to upper bound the same probability.

$X \sim \text{Bin}(n = 20, p = 0.2)$:

$$\mathbb{E}[X] = np = 20 \cdot 0.2 = 4$$

$$\mathsf{Var}(X) = np(1-p) = 20 \cdot 0.2 \cdot (1 - 0.2) = 3.2$$

Chebyshev's inequality is symmetric about the mean (difference of 12; 4 ± 12 gives the interval $[-8, 16]$):

$$\begin{aligned}
\mathbb{P}(X \geq 16) &\leq \mathbb{P}(X \geq 16 \cup X \leq -8) && \text{[adding another event can only increase probability]} \\
&= \mathbb{P}(|X - 4| \geq 12) && \text{[def of abs value]} \\
&= \mathbb{P}(|X - \mathbb{E}[X]| \geq 12) && [\mathbb{E}[X] = 4] \\
&\leq \frac{\mathsf{Var}(X)}{12^2} && \text{[Chebyshev's inequality]} \\
&= \frac{3.2}{12^2} = \frac{1}{45}
\end{aligned}$$

[thanks: UW CS312]

- Hoeffding's inequality: Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables, each taking values in the interval $[a_i, b_i]$.Let $\hat{\mu}$ denote the sample mean $\frac{1}{n}\sum_{i=1}^{n} Z_i$. Then

$$\mathbb{P}[\hat{\mu} \geq \mu_Z + t] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

An important special case is when the Z_i are identically distributed copies of Z and take values in $[0, 1]$. Then we have

ϵ δ

$$\mathbb{P}[\hat{\mu} \geq \mu_Z + t] \leq \exp\left(-2nt^2\right) .$$

when random variables are bounded, sample averages concentrate around their mean value exponentially quickly.

With probability at least 1-δ,

$$\hat{\mu} - \mu_Z \leq \epsilon$$

two-sided: $\mathbb{P}[|\hat{\mu} - \mu_z| \leq \epsilon] \geq 1 - 2\exp(-2n\epsilon^2)$ ← Use this!

one-sided: $\mathbb{P}[\hat{\mu} - \mu_z \leq \epsilon] \geq 1 - \exp(-2n\epsilon^2)$

Example application of a concentration inequality

A person’s height (0, 9] (unit: feet, 1 feet = 12 inches ~ 30 cm)

Sample 30,000 individuals (randomly!) $\{h_1, \ldots h_{30,000}\}$

Hoeffding’s inequality ➔

With 83% probability, sample mean μ'_h is within one inch of true mean μ_h

When random variables have:

- Low variance or are tightly bounded, small experiments quickly reveal insights about the population.
- Large variances or effectively unbounded, the number of samples required for high precision estimates might be impractical and *our estimators and algorithms and predictions* may need to be rethought.

Two scaling regimes

$R_S[f]$ large - generalisation gap decreasing at 1/sqrt(n)

$R_S[f]$ small- generalisation gap decreasing at 1/n

Why?

Consider a *single* prediction function *f*, chosen *independently* of the sample *S*

Note: this is not a random predictor

Hoeffding's inequality ➜

$$\mathbb{P}[R[f] - R_S[f] \geq \epsilon] \leq \exp\left(-2n\epsilon^2\right) .$$

With probability 1-δ,

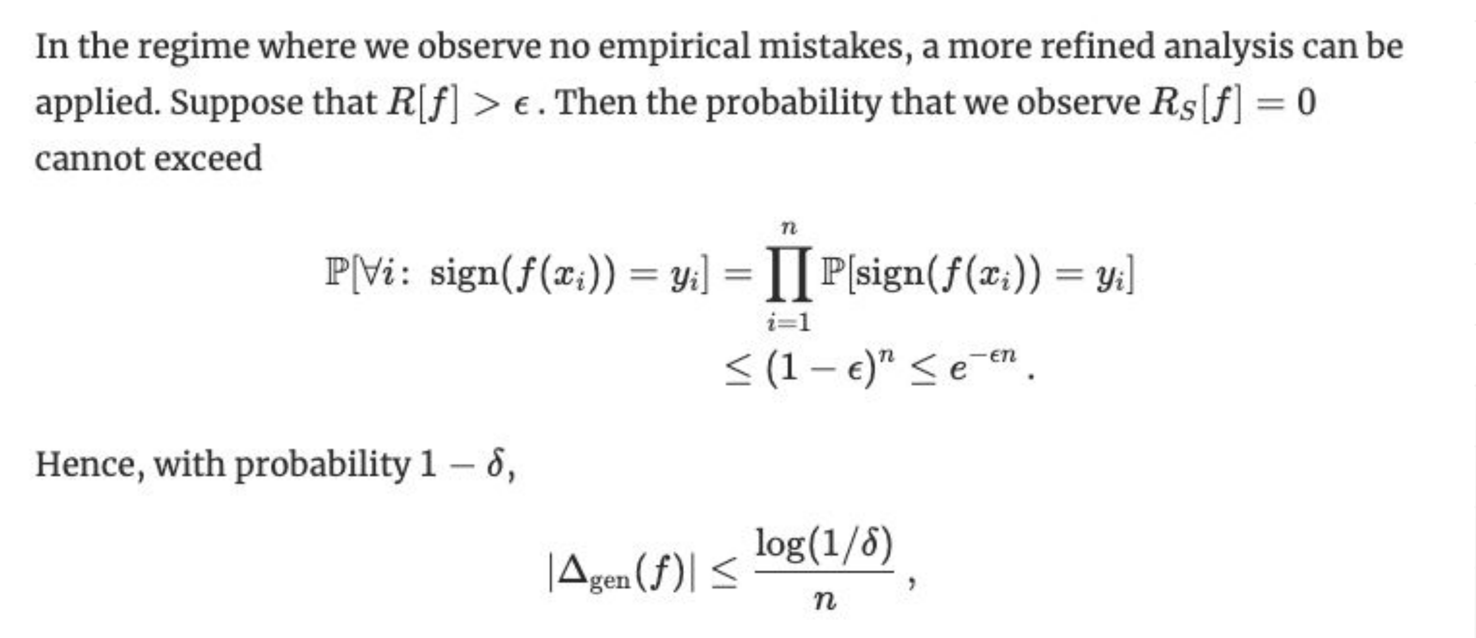
$$|\Delta_{\text{gen}}(f)| \leq \sqrt{\frac{\log(1/\delta)}{2n}} .$$

$$\mathbb{P}[|\hat{\mu} - \mu_z| \leq \epsilon] \geq 1 - 2\exp(-2n\epsilon^2) \xrightarrow[]{\text{Looser bound}} |\Delta_{gen}| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \xrightarrow[2\text{<=}1/\delta]{\text{Looser bound}} |\Delta_{gen}| \leq \sqrt{\frac{\log(1/\delta)}{n}}$$

Two scaling regimes

$R_S[f]$ large - generalisation gap decreasing at 1/sqrt(n)

$R_S[f]$ small- generalisation gap decreasing at 1/n



In the regime where we observe no empirical mistakes, a more refined analysis can be applied. Suppose that $R[f] > \epsilon$. Then the probability that we observe $R_S[f] = 0$ cannot exceed

$$\begin{aligned} \mathbb{P}[\forall i:\ \operatorname{sign}(f(x_i)) = y_i] &= \prod_{i=1}^{n} \mathbb{P}[\operatorname{sign}(f(x_i)) = y_i] \\ &\le (1-\epsilon)^n \le e^{-\epsilon n} . \end{aligned}$$

Hence, with probability $1-\delta$,

$$|\Delta_{\text{gen}}(f)| \le \frac{\log(1/\delta)}{n} ,$$

(a bite-sized intro to) Generalisation

The notion of generalization gap

Overparameterization: empirical phenomena

Prelude: three inequalities – that we'll need later

Theories of generalization

- Algorithmic stability
- Model complexity and uniform convergence
- Margin bounds
- Generalization from algorithms

Algorithmic stability

generalization arises when ***models are insensitive to perturbations in the data*** on which they are trained.

Specifically: how sensitive an algorithm is to changes in a **single** training example.

Three ways of data perturbation, all yielding similar generalisation bounds.

- **Resample a single data point**
- Leave one data point out
- A single data point is arbitrarily corrupted (adversarial scenario)

Some notations

A sample of n data points $S = ((x_1, y_1), \ldots\ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n .$

A labeled example $z = (x, y).$ $loss(f, z) = loss(f(x), y)$

Z ~ distribution of (X, Y)

n hybrid samples $S^{(i)} = (Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n)$

$$S = (Z_1, \ldots, Z_n)$$
$$S' = (Z'_1, \ldots, Z'_n)$$

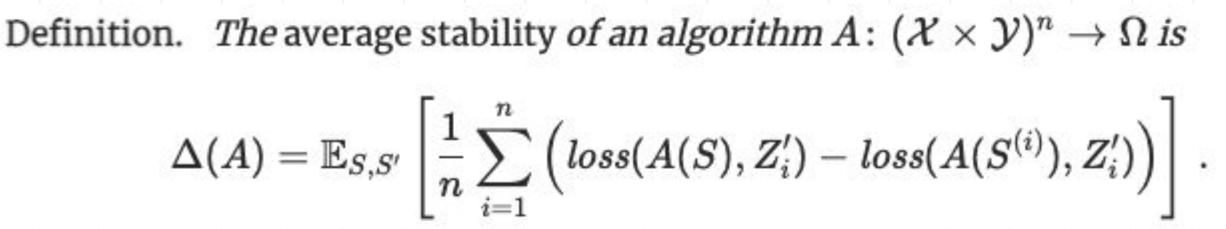
A learning algorithm $A\colon (\mathcal{X} \times \mathcal{Y})^n \to \Omega$

Space of data samples

space of function f

A is assumed to minimize R_s – optimisation bounds will be mentioned later

Average stability and its link to Δ_{gen}



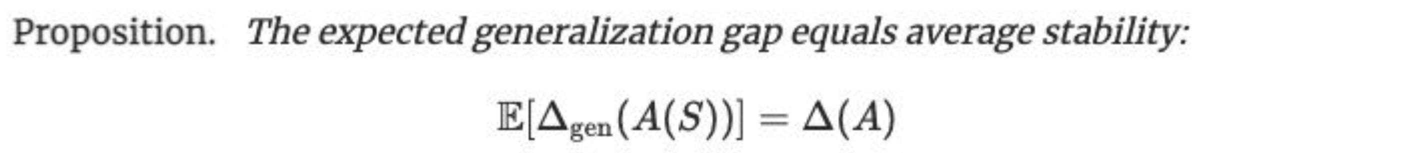
Definition. *The* average stability *of an algorithm* $A\colon (\mathcal{X}\times\mathcal{Y})^n \to \Omega$ *is*

$$\Delta(A) = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathit{loss}(A(S),Z_i') - \mathit{loss}(A(S^{(i)}),Z_i')\right)\right].$$

Two interpretations

change w.r.t. one example;

change w.r.t seen vs unseen examples



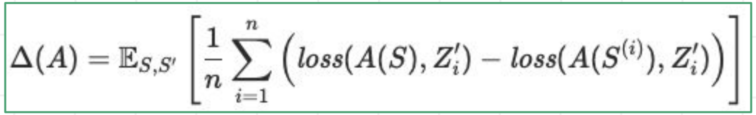
Proposition. *The expected generalization gap equals average stability:*

$$\mathbb{E}[\Delta_{\text{gen}}(A(S))] = \Delta(A)$$

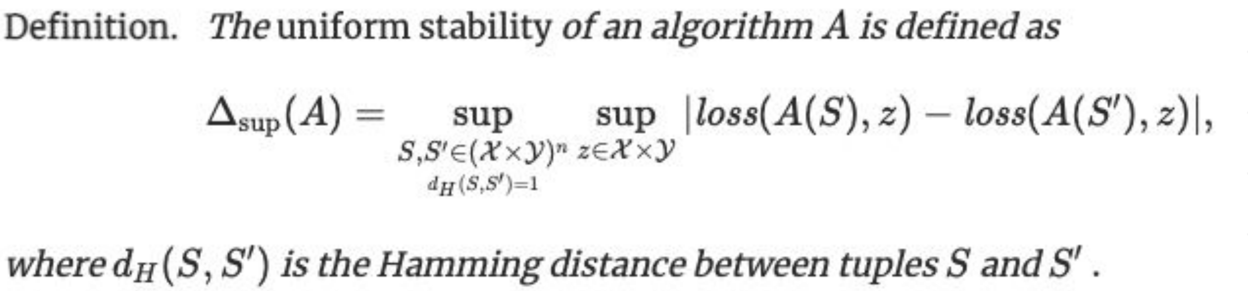
[proof omitted, see book]

Uniform stability

Average stability



$$\Delta(A) = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n}\left(loss(A(S), Z_i') - loss(A(S^{(i)}), Z_i')\right)\right]$$



Definition. *The* uniform stability *of an algorithm A is defined as*

$$\Delta_{\sup}(A) = \sup_{\substack{S,S' \in (\mathcal{X}\times\mathcal{Y})^n \\ d_H(S,S')=1}} \sup_{z\in\mathcal{X}\times\mathcal{Y}} |loss(A(S), z) - loss(A(S'), z)|,$$

where $d_H(S, S')$ *is the Hamming distance between tuples S and S′ .*

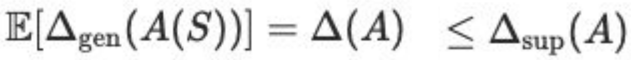
The two *sups* here – worst case scenario

Why is this called uniform? – will hold for every A, (X, Y)

z has nothing to do with S or S', but is sampled from the same distribution (X, Y)

The **worst-case difference** in the predictions of the learning algorithm run on two arbitrary datasets that differ in exactly one point.

Uniform stability upper bounds generalization gap (in expectation)
See book for proof.



$$\mathbb{E}[\Delta_{\text{gen}}(A(S))] = \Delta(A) \quad \leq \Delta_{\sup}(A)$$

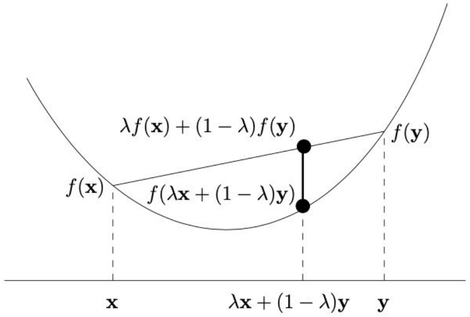
Strongly convex functions

[thanks: EPFL OptML, eq numbers therein]

Convex functions: lower-bounded by tangent lines.

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$


$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$
$f(\mathbf{y})$
$f(\mathbf{x})$
$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y})$
$\mathbf{x}$
$\lambda \mathbf{x} + (1-\lambda)\mathbf{y}$
$\mathbf{y}$

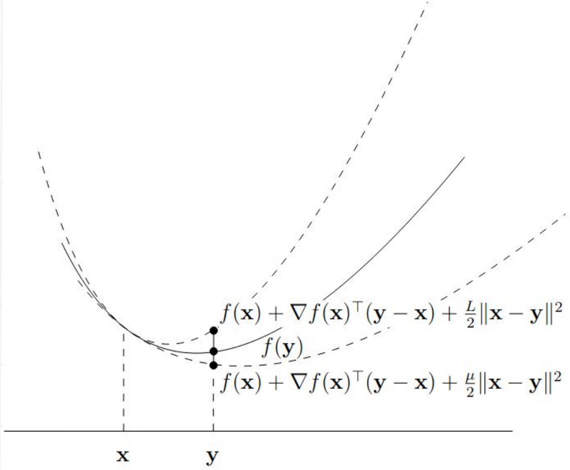

$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$
$f(\mathbf{y})$
$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$
$\mathbf{x}$
$\mathbf{y}$

Figure 2.3: A smooth and strongly convex function

Strictly convex functions $$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \qquad (1.8)$$

Strongly convex functions, $\mu > 0$ $$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \qquad (2.19)$$

Nice properties:

- Lower-bounded by another quadratic function.
- There is a unique global minimum
- It’s “fast” to find

Strongly convex and L-Lipschitz loss functions

Goal: show that strong convexity of the loss function is sufficient for the uniform stability of empirical risk minimization.

Two assumptions needed:

Loss function differentiable and strongly convex

$$loss(w', z) \geq loss(w, z) + \langle \nabla loss(w, z), w' - w \rangle + \frac{\mu}{2}\|w - w'\|^2$$

If $\mathbf{\Phi}$:R^d→R is μ-strongly convex and w* is a stationary point (and hence global minimum)

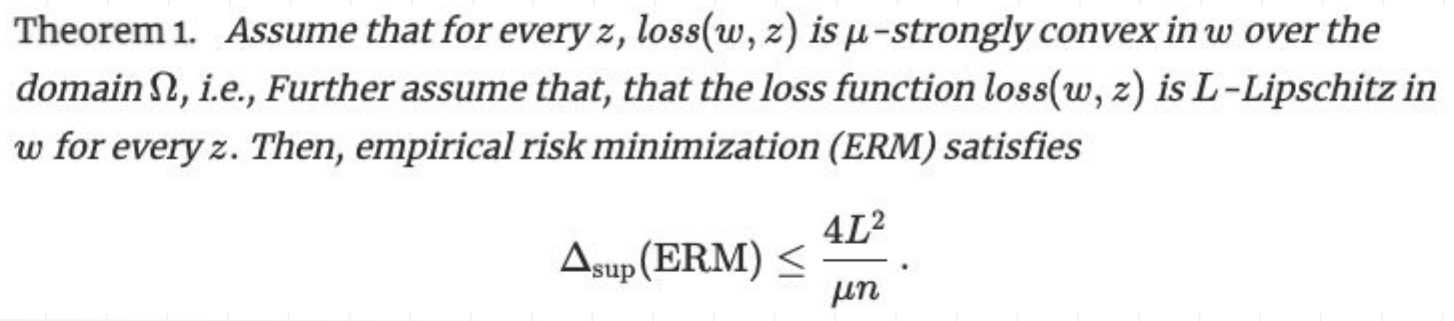
$$\Phi(w) - \Phi(w^*) \geq \frac{\mu}{2}\|w - w^*\|^2$$

$loss(w, z)$ is L-*Lipschitz* in w for every z

$$\|\nabla loss(w, z)\| \leq L \qquad |loss(w, z) - loss(w', z)| \leq L\|w - w'\|.$$

Stability of empirical risk minimisation

See proof in MLstory



Theorem 1. *Assume that for every z, loss(w,z) is μ-strongly convex in w over the domain Ω, i.e., Further assume that, that the loss function loss(w,z) is L-Lipschitz in w for every z. Then, empirical risk minimization (ERM) satisfies*

$$\Delta_{\text{sup}}(\text{ERM}) \le \frac{4L^2}{\mu n}\,.$$

There is no explicit reference to model class. But what is implied here?

https://math.stackexchange.com/questions/1106154/any-example-of-strongly-convex-functions-whose-gradients-are-lipschitz-continuou

What about regularisation?

$$r(w,z) = loss(w,z) + \frac{\mu}{2}\|w\|^2$$

L_2 regularisation turns convex loss into a μ-strongly convex one

assume $\|w\| \le B$ set $\mu = \frac{L}{B\sqrt{n}}$ $\frac{\mu}{2}\|w\|^2$ at most $O(\frac{LB}{\sqrt{n}})$

$\boxed{\Delta_{\text{sup}}(\text{ERM}) \le \frac{4L^2}{\mu n}} \longrightarrow$ the generalization gap will also be $O(\frac{LB}{\sqrt{n}})$

Model complexity and uniform convergence

Uniform convergence: bounding the generalization gap from above for all functions in a function class

Model complexity: counting the number of different functions that can be described with the given model parameters.

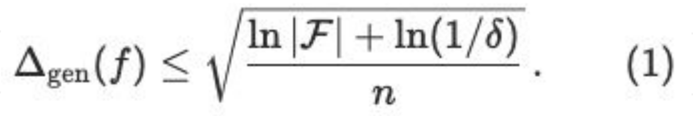
Assume loss function bounded in [0, 1], apply Hoeffding's

$$\mathbb{P}\left[R_S[f] > R[f] + t\right] \leq \exp(-2nt^2)$$

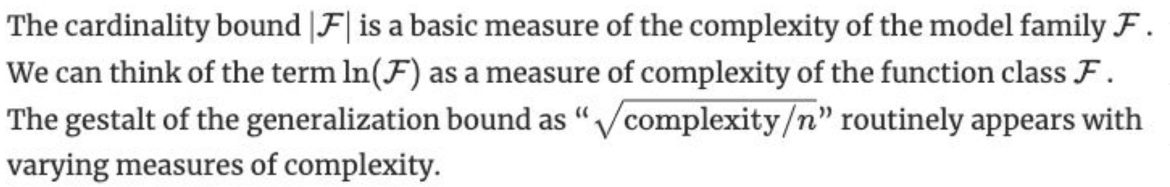
For data-independent prediction function f

With probability 1-δ, $|\Delta_{\text{gen}}(f)| \leq \sqrt{\frac{\log(1/\delta)}{\cancel{2}n}}$.

With probability $1-\delta,\ \forall f \in \mathcal{F}$



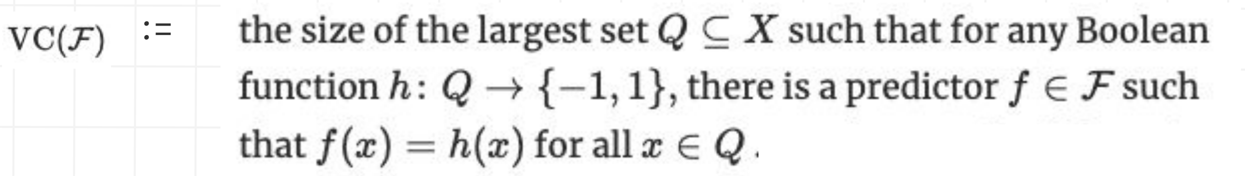
$$\Delta_{\text{gen}}(f) \leq \sqrt{\frac{\ln|\mathcal{F}| + \ln(1/\delta)}{n}}. \qquad (1)$$



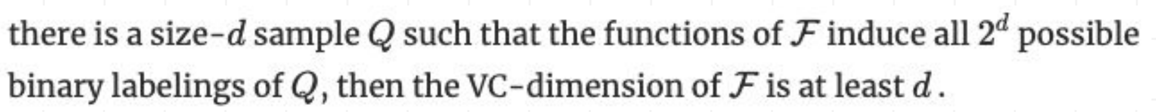
The cardinality bound $|\mathcal{F}|$ is a basic measure of the complexity of the model family $\mathcal{F}$. We can think of the term $\ln(\mathcal{F})$ as a measure of complexity of the function class $\mathcal{F}$. The gestalt of the generalization bound as "$\sqrt{\text{complexity}/n}$" routinely appears with varying measures of complexity.

VC Dimension (Vapnik-Chervonenkis)

Uniform convergence:= bounding the generalization gap from above for all functions in a function class. What happens if |F| infinite?



$\mathrm{VC}(\mathcal{F})$:= the size of the largest set $Q \subseteq X$ such that for any Boolean function $h\colon Q \to \{-1, 1\}$, there is a predictor $f \in \mathcal{F}$ such that $f(x) = h(x)$ for all $x \in Q$.



there is a size-d sample Q such that the functions of $\mathcal{F}$ induce all 2^d possible binary labelings of Q, then the VC-dimension of $\mathcal{F}$ is at least d.

The VC-dimension measures the ability of the model class to conform to an arbitrary labeling of a set of points.

Example: linear models over R^{d} has a VC dimension of d – same as number of model parameters.

VC inequalities

$$\Delta_{\text{gen}}(f) \leq \sqrt{\frac{\mathrm{VC}(\mathcal{F}) \ln n + \ln(1/\delta)}{n}}. \qquad (2)$$

Slower rate here!

Consider all hyperplanes in R^d with norm at most γ^{-1}, data has bounded norm ||x|| <= D
The VC dimension of these hyperplanes is D$^2/\gamma^2$

$$\Delta_{\text{gen}}(f) \leq \sqrt{\frac{D^2 \ln n + \gamma^2 \ln(1/\delta)}{\gamma^2 n}}.$$

d does not appear, ‘free’ from curse of dimensionality

Summary of risk and bounds

$$\mathbb{E}[\Delta_{\text{gen}}(A(S))] = \Delta(A) \quad \leq \Delta_{\text{sup}}(A)$$

Algorithmic stability: generalization arises when ***models are insensitive to perturbations in the data*** on which they are trained.

$$\Delta_{\text{sup}}(\text{ERM}) \leq \frac{4L^2}{\mu n} \,.$$

VC dimension and Rademacher complexity: how small generalization gaps can arise when we ***restrict the complexity of models*** we wish to fit to data.

$$\Delta_{\text{gen}}(f) \leq \sqrt{\frac{\ln|\mathcal{F}| + \ln(1/\delta)}{n}} \,.$$

Margin bounds: whenever the ***data is easily separable***, good generalization will occur.

$$R[f] - R_S^\theta[f] \leq 4\frac{\mathfrak{R}(\mathcal{W}_B)}{\theta} + O\left(\frac{\log(1/\delta)}{\sqrt{n}}\right)$$

Optimization: how ***choice of an algorithmic scheme*** itself can yield models with desired generalization properties

$$\Delta_{\text{sup}}(\text{SGM}) \leq \frac{2L^2}{n} \sum_{t=1}^{T} \eta_t \,.$$

Summary - (a bite-sized intro to) Generalisation

The notion of generalization gap

Overparameterization: empirical phenomena

Prelude: three inequalities – that we’ll need

Theories of generalization

- Algorithmic stability
- Model complexity and uniform convergence
- Margin bounds
- Generalization from algorithms