Inductive Principles

Robert C. Williamson*
Australian National University
Canberra, 0200 ACT, Australia

(Bowdlerised Edition)

* Joint work with **Ralf Herbrich**, Microsoft Research Cambridge

Outline

- Induction (what's "inductive"?)
- Inductive Principles (what's the "principle"?)
- Empirical Risk Minimization
- Key Theorem of Learning Theory
- Conditioning on the Data

After the break, we will move on to the more technical part of the talk . . .

- "Conditioning on the data" in a Frequentist (PAC) setting — The Luckiness Framework
- A new approach — Algorithmic Luckiness

Induction

Hume's problem is how to justify Induction: the inference (discovery of laws) from empirical data.

> If we take in our hand any volume - of divinity or school metaphysics, for instance- let us ask. Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning containing matter of fact and existence? No. Commit it then to the flames, for it can contain nothing but sophistry and illusion.
>
> — David Hume

Impossible "for all is but a woven web of guesses".

Popper's key insight: scientific theories do not lead to certain knowledge; merely approximations to the truth. Thus no "justification"

We can, however, reason logically about the process of scientific discovery. Doing so shows one should prefer a more refutable theory over a less refutable one.

Induction

> [W]e can always construct our machine so that it starts issuing probabilistic predictions only after the 1000th event, say, or after any other number n which we may choose, bearing in mind the number of different properties in our ‘world’. (The problem is so trivial that it is not worth making any effort to solve it systematically; for we know, after all, that applications of the simple inductive rule will never give us more that increasingly good aproximations).
>
> — Karl Popper *Realism and the Aim of Science*, p321 (1956,1983)

Based on work by Menger (1924) Popper argued could formally consider the “dimension” of a theory. Although it is impossible to build a general learning machine Popper clearly admitted the possibility *within a constrained framework* of building a machine and being able to probabilistically reason about its performance.

We will study learning machines, and not induction in general.

Learning Problem

The players . . . threw these abstract formulas at one another displaying the sequences and possibilities of their science.

— Herman Hesse: The Glass Bead Game

Given:

- A training sample $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) = (z_1, \ldots, z_m) \in (\mathcal{X} \times \mathcal{Y})^m = \mathcal{Z}^m$ drawn iid from P_Z (unknown).
- A deterministic learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$.
- A loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$.

Question: How can one tell whether $\mathcal{A}$ is good or not?

One Possible Solution

Find a (probabilistic) bound on $R_l[\mathcal{A}(z)] := \mathbf{E}_{XY}[l(\mathcal{A}(z)(X), Y)]$, that is, a function ψ such that

$$\mathbf{P}_{Z^m}(R_l[\mathcal{A}(\mathbf{Z})] \leq \psi(\mathcal{A}, \mathbf{Z}, \delta)) \geq 1 - \delta\,.$$

Why do we want such bounds?

A bound such as

$$\mathbf{P}_{Z^m}\left(R_l\left[\mathcal{A}\left(\mathbf{Z}\right)\right] \leq \psi\left(\mathcal{A}, \mathbf{Z}, \delta\right)\right) \geq 1-\delta$$

is not an end in itself; it suggests how to adjust the parameters (or knobs) of the learning algorithm $\mathcal{A}$.

Thus the *closer* the analysis is to the algorithm the more insightful we would hope it to be.

Difficulty: Given $\boldsymbol{z} \in \mathcal{Z}^m$ how can the algorithm $\mathcal{A}$ choose an hypothesis that achieves a small value of $R_l[\mathcal{A}(\boldsymbol{z})]$?

Key point: Given $\boldsymbol{z} \in \mathcal{Z}^m$. No chance of computing $R_l[\mathcal{A}(\boldsymbol{z})]$ even in principle because we do not know $\mathbf{P}_Z$.

A Recipe for Generalisation Error Bounds

1. Relate the prediction error $R_l\left[h\right]$ to some empirical quantity, e.g. training error

$$\widehat{R}_l\left[h, \boldsymbol{z}\right] := \frac{1}{|\boldsymbol{z}|} \sum_{(x,y)\in\boldsymbol{z}} l\left(h\left(x\right), y\right),$$

that converges exponentially to $R_l\left[h\right]$ for any h.

2. Apply the basic lemma to the difference of the prediction error and the empirical quantity (training error). This introduces a ghost sample.

3. Fully exploit the independence assumption of $\boldsymbol{z}$ by using a technique known as symmetrisation by permutation: (probability is over double sample Z^{2m})

$$\mathbf{\mathsf{P}}_{\mathsf{Z}^{2m}}\left(\Upsilon\left(\mathbf{\mathsf{Z}}\right)\right) = \mathbf{\mathsf{E}}_{\mathsf{I}}\left[\mathbf{\mathsf{P}}_{\mathsf{Z}^{2m}|\mathsf{I}=\mathsf{i}}\left(\Pi_{\mathsf{i}}\left(\Upsilon\left(\mathbf{\mathsf{Z}}\right)\right)\right)\right] = \mathbf{\mathsf{E}}_{\mathsf{Z}^{2m}}\left[\mathbf{\mathsf{P}}_{\mathsf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}\left(\Pi_{\mathsf{I}}\left(\Upsilon\left(\boldsymbol{z}\right)\right)\right)\right].$$

4. Since $\boldsymbol{z} \in \mathcal{Z}^{2m}$ is fixed, we can construct a cover w.r.t. the loss l and apply the union bound.

Inductive Principle — Empirical Risk Minimization

This is a great algorithm to analyse.
— Ralf Herbrich

Possibility of computing such bounds motivates:
Empirical Risk Minimization Algorithm

$$\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}} \; : \; \mathcal{Z}^{(\infty)} \to \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$$

$$\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}} \; : \; \boldsymbol{z} \mapsto \operatorname*{arg\,min}_{h \in \mathcal{H}} \widehat{R}_l\left[h, \boldsymbol{z}\right].$$

The “principle” is to minimize the empirical surrogate $\widehat{R}_l[h, \boldsymbol{z}]$ of $R_l[h, \boldsymbol{z}]$.

Thus depending on the choice of sets $\mathcal{H}$, get a family of *empirical risk minimization algorithms*.

Note that the algorithm $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$ has one “knob”: the class of functions $\mathcal{H}$.

How to choose $\mathcal{H}$? Want $\mathcal{H}$ as large as possible to ensure a good approximation of the underlying data generating process. Pay a price . . .

Consistency and Strict Consistency

Question: Does $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$ "work"?
More precise question: Is $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$ consistent?

Assume that $\mathcal{H}$ and l are such that for any $\boldsymbol{z} \in \mathcal{Z}^m$

$$\widehat{R}_l \left[\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}(\boldsymbol{z}), \boldsymbol{z}\right] = \inf_{h\in\mathcal{H}} \widehat{R}_l\, [h(\boldsymbol{z}), \boldsymbol{z}]$$

and that for all $h \in \mathcal{H}$, $A \leq R_l[h] \leq B$. Let

$$\mathcal{H}(c) := \{h \in \mathcal{H} \colon R_l[h] \geq c\}.$$

Say that $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$ is *strictly (nontrivially) consistent* if for all $c \geq 0$, for all $\varepsilon > 0$,

$$\lim_{m\to\infty} \mathbf{P}_{\mathsf{Z}^m} \left(\left| \widehat{R}_l \left[\mathcal{A}^{\mathcal{H}(c)}_{\mathrm{emp}}(\boldsymbol{z}), \boldsymbol{z}\right] - c \right| > \varepsilon \right) = 0.$$

Need a definition like this to rule out "coding" the identity of a function into one observation: can construct such artificial function classes of arbitrary complexity which can be learned using $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$ with only one observation.

Consistency of ERM

$\mathcal{A}_{\mathrm{erm}}^{\mathcal{H}}$ is strictly consistent

$\Leftrightarrow$

$$\forall \varepsilon > 0 \quad \lim_{m\to\infty} \mathbf{P}_{\mathsf{Z}^m} \left\{ \sup_{h\in\mathcal{H}} \left(R[h] - \widehat{R}_l[h,\boldsymbol{z}]\right) > \varepsilon \right\} = 0 \qquad \star$$

$$\forall \varepsilon > 0 \quad \lim_{m\to\infty} \mathbf{P}_{\mathsf{Z}^m} \left\{ \sup_{h\in\mathcal{H}} \left|R_l[h] - \widehat{R}_l[h,\boldsymbol{z}]\right| > \varepsilon \right\} = 0 \qquad \star\star$$

$\Leftrightarrow$

$$\forall \varepsilon > 0 \quad \lim_{m\to\infty} \frac{1}{m} \mathbf{E}_{\mathsf{Z}^m} \log \underbrace{\mathcal{N}\left(\varepsilon, \mathcal{H}, \ell_1(\boldsymbol{z})\right)} = 0$$

Covering number of $\mathcal{H}$ at scale ε w.r.t. to the $\ell_1(\boldsymbol{z})$ metric: for $h \in \mathcal{H}$, $\|h\|_{\ell_1(\boldsymbol{z})} := \frac{1}{m}\sum_{z\in\boldsymbol{z}} |h(z)|$.

The effective gap in the reasoning implicit in the difference between $\star$ and $\star\star$ can be plugged using a more complex notion of cover — a one sided bracket cover. I am unaware of any results on the relative sizes of such covering numbers compared to $\mathcal{N}\left(\varepsilon, \mathcal{H}, \ell_1(\boldsymbol{z})\right)$.

So What?

The big deal is that (modulo the small gap mentioned)

$$\mathbf{E}_{Z^m} \log \mathcal{N}(\varepsilon, \mathcal{H}, \ell_1(\boldsymbol{z}))$$

is the "right" quantity to study for understanding the effect of the $\mathcal{H}$ knob on $\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}$. (Why it is worth fussing with strict consistency.)

Thus we know how to understand the effect of the "knob" $\mathcal{H}$.

Note it is impossible to compute (even in principle) since we do not know $\mathbf{P}_{Z^m}$ (the distribution from which $\boldsymbol{z}$ is drawn).

Can upper bound by $\sup_{\boldsymbol{z} \in \mathcal{Z}^m} \log \mathcal{N}(\varepsilon, \mathcal{H}, \ell_1(\boldsymbol{z}))$ which can be effectively bounded.

Leads to "generalization bounds" of the form: for $\boldsymbol{z} \in \mathcal{Z}^m$

$$\mathbf{P}_{Z^m}\left(R_l\left[\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}(\mathbf{Z})\right] \leq \psi_l(\mathcal{H}, \mathbf{Z}, \delta)\right) \geq 1 - \delta\,.$$

The “Key Theorem” in Learning Theory

“ERM is strictly consistent iff covering numbers behave nicely”

Observe that whilst we set out to understand the behaviour of $\mathcal{A}^{\mathcal{H}}_{\text{erm}}$ our bounds are in fact for

$$\mathcal{A}^{\mathcal{H}}_{\text{worst}} := \boldsymbol{z} \mapsto \arg \max_{h \in S(\mathcal{H}, \boldsymbol{z})} R_l[h]$$

where

$$S(\mathcal{H}, \boldsymbol{z}) = \left\{ h \in \mathcal{H} \colon \widehat{R}_l[h, \boldsymbol{z}] = \widehat{R}_l[\mathcal{A}^{\mathcal{H}}_{\text{erm}}, \boldsymbol{z}] \right\}.$$

Consequently the bounds are very loose.

Furthermore $\mathcal{A}^{\mathcal{H}}_{\text{erm}}$ could perform as poorly as $\mathcal{A}^{\mathcal{H}}_{\text{worst}}$ (what is there to stop it?).

Conclusion: behaviour of covering numbers is the crucial quantity for this inductive principle (algorithm). Suggests to make $\mathcal{H}$ as small as possible.

Another Algorithm: SRM

An obvious difficulty with $\mathcal{A}_{\mathrm{erm}}^{\mathcal{H}}$ is that if one chooses $\mathcal{H}$ badly, the algorithm has no hope of approximating the data.

Suppose for $\boldsymbol{z} \in \mathcal{Z}^m$, we know

$$\mathbf{P}_{\mathsf{Z}^m}\left(R_l\left[\mathcal{A}_{\mathrm{erm}}^{\mathcal{H}}(\mathbf{Z})\right] \leq \psi_l(\mathcal{H}, \mathbf{Z}, \delta)\right) \geq 1 - \delta\,.$$

Given a sequence of nonnegative numbers $\boldsymbol{\delta} = (\delta_i)_{i\in\mathbb{N}}$ such that $\sum_i \delta_i = \delta$ and a sequence of hypothesis classes $\boldsymbol{\mathcal{H}} = (\mathcal{H}_i)_{i\in\mathbb{N}}$

$$i^* = i^*(\boldsymbol{z}, \boldsymbol{\mathcal{H}}, \boldsymbol{\delta}, \psi) \;:=\; \operatorname*{arg\,min}_{i\in\mathbb{N}} \psi_l(\mathcal{H}_i, \boldsymbol{z}, \delta_i)$$

$$\mathcal{A}_{\mathrm{srm}}^{\boldsymbol{\mathcal{H}},\boldsymbol{\delta}}(\boldsymbol{z}) \;:=\; \mathcal{A}_{\mathrm{erm}}^{\mathcal{H}_{i^*}}(\boldsymbol{z}).$$

Bounds for $\mathcal{A}_{\mathrm{srm}}^{\boldsymbol{\mathcal{H}},\boldsymbol{\delta}}$

By the definition of $\mathcal{A}_{\mathrm{srm}}^{\boldsymbol{\mathcal{H}},\boldsymbol{\delta}}$ it comes with a performance bound already. For $i \in \mathbb{N}$, with probability at least $1-\delta_i$ over a random draw of $\boldsymbol{z}$,

$$R\left[\mathcal{A}_{\mathrm{erm}}^{\mathcal{H}_i}(\boldsymbol{z})\right] \leq \psi_l(\mathcal{H}_i, \boldsymbol{z}, \delta_i)$$

Thus the union bound ensures that with probability at least $1-\delta$ over a random draw of $\boldsymbol{z}$, for all $i \in \mathbb{N}$

$$R\left[\mathcal{A}_{\mathrm{erm}}^{\mathcal{H}_i}(\boldsymbol{z})\right] \leq \psi_l(\mathcal{H}_i, \boldsymbol{z}, \delta_i)$$

and thus with probability at least $1-\delta$ over a random draw of $\boldsymbol{z}$,

$$R\left[\mathcal{A}_{\mathrm{srm}}^{\boldsymbol{\mathcal{H}},\boldsymbol{\delta}}(\boldsymbol{z})\right] \leq \psi_l(\mathcal{H}_{i^*}, \boldsymbol{z}, \delta_{i^*}) \qquad \spadesuit$$

Algorithm Independence of Bound

"You must *condition on the data*!"
— (Spirit of) Thomas Bayes

Classical bound takes form: with probability at least $1-\delta$ over a random draw of $\boldsymbol{z} \in \mathcal{Z}^m$ according to $\mathbf{P}_{\mathsf{Z}^m}$,

$$R_l[\mathcal{A}^{\mathcal{H}}_{\mathrm{erm}}(\boldsymbol{z})] \leq \psi(\mathcal{H}, \boldsymbol{z}, \delta).$$

Thus *any* algorithm

$$\mathcal{A}^{\mathcal{H}}_{\mathrm{any}} : \mathcal{Z}^m \to \mathcal{H}$$

for which $\widehat{R}_l[\mathcal{A}^{\mathcal{H}}_{\mathrm{any}}(\boldsymbol{z}), \boldsymbol{z}] = 0$ has the same bound on performance.

This is good because one gets a general theory.

It is *bad* because the same bound holds for the *worst* algorithm.

Bayesians would say the problem is that we are not *conditioning on the data.*

Why Conditioning on the Data is Important

> That bayesians and frequentists are willing to discuss these matters is an important first step toward developing a theory that synthesizes both unconditional and conditional inference.
>
> — George Casella (1988)

Why condition — a simple example.
Suppose $X = (X_1, X_2)$, X_1, X_2 iid according to

$$\mathbf{P}_\theta(X_i = \theta - 1) = \mathbf{P}_\theta(X_i = \theta + 1) = \frac{1}{2}$$

for $-\infty < \theta < \infty$.
Consider the "confidence procedure"

$$C(x) := \begin{cases} \frac{x_1+x_2}{2} & \text{if } |x_1 - x_2| = 2 \\ x_1 - 1 & \text{if } |x_1 - x_2| = 0 \end{cases}$$

where $x = (x_1, x_2)$. Can check that

$$\mathbf{P}_\theta(C(X) \text{ contains } \theta) = 0.75 \quad \forall\theta$$

so we would be happy using $C(X)$ according to standard frequentist notions of acceptability.

Why Conditioning on the Data is Important

But after one sees the data:

If $|x_1 - x_2| = 2$ know *for certain* that $\theta \in C(X)$

If $|x_1 - x_2| = 0$, equally unsure whether $\theta = x_1 - 1$ or $x_1 + 1$.

Statisticians have expended considerable effort to develop procedures that have frequentist guarantees of performance *and* which can condition on the data to exploit a lucky observation.

Bayesian methods intrinsically condition on the data, but offer no frequentist guarantees of performance (most Bayesians would say this is no problem because such guarantees are neither necessary nor useful).

Many subtleties. To date really only for simple parameter estimation.

Something like this is needed in order to provide frequentist guarantees of performance for learning algorithms that do more than merely minimize and empirical risk functional.

Maximum Margin Algorithm

If there exists one separating hyperplane then there exist many of them. Why not choose the optimal one?

— Vladimir Vapnik

Maximum Margin Classifier: $\mathcal{H}_i(\boldsymbol{z})$ comprises hyperplanes $h_{\mathbf{w}}$ achieving margin $\gamma_{\boldsymbol{z}}(h_{\mathbf{w}}) = \gamma_i$ on $\boldsymbol{z}$. Here

$$\gamma_{\boldsymbol{z}}(h_{\mathbf{w}}) := \max_{(x_i,y_i)\in\boldsymbol{z}} y_i\langle \mathbf{w}, x_i\rangle / \|\mathbf{w}\|.$$

For linear hyperplanes, the risk bound is of the form

$$\psi(i) \leq \frac{c}{{\gamma_i}^2}\log^2(m) + c\log(1/\delta).$$

Maximum Margin algorithm: For $\mathcal{H}$ the set of linear hyperplanes.

$$\mathcal{A}_{\mathrm{MM}} := \boldsymbol{z} \mapsto \arg\max_{h\in\mathcal{H}} \gamma_{\boldsymbol{z}}(h)$$

("Optimal" only in the sense that it optimizes the particular ψ function used.) Try to understand $\mathcal{A}_{\mathrm{MM}}$ as an instance of $\mathcal{A}_{\mathrm{srm}}^{\boldsymbol{\mathcal{H}},\boldsymbol{\delta}}$.

Data Dependent SRM: $\mathcal{A}_{\text{dsrm}}^{\boldsymbol{\mathcal{H}}(\boldsymbol{z}),\boldsymbol{\delta}}$

New algorithm: *Penalize complexity of* $\mathcal{H}(\boldsymbol{z})$ *as if independent of* $\boldsymbol{z}$. Consider $\boldsymbol{\mathcal{H}}(\boldsymbol{z}) = (\mathcal{H}_i(\boldsymbol{z}))_i$. Suppose for data-independent $\mathcal{H}_i$ and $h_{\text{emp}}^i = \mathcal{A}_{\text{erm}}^{\mathcal{H}_i}(\boldsymbol{z})$ have a bound

$$R_l[h_{\text{emp}}^i] \leq \psi(\mathcal{H}, \boldsymbol{z}, \delta_i) =: \chi(i)$$

Let

$$i^* := \arg\min_i \chi(i) \qquad\qquad \mathcal{A}_{\text{dsrm}}^{\boldsymbol{\mathcal{H}}(\boldsymbol{z}),\boldsymbol{\delta}}(\boldsymbol{z}) := \mathcal{A}_{\text{erm}}^{\mathcal{H}_{i^*}(\boldsymbol{z})}(\boldsymbol{z})$$

Gist: penalize complexity ignoring data-dependence; apply SRM.

Problem: how to rigorously justify?

How to conceptualize what's going on?

Adherents of the various non-NPW schools take advantage of "lucky observations" to make more conclusive sounding statements than they would for "unlucky outcomes".

— Jack Kieffer: Conditional Confidence Statements . . . 1977

Can not strictly justify the algorithm as an application of SRM.

If as well as $\widehat{R}_l[h, \boldsymbol{z}] = 0$ we have $\gamma_{\boldsymbol{z}}(h) = \gamma \gg 0$, then $R_l[h]$ is small.

We are *lucky* if our data is like this.

We want to condition on the data like Bayesians do.

Would like to capture this notion in a general formal way.

BREAK

Have covered

- Induction (what's "inductive"?)
- Inductive Principles (what's the "principle"?)
- Empirical Risk Minimization
- Key Theorem of Learning Theory
- Conditioning on the Data

Yet to come:

- "Conditioning on the data" in a Frequentist (PAC) setting — The Luckiness Framework
- A new approach — Algorithmic Luckiness

Luckiness

I should be so lucky; lucky, lucky, lucky.
I should be so lucky in love.
— Kylie Minogue

The margin $\gamma_{\boldsymbol{z}}(h_{\mathbf{w}})$ measures how *lucky* $h_{\mathbf{w}}$ is on $\boldsymbol{z}$.
In general $L\colon \mathcal{H} \times \mathcal{Z}^m \to \mathbb{R}$.
Would like a bound that says with probability at least $1-\delta$ over a random draw of $\boldsymbol{z}$ according to $\mathbf{P}_{\mathsf{Z}^m}$ if $\widehat{R}_l[h, \boldsymbol{z}] = 0$ and $\omega(L(h, \boldsymbol{z}), \delta) \leq 2^d$ then

$$R_l[h] \leq \psi(m, d) \qquad \diamondsuit$$

The parameter d is an *effective complexity*.

There needs to be some restrictions on L: if we "use up" all of the information in the sample estimating its luckiness there is "none left" to estimate $R_l[h]$.
A bound like $\diamondsuit$ is a bound for the algorithm

$$\mathcal{A}^{L,\mathcal{H}}_{\text{lucky}} := \boldsymbol{z} \mapsto \arg\min_{h \in \mathcal{H}} \psi(m, \log \omega(L(h, \boldsymbol{z}), \delta))$$

Luckiness (continued)

Given a *luckiness function* $L\colon \mathcal{H} \times \mathcal{Z}^m \to \mathbb{R}$, the *level* is

$$\ell_L(h, \boldsymbol{z}) := \left|\{(l(g(x_i), y_i))_{i=1}^m \colon L(g, \boldsymbol{z}) \geq L(h, \boldsymbol{z})\}\right| ,$$

the number of dichotomies induced on $\boldsymbol{z}$ by functions at least as lucky as h. Require L to be well behaved:

L is *probably smooth* w.r.t. $\omega\colon \mathbb{R} \times (0, 1] \to \mathbb{N}$ if for all $m \in \mathbb{N}$ all distributions P_Z and all $\delta \in (0, 1]$

$$\mathsf{P}_{Z^{2m}}(\exists h \in \mathcal{H}\colon \ell_L(h, \mathbf{Z}_{[1:2m]}) > \omega(L(h, \mathbf{Z}_{[1:m]})), \delta)) \leq \delta\ .$$

If L is probably smooth w.r.t. ω, $\boldsymbol{\delta} = (\delta_i)_i$, $\sum_i \delta_i = \delta$, with probability at least $1 - \delta$ over a random draw of $\boldsymbol{z}$, if $\widehat{R}_l[h, \boldsymbol{z}] = 0$ and $\omega(L(h, \boldsymbol{z}), \delta_d/4) \leq 2^d$ then

$$R_l[h] \leq \frac{2}{m}\left(d + \log_2(4/\delta_d)\right) \qquad \heartsuit$$

Effectively $\boldsymbol{\mathcal{H}}(\boldsymbol{z}) = (\mathcal{H}_i(\boldsymbol{z}))_i$ with $\mathcal{H}_i(\boldsymbol{z}) = \left\{h \in \mathcal{H}\colon \omega(L(h, \boldsymbol{z}), \delta_i/4) \leq 2^i\right\}$.

Comments on Luckiness

- Can put $\mathcal{A}_{\mathrm{MM}}$ into this framework.
- The luckiness function L is how we *encode our prior knowledge*. We weight with δ_i the ith data-dependent hypothesis class

$$\mathcal{H}_i(\boldsymbol{z}) = \left\{h \in \mathcal{H} \colon \omega(L(h, \boldsymbol{z}), \delta_i/4) \leq 2^i\right\}$$

- Key practical difficulty is showing L is probably smooth w.r.t. a "good" ω — the smaller the ω the tighter $\heartsuit$ is.
- **Problem:** Still do not pay enough attention to the algorithm, which motivates . . .

Algorithmic Luckiness — Foundation

It is possible to prove a basic lemma for algorithms which means that the symmetrisation by permutation step only considers all hypotheses that can be learned using $\mathcal{A}$.

Basic lemma says:

$$\mathbf{P}_{\mathsf{Z}^m}\left(R_l\left[\mathcal{A}\left(\mathbf{Z}\right)\right] - \widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}\right), \mathbf{Z}\right] > \varepsilon\right) <$$

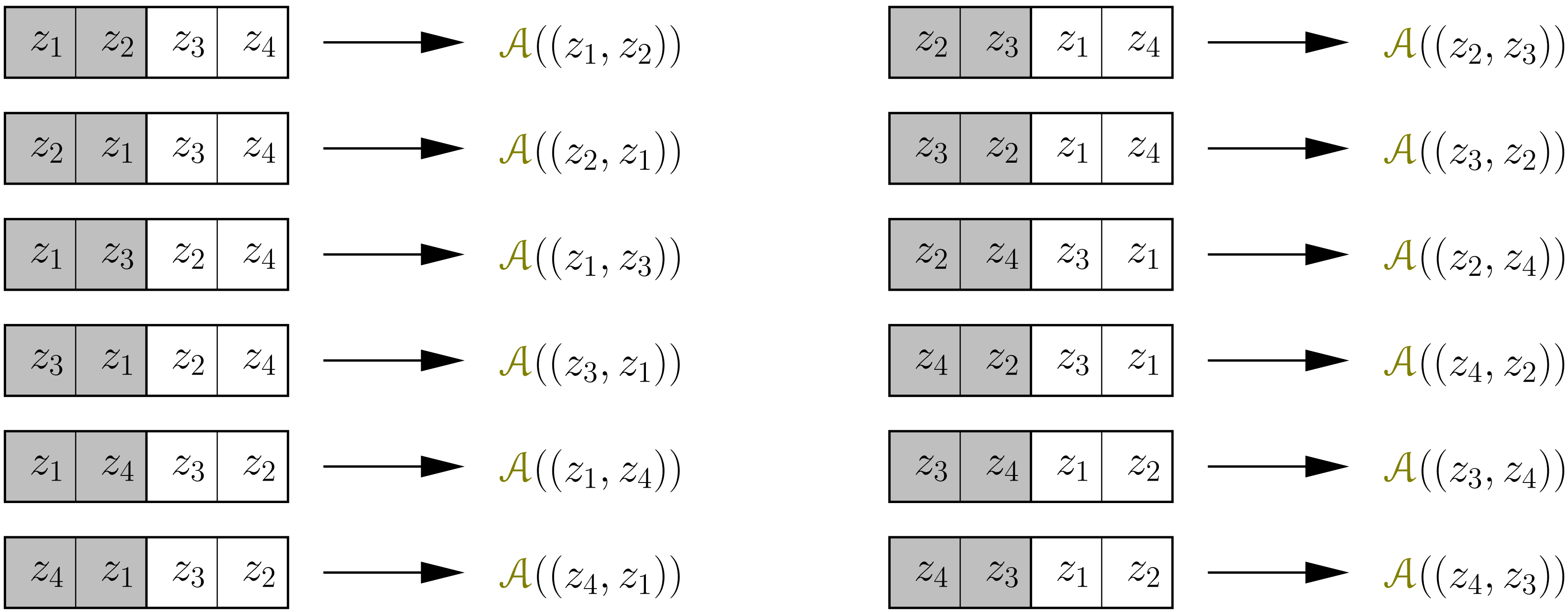
$$2 \cdot \mathbf{P}_{\mathsf{Z}^{2m}}\left(\widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[(m+1):2m]}\right] - \widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[1:m]}\right] > \frac{\varepsilon}{2}\right)$$

Again, we introduce an ordering between the at most $(2m)!$ hypotheses using an algorithmic luckiness $L(\mathcal{A}, \boldsymbol{z})$. This gives

$$\mathcal{H}\left(\mathcal{A}, L, \boldsymbol{z}\right) := \left\{\mathcal{A}\left(\Pi_{\mathbf{i}}\left(\boldsymbol{z}\right)_{[1:m]}\right) \mid \mathbf{i} \in \mathcal{I}\left(\mathcal{A}, L, \boldsymbol{z}\right)\right\},$$

$$\mathcal{I}\left(\mathcal{A}, L, \boldsymbol{z}\right) := \left\{\mathbf{i} \,\middle|\, L\left(\mathcal{A}, \Pi_{\mathbf{i}}\left(\boldsymbol{z}\right)_{[1:m]}\right) \geq L\left(\mathcal{A}, \boldsymbol{z}_{[1:m]}\right)\right\}.$$

Illustration of Basic Lemma when $m = 2$



Consider the simple case of $m = 2$. Consider all the hypotheses generated by $\mathcal{A}$ and take $\mathcal{H}(\mathcal{A}, L, \boldsymbol{z})$ to be those so generated that are at least as lucky as $\mathcal{A}((z_1, z_2))$ where the luckiness is measured on (z_1, z_2).

Algorithmic Luckiness — ω-smallness

Algorithmic luckiness $\mathcal{H}(\mathcal{A}, L, \boldsymbol{z})$ is a function of $\mathcal{A}$ directly.

Need to be able to bound the covering number $\mathcal{N}$ of hypotheses $h \in \mathcal{H}(\mathcal{A}, L, \boldsymbol{z})$ on the double sample $\boldsymbol{z}$ only using the luckiness on the first half, i.e. the training sample.

ω**–smallness of** L: *Given an algorithm* $\mathcal{A}$ *and a loss* l*, the algorithmic luckiness* L *is* ω*–small at scale* τ*, if for all* δ

$$\mathbf{P}_{\mathsf{Z}^{2m}}\left(\mathcal{N}\left(\tau, \mathcal{H}\left(\mathcal{A}, L, \mathbf{Z}\right), \ell_{l,1}\left(\mathbf{Z}\right)\right) > \omega\left(L\left(\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right)\right), \delta, \tau\right)\right) < \delta \,.$$

To prove this property we can only exploit that $\mathbf{P}_{\mathsf{Z}^{2m}}$ is a product measure.

Algorithmic Luckiness — Main Result

And the winner is . . . Lucky!

— Britney Spears

Algorithmic Luckiness Bound: *For all $[0,1]$-valued loss functions l, for all ω–small algorithmic luckiness functions L w.r.t. $\mathcal{A}$, for all τ, with probability at least $1-\delta$ over $\boldsymbol{z} \in \mathcal{Z}^m$,*

$$R_l[\mathcal{A}(\boldsymbol{z})] \leq \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] + \sqrt{\frac{8}{m}\left(\lceil d \rceil + \log_2\left(\frac{4m}{\delta}\right)\right)} + 4\tau \qquad \clubsuit$$

where $d = \log\left(\omega\left(L(\mathcal{A}, \boldsymbol{z}), \frac{\delta}{4m}, \tau\right)\right)$. If l is $\{0,1\}$-valued, then whenever $\mathcal{A}(\boldsymbol{z})$ has zero training error, $\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] = 0$, for $\tau = 1/2m$

$$R_l[\mathcal{A}(\boldsymbol{z})] \leq \frac{2}{m}\left(\lceil d \rceil + \log_2\left(\frac{4m}{\delta}\right)\right). \qquad \clubsuit$$

Application — Classical VC Setting

If $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ maps to a hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, we know that

$$\mathcal{H}(\mathcal{A}, L, \boldsymbol{z}) \subseteq \mathcal{H}$$

regardless of $\boldsymbol{z}$ and L.

VC-dimension is not most important quantity.

— Vladimir Vapnik: Dagstuhl, Germany (July 2001)

Thus, for the zero-one loss $l(\hat{y}, y) = \mathbb{I}_{\hat{y} \neq y}$ the growth function is an upper bound on $\mathcal{N}\left(|\boldsymbol{z}|^{-1}, \mathcal{H}(\mathcal{A}, L, \boldsymbol{z}), \ell_{l,1}(\boldsymbol{z})\right)$ and can thus serve as a ω function.

Neither the serendipity of the sample nor the properties of the algorithm $\mathcal{A}$ have been exploited!

As is widely known, this results in bounds which are quite loose.

Application — Compression Bounds

I used to be God. . .

— Manfred Warmuth: Dagstuhl, Germany (2001)

Sparsity luckiness: *If* $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ *is a compression scheme, that is,* $\mathcal{A}(\boldsymbol{z}) = \mathcal{R}(\mathcal{C}(\boldsymbol{z}))$*, then*

$$L_{\text{sparse}}(\mathcal{A}, \boldsymbol{z}) := -|\mathcal{C}(\boldsymbol{z})|$$

is ω*-small at any scale* τ*, where*

$$\omega(L, \delta, \tau) = \left(\frac{2em}{-L}\right)^{-L}.$$

Plugging this result into ♣ gives a new compression result for regression as well as resembling the original result of Littlestone and Warmuth (1986).

Sparsity Luckiness — Proof

You must be my lucky star . . .
But I'm the luckiest by far.
— Madonna

Since we only have to consider permutations Π_{i} where

$$\left|\mathcal{C}\left(\Pi_{\mathrm{i}}\left(\boldsymbol{z}\right)_{[1:m]}\right)\right| \leq \left|\mathcal{C}\left(\boldsymbol{z}_{[1:m]}\right)\right| =: -L_0$$

we know that the permutation invariant reconstruction function $\mathcal{R}$ never uses more than $-L_0$ examples.

The number of different choices of no more than $-L_0$ examples out of $2m$ (double sample size) is given by

$$\sum_{i=0}^{-L_0} \binom{2m}{i} \leq \left(\frac{2em}{-L_0}\right)^{-L_0} .$$

Application — Kernel Classifiers

Consider learning algorithms for kernel classifiers, that is,

$$\mathcal{H}_{\boldsymbol{\phi}} := \{\boldsymbol{x} \mapsto \langle \boldsymbol{\phi}(\boldsymbol{x}), \mathbf{w} \rangle \mid \mathbf{w} \in \mathcal{K}\}, \qquad \boldsymbol{\phi} : \mathcal{X} \to \mathcal{K} \subseteq \ell_2^n .$$

Assume that the learning algorithms $\mathcal{A}$ have the property that

$$\mathcal{A} : \boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle \boldsymbol{\phi}(\boldsymbol{x}), \mathbf{w}_{\boldsymbol{z}} \rangle \text{ where } \mathbf{w}_{\boldsymbol{z}} = \sum_{x_i \in \boldsymbol{x}} \hat{\alpha}_i \boldsymbol{\phi}(x_i) \text{ and } \|\mathbf{w}_{\boldsymbol{z}}\| = 1 .$$

Examples are SVMs, BPMs and the perceptron algorithm.

Let the (normalised) margin $\Gamma(\boldsymbol{z})$ be defined by

$$\Gamma(\boldsymbol{z}) := \min_{(x_i, y_i) \in \boldsymbol{z}} \frac{y_i \langle \boldsymbol{\phi}(x_i), \mathbf{w}_{\boldsymbol{z}} \rangle}{\|\boldsymbol{\phi}(x_i)\| \cdot \|\mathbf{w}_{\boldsymbol{z}}\|} .$$

Application — Kernel Classifiers (cont.)

Margin Luckiness: *Let* $\varepsilon_i(\boldsymbol{x})$ *be the smallest* $\epsilon > 0$ *such that* $\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_m)\}$ *can be covered by at most* i *balls of radius less than or equal to* ϵ. *For the loss* $l(\hat{y}, y) = \mathbb{I}_{y\hat{y} \leq 0}$, *the luckiness function*

$$L_{\text{margin}}(\mathcal{A}, \boldsymbol{z}) = -\min\left\{ i \in \mathbb{N} \;\middle|\; i \geq \left(\frac{\varepsilon_i(\boldsymbol{x}) \sum_{j=1}^{m} |\hat{\alpha}_j|}{\Gamma(\boldsymbol{z})} \right)^2 \right\}$$

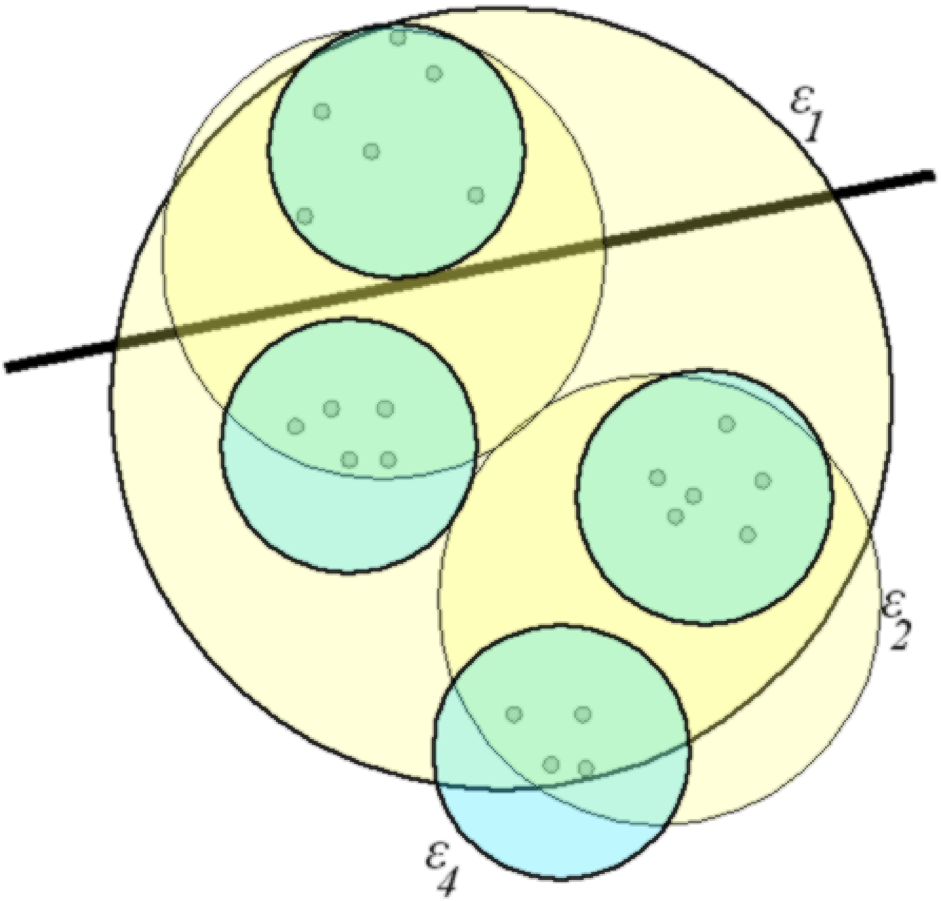
is ω*–small at scale* $1/2m$ *where*

$$\omega(L, \delta, 1/2m) = \left(\frac{2em}{-L} \right)^{-2L}.$$

The bound comprises 3 main terms: *margin* $\Gamma(\boldsymbol{z})$, *sparsity surrogate* $\sum_{j=1}^{m} |\hat{\alpha}_j|$ and a factor depending on the *distribution of the data* $\varepsilon_i(\boldsymbol{x})$.

Kernel Classifiers — Meaning of $\varepsilon_i(\boldsymbol{x})$

The sequence $(\varepsilon_i(\boldsymbol{x}))_i$ measures how clumpy the data is.



A small number of small clumps means $\varepsilon_i(\boldsymbol{x})$ is small for small i.

Compare with the idea of "kernel alignment".

Application — Kernel Classifiers (proof)

Makovoz theorem shows that for all $\boldsymbol{z} \in \mathcal{Z}^m$ there exists a weight vector $\tilde{\mathbf{w}} = \sum_{i=1}^m \tilde{\alpha}_i \boldsymbol{\phi}(x_i)$ such that

$$\|\tilde{\mathbf{w}} - \mathbf{w}_{\boldsymbol{z}}\|^2 \leq \Gamma^2(\boldsymbol{z})$$

and $\|\tilde{\boldsymbol{\alpha}}\|_0 \leq -L_{\text{margin}}(\mathcal{A}, \boldsymbol{z}) =: -L_0$.

It follows that $\langle \mathbf{w}_{\boldsymbol{z}}, \tilde{\mathbf{w}} / \|\tilde{\mathbf{w}}\| \rangle \geq \sqrt{1 - \Gamma^2(\boldsymbol{z})}$; that is, $\tilde{\mathbf{w}}$ still correctly classifies $\boldsymbol{z}$.

For every of the no more than $\left(\frac{2em}{-L_0}\right)^{-L_0}$ many subsamples $\tilde{\boldsymbol{z}} \subseteq \boldsymbol{z}$, $\tilde{\mathbf{w}}$ lives in a space of dimension no more than $-L_0$.

By an application of the growth function bound, each $\tilde{\mathbf{w}}$ can achieve no more than $\left(\frac{2em}{-L_0}\right)^{-L_0}$ many dichotomies on $\boldsymbol{z}$.

Discussion
Discussion

Algorithmic luckiness framework differs from classical statistical learning theory aproaches in that it does not use the crude step of viewing algorithms just in terms of their hypothesis space.

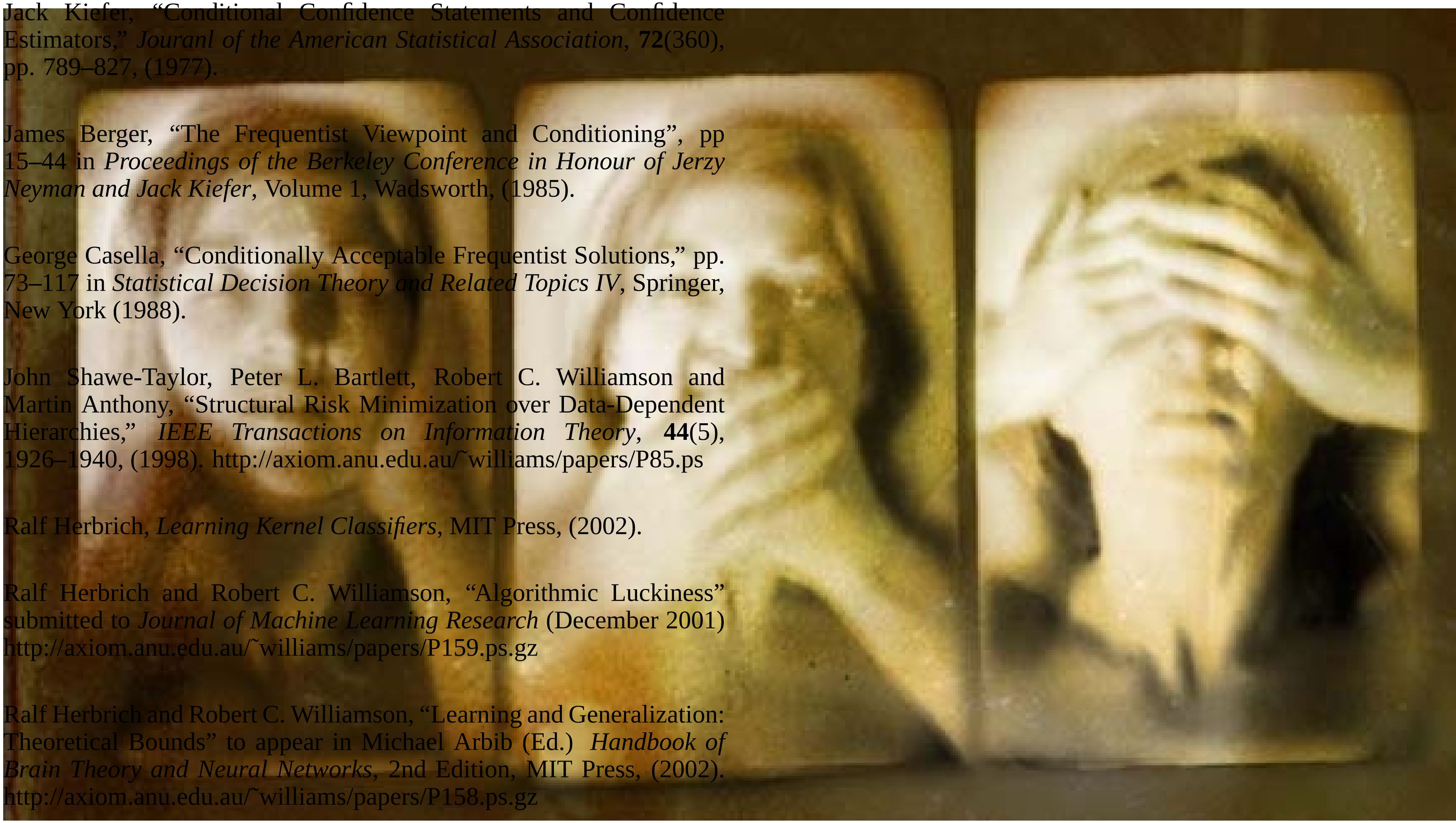
Generalization of standard VC results. Get agnostic and realizable bounds.

Example of maximum margin algorithm illustrates that the framework has the power to develop new insights into what makes algorithms perform well.

Main point is that it provides new theoretical tools for understanding algorithms "smarter" that Empirical Risk Minimization.

Hope is that by analysing algorithms in this (or related) ways, will be able to better discern the features about particular learning problems that make them easy or difficult.

References and Slides



Jack Kiefer, "Conditional Confidence Statements and Confidence Estimators," *Jouranl of the American Statistical Association*, **72**(360), pp. 789–827, (1977).

James Berger, "The Frequentist Viewpoint and Conditioning", pp 15–44 in *Proceedings of the Berkeley Conference in Honour of Jerzy Neyman and Jack Kiefer*, Volume 1, Wadsworth, (1985).

George Casella, "Conditionally Acceptable Frequentist Solutions," pp. 73–117 in *Statistical Decision Theory and Related Topics IV*, Springer, New York (1988).

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson and Martin Anthony, "Structural Risk Minimization over Data-Dependent Hierarchies," *IEEE Transactions on Information Theory*, **44**(5), 1926–1940, (1998). http://axiom.anu.edu.au/˜williams/papers/P85.ps

Ralf Herbrich, *Learning Kernel Classifiers*, MIT Press, (2002).

Ralf Herbrich and Robert C. Williamson, "Algorithmic Luckiness" submitted to *Journal of Machine Learning Research* (December 2001) http://axiom.anu.edu.au/˜williams/papers/P159.ps.gz

Ralf Herbrich and Robert C. Williamson, "Learning and Generalization: Theoretical Bounds" to appear in Michael Arbib (Ed.) *Handbook of Brain Theory and Neural Networks*, 2nd Edition, MIT Press, (2002). http://axiom.anu.edu.au/˜williams/papers/P158.ps.gz

Slides at **http://axiom.anu.edu.au/˜williams/papers/P156.pdf**