Data-derived estimation

Venkat Anantharam
University of California, Berkeley

From Local to Global Information
Honolulu, Hawaii
February 5 -7, 2020

Partly joint work with
Narayana Prasad Santhanam
University of Hawaii, Manoa
and
Wojtek Szpankowski
Purdue University

Outline

1. Insurance
2. Compression
3. Learning

General framework

- Let $\mathcal{P}$ be a collection of probability distributions on some measurable space. Throughout this talk we will take this space to be the integers $\mathbb{N}$ with the usual σ-field.
- We observe $X_1, X_2, \ldots$ drawn i.i.d. with distribution $p \in \mathcal{P}$. However, we do not know which p is in effect.
- At some time we make a decision (decide some predicate is satisfied). We may be wrong with some nonzero probability.
- Satisfying the predicate may also involve active participation of the observer, e.g. using some algorithm .

Strong, weak and data-derived weak

- If the time at which we decide that the predicate holds is uniform over all $p \in \mathcal{P}$, the class $\mathcal{P}$ is strongly amenable to the predicate.
- If, for every $p \in \mathcal{P}$, there is a time at which the predicate is satisfied, the class $\mathcal{P}$ is weakly amenable to the predicate.
- We propose a novel notion of what it means for the class $\mathcal{P}$ to be data-derived weakly amenable to the predicate. Basically, the time at which we decide the predicate holds should be based on the observed data.
- More generally, these notions are applied to scenarios where there are multiple predicates.

Outline

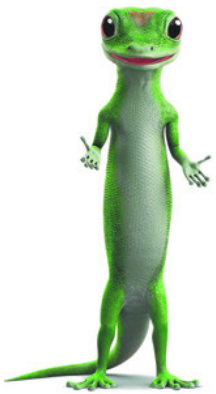
1 Insurance

2 Compression

3 Learning

Motivating context

Insurer



- The loss distribution is assumed to be a member of a class $\mathcal{P}$ but otherwise not known to the insurer. We assume i.i.d. losses over time.
- The premium setting strategy should work universally over loss distributions in this class
- The insurer can observe losses for a while before deciding to write a contract.
- We ask: which classes $\mathcal{P}$ are insurable?

Insurability as tail domination

- A tail domination strategy is a sequence of functions $\Phi := (\Phi_n, n \geq 0)$, where $\Phi_n : \mathbb{R}^n \to \mathbb{R}_+$.
- Intuitively, the aim is to make sure that for all n, after an initial observation period, we will have, with only a small probability of error,

$$\Phi_n(X_1, \ldots, X_n) > X_{n+1}.$$

- Insurability is basically equivalent to tail domination, except that the insurer needs to adjust its premiums to correct for the built up excess from past premiums after claims have been paid out.

Strong insurability of $\mathcal{P}$

- $\mathcal{P}$ is called strongly insurable if, for every $\eta > 0$, there exists a tail domination strategy $\Phi^\eta := (\Phi^\eta_n.n \geq 0)$ and a time $n_0(\eta, \Phi^\eta)$ such that, for all $p \in \mathcal{P}$, we have

$$P_p(\text{ there exists } n \geq n_0(\eta, \Phi^\eta) \text{ such that } \Phi^\eta_n(X_1, \ldots, X_n) \leq X_{n+1}) < \eta.$$

Characterizing strong insurability

- $\mathcal{P}$ is called tight if, for every $\epsilon > 0$, there is some $K^\epsilon < \infty$ such that
$$p(X > K^\epsilon) < \epsilon \text{ for all } p \in \mathcal{P}.$$
- Tightness implies strong insurability.
- Tightness is not necessary for strong insurability.
Example: Let $\mathcal{P}$ be the collection of uniform distributions on $\{1, \ldots, 2N\}$ for $N \geq 1$, call these u_N.
- $$P_{u_N}(\max_{1 \leq i \leq n} X_i \leq N) = (\frac{1}{2})^n.$$
- Observe $n := \lceil \log_2 \frac{1}{\eta} \rceil$ samples, then set the tail dominator to $2 \max_{1 \leq i \leq n} X_i + 1$.

Weak insurability of $\mathcal{P}$

- $\mathcal{P}$ is called weakly insurable if, for every $\eta > 0$, there exists a tail domination strategy $\Phi^\eta := (\Phi^\eta_n . n \geq 0)$ such that, for every $p \in \mathcal{P}$, there exists a time $n_0(\eta, \Phi^\eta, p)$ such that

$$P_p(\text{ there exists } n \geq n_0(\eta, \Phi^\eta, p) \text{ such that } \Phi^\eta_n(X_1, \ldots, X_n) \leq X_{n+1}) < \eta.$$

Characterizing weak insurability

- Recent results of Chonglong Wu and Narayana Santhanam indicate that if $\mathcal{P}$ is a countable union of tight classes then $\mathcal{P}$ is weakly insurable.

Data-derived weak insurability of $\mathcal{P}$

- Let $\mathbb{N}^*$ denote the set of all finite strings of integers (including the empty string).

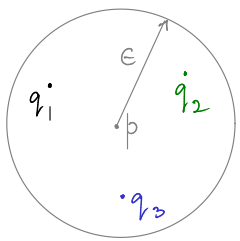
- A stopping rule is a map $\tau : \mathbb{N}^* \to \{0, 1\}$ such that
(i) If $\tau(x) = 1$ and x is a prefix of y then $\tau(y) = 1$.
(ii) $P_p(\tau(X^n) \text{ is eventually equal to } 1) = 1$ for all $p \in \mathcal{P}$.

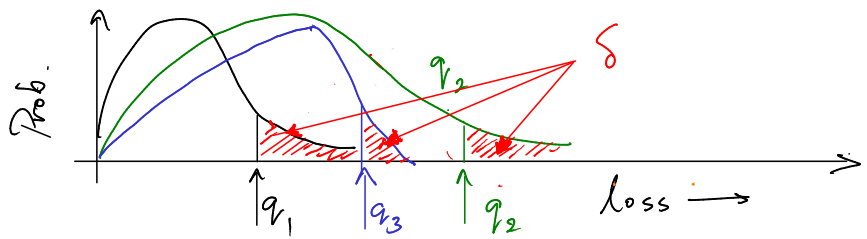
- $\mathcal{P}$ is called data-derived weak insurable if, for every $\eta > 0$, there is a tail domination strategy $\Phi^\eta := (\Phi^\eta_n, n \geq 0)$ and stopping rule τ^η such that

$$P_p(\Phi^\eta_n(X_1, \ldots, X_n) \leq X_{n+1} \text{ for some } n \text{ such that } \tau^\eta(X_1, \ldots, X_n) = 1) < \eta.$$

Characterizing data-derived weak insurability.

- A probability distribution $p \in \mathcal{P}$ is called deceptive for tails if the intersection of every open l^1-neighborhood of p with $\mathcal{P}$ is not tight .





- $\sup_{\text{all } q \in \mathcal{P} \text{ in the neighborhood of } p \in \mathcal{P}}$ $1-\delta$ percentile of q bounded?

  Above statement true for all δ?

- *Theorem*: $\mathcal{P}$ is data-derived weak insurable iff each $p \in \mathcal{P}$ is not deceptive for tails.

- Narayana Santhanam and VA, "Agnostic insurability of model classes", JMLR, pp. 2329 -2355, 2015.

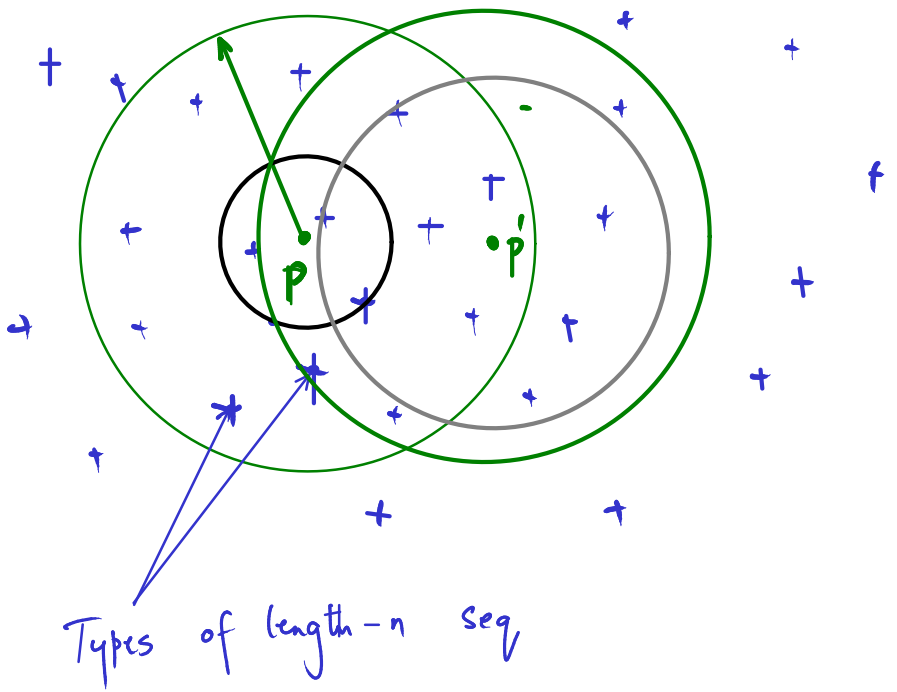
Technical ingredients in the proof

- Let
$$J(p,q) := D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2}) \ .$$
where
$$D(p\|q) := \sum_i p(i) \log \frac{p(i)}{q(i)} \ .$$
- For any two probability distributions p and q on $\mathbb{N}$, we have
$$\frac{1}{4 \ln 2}|p-q|_1^2 \leq J(p,q) \leq \frac{1}{\ln 2}|p-q|_1 \ .$$
- For any three probability distributions $p, q, r \in \mathbb{N}$ we have
$$J(p,q) + J(q,r) \geq \frac{\ln 2}{8} J(p,r)^2 \ .$$

Sketch of proof of sufficiency

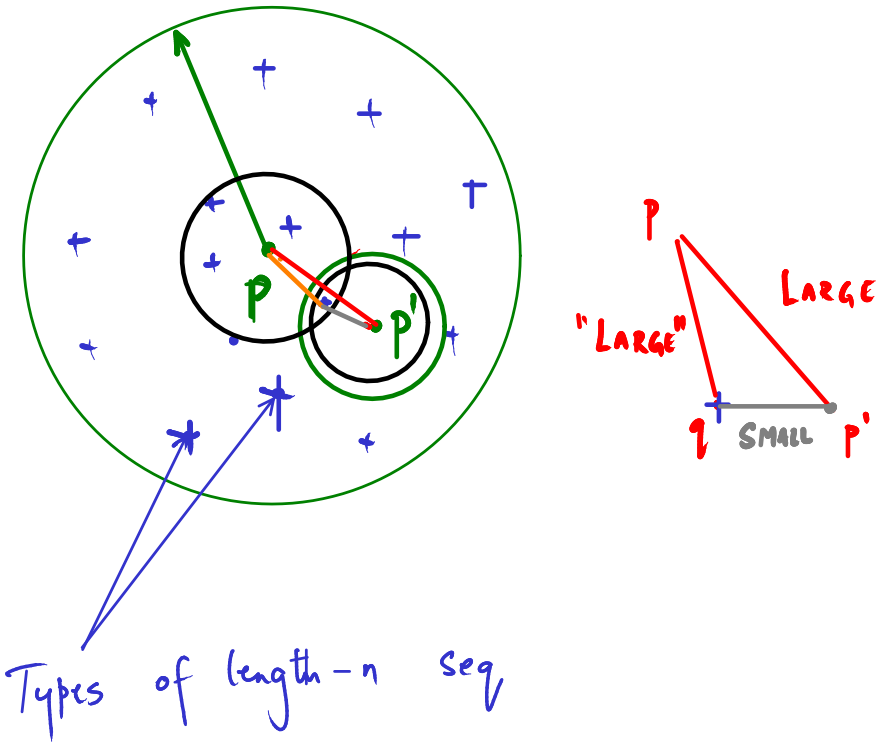
- Each $p \in \mathcal{P}$ has an open l^1-neighborhood of positive radius ϵ_p, called its reach such that, for all percentiles $1-\delta$, there is a universal upper bound on the $1-\delta$ percentile of all $q \in \mathcal{P}$ lying in the reach of p.
- Since l^1 is a separable metric space, it is Lindelöf, so one can cover $\mathcal{P}$ by a countable union of reaches, enumerated in some way.
- When $p \in \mathcal{P}$ is in effect, the insurer computes the *type* (empirical distribution) of the observed losses as time progresses. This will eventually enter the union of this countable collection of reaches.
- The decision that needs to be made is whether to allow oneself to be trapped by a reach.

Good traps



Types captured by p' will not harm p if p is within the reach of p'.

Bad traps



p generates types captured by hostile p' —but such types have low probability

Putting these together gives a constructive proof for the sufficiency condition

Examples

If $\mathcal{P}$ is:

- The set of *all* finite support distributions, then it is not insurable
- The set of all *uniform* distributions, then it is insurable
- The set of all *monotone* distributions (with finite entropy), then it is not insurable
- The set of all distributions with a *uniformly bounded mean*, then it is insurable
- The set of all distributions with a *finite mean*, then it is not insurable

Outline

1 Insurance

2 Compression

3 Learning

Strong, weak and data-derived compressibility

- We restrict attention to classes of probability distributions $\mathcal{P}$ on $\mathbb{N}$ comprised of probability distributions with finite entropy.
- We introduce a novel notion of data-derived weak compressiblity .

Strong compressibility of $\mathcal{P}$

- $\mathcal{P}$ (comprised of probability distributions with finite entropy) is called strongly compressible if there exists a probability measure q on $\mathbb{N}^{\mathbb{N}}$ such that

$$\limsup_{n\to\infty} \sup_{p\in\mathcal{P}} \frac{1}{n}\mathbb{E}_p \log \frac{p(X^n)}{q(X^n)} = 0 \ .$$

- Equivalently, for all $\epsilon > 0$ and all $\eta > 0$, there exists a probability measure $q^{\epsilon,\eta}$ on $\mathbb{N}^{\mathbb{N}}$ and $n_0(\epsilon, \eta, q^{\epsilon,\eta})$ such that, for all $n \geq n_0(\epsilon, \eta, q^{\epsilon,\eta})$ and all $p \in \mathcal{P}$, we have

$$P_p\Big(\sum_{x^n\in\mathbb{N}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \geq \epsilon \text{ for some } n \geq n_0(\epsilon, \eta, q^{\epsilon,\eta})\Big) < \eta$$

- The introduction of η here is meaningless, but is done for future reference
- The equivalence of a single q to multiple $q^{\epsilon,\eta}$ comes from considering $\sum_{k\geq 1, l\geq 1} \frac{1}{k(k+1)l(l+1)} q^{\frac{1}{2^k}, \frac{1}{2^l}}$.

Weak compressibility of $\mathcal{P}$

- $\mathcal{P}$ (comprised of probability distributions with finite entropy) is called weakly compressible if there exists a probability measure q on $\mathbb{N}^{\mathbb{N}}$ such that, for each $p \in \mathcal{P}$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}_p \log \frac{p(X^n)}{q(X^n)} = 0 \ .$$

- Equivalently, for all $\epsilon > 0$ and all $\eta > 0$, there exists a probability measure $q^{\epsilon,\eta}$ on $\mathbb{N}^{\mathbb{N}}$ such that, for all $p \in \mathcal{P}$, there exists $n_0(\epsilon, \eta, q^{\epsilon,\eta}, p)$ such that, for all $n \geq n_0(\epsilon, \eta, q^{\epsilon,\eta}, p)$, we have

$$P_p\Big(\sum_{x^n \in \mathbb{N}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \geq \epsilon \text{ for some } n \geq n_0(\epsilon, \eta, q^{\epsilon,\eta}, p)\Big) < \eta.$$

- The introduction of η here is meaningless.

Data-derived weak compressibility of $\mathcal{P}$

- Let $\mathbb{N}^*$ denote the set of all finite strings of integers (including the empty string).

- A stopping rule is a map $\tau : \mathbb{N}^* \to \{0, 1\}$ such that
(i) If $\tau(x) = 1$ and x is a prefix of y then $\tau(y) = 1$.
(ii) $P_p(\tau(X^n)$ is eventually equal to $1) = 1$ for all $p \in \mathcal{P}$.

- $\mathcal{P}$ (comprised of probability distributions with finite entropy) is called data-derived weakly compressible if for all $\epsilon > 0$ and all $\eta > 0$, there exists a probability measure $q^{\epsilon,\eta}$ on $\mathbb{N}^{\mathbb{N}}$ and a stopping rule $\tau^{\epsilon,\eta,q^{\epsilon,\eta}}$ such that, for all $p \in \mathcal{P}$, we have

$$P_p\Big(\sum_{x^n \in \mathbb{N}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \geq \epsilon \text{ for } n \text{ s.t. } \tau^{\epsilon,\eta,q^{\epsilon,\eta}}(X_1, \ldots, X_n) = 1\Big) < \eta.$$

Characterizing data-derived weak compressibility

- $p \in \mathcal{P}$ is called deceptive for compression if, for all probability measures q on $\mathbb{N}^{\mathbb{N}}$, we have

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \sup_{p' \in B(p,\epsilon)} \frac{1}{n} \mathbb{E}_{p'} \log \frac{p'(X^n)}{q(X^n)} > 0 \ ,$$

  where $B(p, \epsilon)$ denotes the intersection open l^1-ball of radius ϵ around p with $\mathcal{P}$.

- Here $\mathcal{P}$ is assumed to be comprised of probability distributions on $\mathbb{N}$ with finite entropy.
- *Theorem**: (with Narayana Santhanam, Wojtek Szpankowski and Aleksandr Kavcic)
  $\mathcal{P}$ is data-derived weakly compressible iff no $p \in \mathcal{P}$ is deceptive for compression.
- Waiting (for five years already and still counting) for Narayana Santhanam to write the journal paper.

Proofs

- For the sufficiency proof, assume that each $p \in \mathcal{P}$ is not deceptive for compression.
- Then, for each $p \in \mathcal{P}$, there is some probability measure q_p on $\mathbb{N}^{\mathbb{N}}$ such that, for all $\delta > 0$, there is some $\epsilon_{p,\delta} > 0$ and $n_0(p, \delta)$ such that, for all $n \geq n_0(p, \delta)$ and all $p' \in B(p, \epsilon_{p,\delta})$, we have

$$\frac{1}{n}\mathbb{E}_{p'} \log \frac{p'(X^n)}{q_p(X^n)} < \delta.$$

- What this means is that treating q_p as the compression measure for all $p' \in B(p, \epsilon_{p,\delta})$ is good enough to get compression redundancy less than δ.
- Hence, for a given $\delta > 0$ we define $\epsilon_{p,\delta}$ as above to now be the reach of $p \in \mathcal{P}$.
- The proof is then somewhat different from, but of the same flavor as, the proof of sufficiency for the insurability theorem.

Examples

If $\mathcal{P}$ is:

- The set of all *uniform* distributions, then it is d.w.c. but it is not strongly compressible.
- The set of all distributions with a *finite mean*, then it is weakly compressible but it is not d.w.c..
- The set of all *monotone* distributions with finite entropy, then it is weakly compressible but it is not d.w.c..
- The set of all *monotone* distributions with finite entropy, and with the uniformly bounded second moment of the self information, i.e. $\mathbb{E}_p[(\log \frac{1}{p(X)})^2] \leq h < \infty$, then then it is strongly compressible.

Outline

1 Insurance

2 Compression

3 Learning

Rademacher complexity

- Let $A \subseteq \mathbb{R}^n$.
- Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. uniform $\{\pm 1\}$-valued random variables.
- $E_\epsilon[\sup_{a \in A} \sum_{i=1}^n \epsilon_i a_i]$ is called the Rademacher complexity of the set A.
- Note that $\sup_{a \in A} \sum_{i=1}^n \epsilon_i a_i$ is sub-Gaussian with parameter $4 \sum_{i=1}^n \sup_{a \in A} a_i^2$.

Glivenko-Cantelli class of functions

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on a standard measurable space $(\mathbb{X}, \mathcal{X})$.
- Let p be a probability distribution on $(\mathbb{X}, \mathcal{X})$.
- If $X_1, \ldots, X_n$ are i.i.d. with law p, we write p_n for their empirical distribution.
  (Note that p_n is a random variable.)
- $\|p_n - p\|_{\mathcal{F}}$ denotes $\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - E_p[f(X)]|$. This is sub-Gaussian with parameter $\frac{1}{n}$.
- If $\|p_n - p\|_{\mathcal{F}} \xrightarrow{P} 0$ as $n \to \infty$ we say that p verifies a uniform law of large numbers (uniformity is over $\mathcal{F}$, but p is fixed). and $\mathcal{F}$ is then called a weak Glivenko-Cantelli class of functions for p.
- If $\|p_n - p\|_{\mathcal{F}} \xrightarrow{a.s.} 0$ as $n \to \infty$ then $\mathcal{F}$ is then called a strong Glivenko-Cantelli class of functions for p.

Rademacher complexity of $\mathcal{F}$ with respect to p

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$, and p be a probability distribution on $(\mathbb{X}, \mathcal{X})$.
- Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. uniform $\{\pm 1\}$-valued random variables.
- If $X_1, \ldots, X_n$ are i.i.d. with law p, consider $\{(f(X_1), \ldots, f(X_n))^T \in \mathbb{R}^n\}$, and let

$$Rad_n(\mathcal{F}, p) := E_p[E_\epsilon[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i)]]$$

(Note the normalization by $\frac{1}{n}$.)

Rademacher complexity and Glivenko Cantelli classes

- $$\frac{1}{2}Rad_n(\mathcal{F}, p) - \sqrt{\frac{\log 2}{2n}} \leq E_p\|p_n - p\|_{\mathcal{F}} \leq 2Rad_n(\mathcal{F}, p)$$
- Since $\|p_n - p\|_{\mathcal{F}}$ is sub-Gaussian with parameter $\frac{1}{n}$, we have, for all $\delta > 0$,

  $$E_p\|p_n - p\|_{\mathcal{F}} - \delta \leq \|p_n - p\|_{\mathcal{F}} \leq E_p\|p_n - p\|_{\mathcal{F}} + \delta,$$

  with probability at least $1 - 2\exp(-2\delta^2 n)$.
- Hence $\mathcal{F}$ is a strong Glivenko Cantelli class for p iff $Rad_n(\mathcal{F}, p) \to 0$ as $n \to \infty$.
- Also, since $\|p_n - p\|_{\mathcal{F}} \leq 2$ under our assumptions, if $\|p_n - p\|_{\mathcal{F}} \xrightarrow{P} 0$ as $n \to \infty$, we have $Rad_n(\mathcal{F}, p) \to 0$ as $n \to \infty$. Thus we will not need to distinguish between weak and strong Glivenko-Cantelli classes under our assumptions.

Uniform Glivenko-Cantelli class of functions

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.
- $\mathcal{F}$ is called a uniform Glivenko-Cantelli class of functions if, for all $\epsilon > 0$, we have

$$\sup_p P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) \to 0$$

  as $n \to \infty$
- Equivalently, for all $\epsilon > 0$ and $\eta > 0$, there is $n_0(\epsilon, \eta)$ such that, for all p, we have

$$P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) < \eta, \text{ for all } n \geq n_0(\epsilon, \eta).$$

- Notice the strong sense in which the predicate is satisfied.

Universal Glivenko-Cantelli class of functions

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $\mathbb{N}$.
- $\mathcal{F}$ is called a universal Glivenko-Cantelli class of functions if it is a Glivenko-Cantelli class for each probability distribution p on $\mathbb{N}$.
- Equivalently, for all $\epsilon > 0$ and $\eta > 0$, and all p, there is $n_0(\epsilon, \eta, p)$ such that

$$P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) < \eta \text{ for all } n \geq n_0(\epsilon, \eta, p).$$

- Notice the weak sense in which the predicate is satisfied.

$\mathcal{P}$ admitting $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense

- Let $\mathcal{P}$ be a collection of probability distributions on $(\mathbb{X}, \mathcal{X})$.
- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.
- We say that $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense if, for all $\epsilon > 0$, we have

$$\sup_{p \in \mathcal{P}} P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) \to 0$$

  as $n \to \infty$
- Equivalently, for all $\epsilon > 0$ and $\eta > 0$, there is $n_0(\epsilon, \eta)$ such that, for all $p \in \mathcal{P}$, we have

$$P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) < \eta \text{ for all } n \geq n_0(\epsilon, \eta).$$

Characterizing $\mathcal{P}$ that admit $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense

- $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense iff

$$\sup_{p \in \mathcal{P}} Rad_n(\mathcal{F}, p) \to 0$$

as $n \to \infty$.

$\mathcal{P}$ admitting $\mathcal{F}$ as a Glivenko-Cantelli class in the weak sense

- Let $\mathcal{P}$ be a collection of probability distributions on $(\mathbb{X}, \mathcal{X})$.
- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.
- We say that $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the weak sense if $\mathcal{F}$ is a Glivenko-Cantelli class for each $p \in \mathcal{P}$.
- Equivalently, for all $\epsilon > 0$ and $\eta > 0$, and all $p \in \mathcal{P}$, there is $n_0(\epsilon, \eta, p)$ such that

$$P_p(\|p_n - p\|_{\mathcal{F}} \geq \epsilon) < \eta \text{ for all } n \geq n_0(\epsilon, \eta, p).$$

Characterizing $\mathcal{P}$ that admit $\mathcal{F}$ as a Glivenko-Cantelli class in the weak sense

- $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the weak sense iff, for all $p \in \mathcal{P}$, we have

$$Rad_n(\mathcal{F}, p) \to 0$$

as $n \to \infty$.

$\mathcal{P}$ admitting $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense

- Let $\mathbb{X}^*$ denote the set of all finite strings of elements from $\mathbb{X}$ (including the empty string).
- A stopping rule is a map $\tau : \mathbb{X}^* \to \{0, 1\}$ such that
  (i) If $\tau(\bar{x}) = 1$ and $\bar{x}$ is a prefix of $\bar{y}$ then $\tau(\bar{y}) = 1$.
  (ii) $P_p(\tau(X^n)$ is eventually equal to $1) = 1$ for all $p \in \mathcal{P}$.
- We say that $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense if, for all $\delta > 0$ and $\eta > 0$,there is a stopping time $\tau^{\delta,\eta}$ such that, for all $p \in \mathcal{P}$, we have

  $$P_p(P_p(\|p_n - p\|_{\mathcal{F}}) < \delta \text{ for all } n \text{ such that } \tau^{\delta,\eta}(X_1, \ldots, X_n) = 1) \geq 1 - \eta.$$

- If the collection of all probability distributions on $(\mathbb{X}, \mathcal{X})$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense, we say that $\mathcal{F}$ is a Glivenko-Cantelli class in the data-derived weak sense.

Characterizing $\mathcal{P}$ that admit $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense

- $p \in \mathcal{P}$ is called deceptive for $\mathcal{F}$ if we have

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{q \in B(p,\delta)} Rad_n(\mathcal{F}, q) > 0.$$

Theorem *

$\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense iff no $p \in \mathcal{P}$ is deceptive for $\mathcal{F}$.

Proofs

- For the sufficiency proof, assume that each $p \in \mathcal{P}$ is not deceptive for $\mathcal{F}$.
- Then, for each $p \in \mathcal{P}$, for every $\delta > 0$ there is some $\epsilon_{p,\delta} > 0$ and $n_0(p, \delta)$ such that, for all $n \geq n_0(p, \delta)$ and all $q \in B(p, \epsilon_{p,\delta})$, we have

$$Rad_n(\mathcal{F}, q) < \delta.$$

- What this means is that the empirical distribution of the observations will be within δ of the driving distribution, uniformly over $\mathcal{F}$, as long as the driving distribution is within the reach of p.
- Hence, for a given $\delta > 0$ we define $\epsilon_{p,\delta}$ as above to now be the reach of $p \in \mathcal{P}$.
- The proof is then somewhat different from, but of the same flavor as, the proof of sufficiency for the insurability theorem.

Pseudometric

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.
- For $1 \leq r < \infty$ and fixed $x_1, \ldots, x_n \in \mathbb{X}$, one can define a pseudometric on $\mathcal{F}$ by

$$d_{r,x^n}(f,g) := \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|^r \right)^{\frac{1}{r}},$$

and one can define

$$d_{\infty,x^n}(f,g) := \max_{1 \leq i \leq n} |f(x_i) - g(x_i)|.$$

Here $f, g \in \mathcal{F}$.

Metric entropy

- Given $\alpha > 0$, let $\mathcal{N}(\alpha, \mathcal{F}, d)$ denote the α-covering number of $\mathcal{F}$ for any pseudometric d on $\mathcal{F}$, i.e. the smallest cardinality of a collection of α-balls in the pseudometric d that cover $\mathcal{F}$.
- The function $\alpha \to \log \mathcal{N}(\alpha, \mathcal{F}, d)$ is called the metric entropy function of $\mathcal{F}$ under the pseudometric d.
- Let p be a probability distribution on $(\mathbb{X}, \mathcal{X})$. Then we have

$$\frac{C_{down}}{\log n} \sup_{0<\alpha<\infty} \left(\alpha E_p \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, d_{2,X^n})}{n}} \right) \leq Rad_n(\mathcal{F}, p)$$
$$\leq C_{up} E_p \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, d_{2,X^n})}{n}} d\alpha.$$

- The lower bound is called Sudakov's bound and the upper bound is called Dudley's entropy integral bound.

Uniform Glivenko-Cantelli classes and metric entropy

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.

Theorem

$\mathcal{F}$ is a uniform Glivenko-Cantelli class iff for some (and equivalently for all) $1 \leq r \leq \infty$ it holds for all $\alpha > 0$ that

$$\lim_{n\to\infty} \frac{1}{n} \sup_{x^n \in \mathbb{N}^n} \log \mathcal{N}(\alpha, \mathcal{F}, d_{r,x^n}) = 0.$$

R. M. Dudley, E. Giné and J. Zinn. "Uniform and universal Glivenko-Cantelli classes", 1991.

- Let us specialize this result to $\mathbb{N}$ with its usual σ-field.

Universal Glivenko-Cantelli classes and metric entropy

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$.
- Let $d_{l^1(p)}$ denote the pseudometric on $\mathcal{F}$ given by $\sum_{x\in\mathbb{N}} |f(x) - g(x)|p(x)$, where $f, g \in \mathcal{F}$.

Theorem

$\mathcal{F}$ is a universal Glivenko-Cantelli class iff we have, for every $\alpha > 0$ and every probability distribution p on $\mathbb{N}$, that

$$\mathcal{N}(\alpha, \mathcal{F}, d_{l^1(p)}) < \infty.$$

Ramon van Handel. "The universal Glivenko-Cantelli property", 2013.

- If we specialize this result to $\mathbb{N}$ with its usual σ-field, we see that every countable collection $\mathcal{F}$ of $[0,1]$-valued functions on $\mathbb{N}$ is universal Glivenko-Cantelli.

Three questions

- $\mathcal{F}$ a countable collection of $[0,1]$-valued functions on $\mathbb{N}$.

Question 1

What is a metric entropy characterization of when $\mathcal{F}$ is a Glivenko-Cantelli class in the data-derived weak sense?

- Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$.

Question 2

What is a metric entropy characterization of when $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense?

Question 3

What is a metric entropy characterization of when $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense?

Combinatorial characterization of uniform Glivenko-Cantelli classes

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $(\mathbb{X}, \mathcal{X})$ (but specialize to $\mathbb{N}$ for our purposes).
- Let $A \subseteq \mathbb{X}$ be a finite subset, and let $\gamma > 0$. We say that $\mathcal{F}$ shatters A at scale γ if there are real numbers $s(x), x \in A$ such that, for every subset $B \subset A$ there is a function $f \in \mathcal{F}$ such that $f(x) \geq s(x) + \gamma$ for $x \in B$ and $f(x) \leq s(x) - \gamma$ for $x \in A - B$.
- The γ-fat shattering dimension of $\mathcal{F}$ is the supremum over cardinalities of finite sets that are shattered by $\mathcal{F}$ at scale γ.

Theorem

$\mathcal{F}$ is a uniform Glivenko-Cantelli class iff its γ-fat shattering dimension is finite for all $\gamma > 0$.

N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler.

"Scale-sensitive dimensions, uniform covergence and learnability", 1997.

Special case: Shattering by sets

- Let $\mathcal{C}$ be a countable collection of subsets of $(\mathbb{X}, \mathcal{X})$ (but specialize to $\mathbb{N}$ for our purposes).
  Note that we can interpret $\mathcal{C}$ as a special case of $\mathcal{F}$ comprised of $\{0,1\}$-valued as opposed to $[0,1]$-valued functions.
- A finite set $A \subseteq \mathbb{X}$ is said to be shattered by $\mathcal{C}$ if for every $B \subseteq A$ there is a set $C \in \mathcal{C}$ such that $C \cap A = B$.
- The Vapnik-Chervonenkis dimension of $\mathcal{C}$ is the supremum of the cardinalities of finite subsets of $\mathbb{X}$ that are shattered by $\mathcal{C}$.

Theorem

$\mathcal{C}$ is a uniform Glivenko-Cantelli class iff its VC dimension is finite.

Vapnik and Chervonenkis, 1971; Assouad and Dudley, 1989; Dudley, Giné and Zinn, 1991.

Three more questions

- Let $\mathcal{F}$ be a countable collection of $[0,1]$-valued functions on $\mathbb{N}$ (in particular, consider $\mathcal{C}$).

Question 4

What is a combinatorial characterization of when $\mathcal{F}$ is a Glivenko-Cantelli class in the data-derived weak sense?

- Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$.

Question 5

What is a combinatorial characterization of when $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the strong sense?

Question 6

What is a combinatorial characterization of when $\mathcal{P}$ admits $\mathcal{F}$ as a Glivenko-Cantelli class in the data-derived weak sense?

The End

Thank you!