PAC learning, No-Free-Lunch theorem, uniform convergence, advanced Chernoff, application to missing mass estimation

Master 2 Mathématiques Avancées

Aurélien Garivier

2024-2025

Table of contents

PAC learning

PAC learnability: "probably approximately correct"

Definition

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_n$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D_X on $\mathcal{X}$ and for every labelling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, D_X, f$ then when $S = \big((X_1, f(X_1)), \dots, (X_n, f(X_n)\big)$ with $(X_i)_{1 \le i \le n} \overset{iid}{\sim} D_X$,

$$\mathbb{P}\Big(L_{(D_X,f)}\big(\hat{h}_n\big) \ge \epsilon\Big) \le \delta$$

for all $n \ge n_{\mathcal{H}}(\epsilon, \delta)$.

The smallest possible function $n_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

Remark: Valiant's PAC requires also sample complexity and running time polynomial in $1/\epsilon$ and $1/\delta$.

Finite hypothese classes are PAC-learnable

The sample complexity of finite hypothese classes in the realizable case is smaller than $\frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$:

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\epsilon, \delta \in (0,1)$ and let m be an integer that satisfies

$$n \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon} .$$

Then, for any labeling function f and for any distribution D_X on $\mathcal{X}$, under the realizability assumption, with probability at least $1 - \delta$ over the choice of iid sample S of size m, any ERM hypothesis $\hat{h}_n$ is such that

$$L_{(D_X, f)}(\hat{h}_n) \leq \epsilon .$$

Proof

The realizability assumption implies that an ERM $\hat{h}_S$ has empirical risk $L_S\big(\hat{h}_S\big) = 0$. Hence,

$$
\begin{aligned}
\mathbb{P}\left(L\big(\hat{h}_S\big) \geq \epsilon\right) &= D_X^{\otimes n}\Big(\big\{S \in \mathcal{X}^n : \exists h \in \mathcal{H}, L_S(h) = 0 \text{ and } L_D(h) \geq \epsilon\big\}\Big) \\
&= D_X^{\otimes n}\left(\bigcup_{h: L_D(h) \geq \epsilon} S_h\right) \quad \text{where } S_h = \big\{S \in \mathcal{X}^n : L_s(h) = 0\big\} \\
&\leq \sum_{h: L_D(h) \geq \epsilon} D_X^{\otimes n}\big(S_h\big) \\
&= \sum_{h: L_D(h) \geq \epsilon} \prod_{i=1}^{n} \underbrace{D_X\big(\{x \in \mathcal{X} : h(x) = f(x)\}\big)}_{=1 - L_D(h) \leq 1 - \epsilon} \\
&\leq \sum_{h: L_{(D_X, f)}(h) \geq \epsilon} \prod_{i=1}^{n} (1 - \epsilon) \leq \big|\mathcal{H}\big|(1 - \epsilon)^n \leq \big|\mathcal{H}\big| \exp(-n\epsilon) \;.
\end{aligned}
$$

This quantity is smaller than δ for $n \geq \dfrac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$.

Agnostic PAC learnability

Definition

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_n$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D on $\mathcal{X} \times \mathcal{Y}$ when $S = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \overset{iid}{\sim} D$,

$$\mathbb{P}\Big(L_D(\hat{h}_n) \geq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon\Big) \leq \delta$$

for all $m \geq n_{\mathcal{H}}(\epsilon, \delta)$.
The smallest possible function $n_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

If the realizable assumption holds, boils down to PAC learnability.
Otherwise, recall that the best Bayes classifier has a risk not larger than $\min_{h' \in \mathcal{H}} L_D(h')$.

Learning via uniform convergence

Definition

A training set S is called ϵ-representative (wrt domain $\mathcal{X} \times \mathcal{Y}$, hypothese class $\mathcal{H}$, loss function l and distribution D) if

$$\forall h \in \mathcal{H}, \left| L_S(h) - L_D(h) \right| \leq \epsilon \, .$$

Lemma

If S is $\epsilon/2$-representative, then any ERM $\hat{h}_n$ defined by $\hat{h}_n \in \arg\min_{h \in \mathcal{H}} L_S(h)$ satisfies:

$$L_D\left(\hat{h}_n\right) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon \, .$$

Proof: for every $h \in \mathcal{H}$,

$$L_D\left(\hat{h}_n\right) \leq L_S\left(\hat{h}_n\right) + \frac{\epsilon}{2} \leq L_S\left(h\right) + \frac{\epsilon}{2} \leq L_D\left(h\right) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \, .$$

Uniform Convergence Property

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (wrt $\mathcal{X} \times \mathcal{Y}$ and l) if there exists a function $n_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every distribution D over $\mathcal{X} \times \mathcal{Y}$, a sample $S = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \overset{iid}{\sim} D$ of size $m \geq n_{\mathcal{H}}^{UC}(\epsilon, \delta)$ has probability at least $1 - \delta$ to be ϵ-representative.

Corollary

If $\mathcal{H}$ has the uniform convergence property with a function $n_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostically PAC learnable with a sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq n_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$. Furthermore, the ERM is a successful PAC learner for $\mathcal{H}$.

Finite classes are agnostically PAC-learnable

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2} \right\rceil .$$

Moreover, $\mathcal{H}$ is agnostically PAC learnable using an ERM algorithm with sample complexity

$$n_{\mathcal{H}}(\epsilon, \delta) \leq 2 n_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2 \log \frac{2|\mathcal{H}|}{\delta}}{\epsilon^2} \right\rceil .$$

Proof: Hoeffding's inequality and the union bound.

No-Free-Lunch theorems: when learning is not possible

The No-Free-Lunch theorem

Theorem

Let A be any learning algorithm for binary classification over a domain $\mathcal{X}$. If the training set size is $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

- there exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_D(f) = 0$;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes n}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8} \, .$$

Note that the ERM over $\mathcal{H} = \{f\}$, or over any set $\mathcal{H}$ such that $m \geq 8 \log(7|\mathcal{H}|/6)$, is a successful learner in that setting.

Proof

Let $n \in \mathbb{N}$ and $A : (\mathcal{X} \times \{0,1\})^n \to \{0,1\}^{\mathcal{X}}$ be a learning algorithm. By assumption, there exists $C \subset \mathcal{X}$ of size $|C| \geq 2n$. Let $\mathcal{F} = \{0,1\}^C$, and for every $f \in \mathcal{F}$ let $D_f \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})$ be defined by: $D_f(\{x,y\}) = \begin{cases} \frac{1}{2n} \text{ if } y = f(x)\ , \\ 0 \text{ otherwise.} \end{cases}$

For every $f \in \mathcal{F}$, the marginal distribution of X under D_f is $\mathcal{U}(C)$, and the conditional distribution of Y given X is $\delta_{f(X)}$. Hence, $(X_1, \ldots, X_n) \sim \mathcal{U}(C^n)$. For every $s_X \in C^n$, define $s_X^f = (x, f(x))_{x \in s_X}$.

We will prove that $\max_{f \in \mathcal{F}} \mathbb{E}_{S \sim D_f^{\otimes n}} \left[L_{D_f}(A(S)) \right] \geq \dfrac{1}{4}$, which is sufficient: if $P(0 \leq Z \leq 1) = 1$ and $\mathbb{E}[Z] \geq \frac{1}{4}$, then $\mathbb{P}\left(Z \geq \frac{1}{8}\right) \geq \frac{1}{7}$ as $\frac{1}{4} \leq \mathbb{E}[Z] \leq \frac{1}{8}\mathbb{P}\left(Z < \frac{1}{8}\right) + \mathbb{P}(Z \geq 1/8) = \frac{1}{8} + \frac{7}{8}\,\mathbb{P}\left(Z \geq \frac{1}{8}\right)$.

$$\max_{f \in \mathcal{F}} \mathbb{E}_{S \sim D_f^{\otimes n}} \left[L_{D_f}(A(S)) \right] \geq \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{E}_{S \sim D_f^{\otimes n}} \left[L_{D_f}(A(S)) \right] = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{1}{|C^n|} \sum_{s_X \in C^n} L_{D_f}(A(s_X^f))$$

$$= \frac{1}{|C^n|} \sum_{s_X \in C^n} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} L_{D_f}(A(s_X^f)) \geq \min_{s_X \in C^n} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} L_{D_f}(A(s_X^f))\ .$$

For every $s_X \in \mathbb{C}^n$, observe that

$$\frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} L_{D_f}(A(s_X^f)) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{1}{|C|} \sum_{x \in C} \mathbb{1}\{A(s_X^f)(x) \neq f(x)\}$$

$$= \frac{1}{|C|} \sum_{x \in C} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{1}\{A(s_X^f)(x) \neq f(x)\} \geq \frac{1}{|C|} \sum_{x \in C \setminus s_X} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{1}\{A(s_X^f)(x) \neq f(x)\}$$

For $x \in C \setminus s_X$ and $y \in \{0,1\}$, let $\mathcal{F}_x^y = \{f \in \mathcal{F} : f(x) = y\}$; for $f \in \mathcal{F}_x^0$ let $\tilde{f}_x \in \mathcal{F}_x^1$ be s.t. $\forall x' \neq x,\ \tilde{f}_x(x') = f(x')$.

$$\sum_{f \in \mathcal{F}} \mathbb{1}\{A(s_X^f)(x) \neq f(x)\} = \sum_{y \in \{0,1\}} \sum_{f \in \mathcal{F}_x^y} \mathbb{1}\{A(s_X^f)(x) \neq f(x)\} = \sum_{f \in \mathcal{F}_x^0} \mathbb{1}\{A(s_X^f)(x) \neq \underbrace{f(x)}_{=0}\} + \mathbb{1}\{A(s_X^{\tilde{f}_x})(x) \neq \underbrace{\tilde{f}_x(x)}_{=1}\} = \frac{|\mathcal{F}|}{2}$$

since, as $x \notin s_X$, $s_X^{\tilde{f}_x} = s_X^f$ and hence $A(s_X^{\tilde{f}_x}) = A(s_X^f)$. The conclusion comes, as $|C \setminus s_X| \geq |C|/2$.

Consequence: Curse of Dimensionality

Theorem

Let $c > 1$ be a Lipschitz constant. Let A be any learning algorithm for binary classification over a domain $\mathcal{X} = [0,1]^d$. If the training set size is $n \leq c^d/2$, then there exists a distribution $\mathcal{D}$ over $[0,1]^d \times \{0,1\}$ such that:

- $\eta(x)$ is c-Lipschitz;
- the Bayes error of the distribution is 0;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes n}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8} .$$

Uniform convergence for infinite classes: VC dimension

Shattering

Definition

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{c_1, \dots, c_n\} \subset \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of functions $C \to \{0,1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \Big\{(c_1, \dots, c_n) \to \big(h(c_1), \dots, h(c_n)\big) : h \in \mathcal{H}\Big\}\,.$$

Shattering

A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Example:

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \big\{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1 \text{ and } a_2 \leq b_2\big\}$ where

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2\,; \\ 0 & \text{otherwise}\,. \end{cases}$$

VC dimension

Definition

The *Vapnik Chervonenkis dimension* $\text{VCdim}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\text{VCdim}(\mathcal{H}) = \infty$.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then $\mathcal{H}$ is not PAC-learnable.

Proof: for every training size m, there exists a set C of size $2m$ that is shattered by $\mathcal{H}$. By the NFL theorem, for every learning algorithm A there exists a probability distribution D over $\mathcal{X} \times \{0, 1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.

Outline

PAC learning

No-Free-Lunch theorems: when learning is not possible

Uniform convergence for infinite classes: VC dimension

- VC dimension and Sauer's lemma
- Finite VC dimension implies Uniform Convergence
- Finite VC-dimension implies learnability

More on Chernoff's method

- Example: Poisson distribution
- Chernoff's bound for real-valued variables

Application: estimating the missing mass

Shattering

Definition

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{c_1, \ldots, c_n\} \subset \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of functions $C \to \{0,1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \Big\{(c_1, \ldots, c_n) \to \big(h(c_1), \ldots, h(c_n)\big) : h \in \mathcal{H}\Big\}.$$

Shattering

A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Example:

- $\mathcal{H} = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \{\mathbb{1}_{[a_1,b_1]\times[a_2,b_2]} : a_1 \leq b_1 \text{ and } a_2 \leq b_2\}$.

VC dimension

Definition

The *Vapnik Chervonenkis dimension* $\text{VCdim}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\text{VCdim}(\mathcal{H}) = \infty$.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then $\mathcal{H}$ is not PAC-learnable.

Proof: for every training size n, there exists a set C of size $2n$ that is shattered by $\mathcal{H}$. By the NFL theorem, for every learning algorithm A there exists a probability distribution D over $\mathcal{X} \times \{0, 1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.

Fundamental theorem of PAC learning

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Then the following propositions are equivalent:

1. $\mathcal{H}$ has the uniform convergence property,
2. any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$,
3. $\mathcal{H}$ is agnostic PAC learnable,
4. $\mathcal{H}$ is PAC learnable,
5. any ERM rule is a sucessful PAC learner for $\mathcal{H}$,
6. $\mathcal{H}$ has finite VC-dimension.

Fundamental theorem of PAC learning (quantitative version)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Assume thatVCdim$(\mathcal{H}) < \infty$. Then there exist constants C_1, C_2 such that:

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

3. $\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon} \ .$$

Sauer's lemma

Definition

Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as the maximal number of different functions that can be obtained by restricting $\mathcal{H}$ to a set of size m:

$$\tau_{\mathcal{H}}(n) = \max_{C \subset X : |C| = n} \left| \mathcal{H}_C \right| .$$

Note: if $\mathrm{VCdim}(\mathcal{H}) = d$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^n$.

Sauer's lemma

Let $\mathcal{H}$ be a hypothesis class with $d = \mathrm{VCdim}(\mathcal{H}) < \infty$. Then, for all $n \geq d$,

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d} \right)^d .$$

Think of example: $\mathcal{H} = \left\{ \mathbb{1}_{(-\infty, a]} : a \in \mathbb{R} \right\}$ with $d = \mathrm{VCdim}(\mathcal{H}) = 1$.

Proof of Sauer's lemma 1/2

In fact we prove the stronger claim:

$$|\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{n}{i} .$$

where the last inequality holds since no set of size larger than d is shattered by $\mathcal{H}$. The proof is by induction.

n=1: The empty set is always considered to be shattered by $\mathcal{H}$. Hence, either $|\mathcal{H}_C| = 1$ and $d = 0$, inequality $1 \leq 1$, or $d \geq 1$ and the inequality is $2 \leq 2$.

Induction: Let $C = \{c_1, \dots, c_n\}$, and let $C' = \{c_2, \dots, c_n\}$. We note functions like vectors, and we define

$$Y_0 = \Big\{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C \text{ or } (1, y_2, \dots, y_n) \in \mathcal{H}_C\Big\}, \quad \text{and}$$

$$Y_1 = \Big\{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C \text{ and } (1, y_2, \dots, y_n) \in \mathcal{H}_C\Big\} .$$

Then $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Moreover, $Y_0 = \mathcal{H}_{C'}$ and hence by the induction hypothesis:

$$|Y_0| \leq |\mathcal{H}_{C'}| \leq |\{B \subset C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}|$$

Next, define

$$\mathcal{H}' = \left\{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } \forall 1 \leq i \leq n, h'(c_i) = \begin{cases} 1 - h(c_1) \text{ if } i = 1 \\ h(c_i) \text{ otherwise} \end{cases}\right\}$$

Note that $\mathcal{H}'$ shatters $B \subset C'$ iff $\mathcal{H}'$ shatters $B \cup \{c_1\}$, and that $Y_1 = \mathcal{H}'_{C'}$. Hence, by the induction hypothesis,

$$\begin{aligned} |Y_1| = |\mathcal{H}'_{C'}| &\leq |\{B \subset C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subset C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ &= |\{B \subset C : c_1 \in B \text{ and } \mathcal{H}' \text{ shatters } B\}| \leq |\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| . \end{aligned}$$

Overall,

$$|\mathcal{H}_C| = |Y_0| + |Y_1| \leq |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| + |\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| = |\{B \subset C : \mathcal{H} \text{ shatters } B\}| .$$

Proof of Sauer's lemma 2/2

For the last inequality, one may observe that if $n \geq 2d$, defining $N \sim \mathcal{B}(n, 1/2)$, Chernoff's inequality and inequality $\log(u) \geq (u-1)/u$ yield

$$\begin{aligned} -\log \mathbb{P}(N \leq d) \geq n \, \mathrm{kl}\left(\frac{d}{n}, \frac{1}{2}\right) &\geq d \log \frac{2d}{n} + (n-d) \log \frac{2(n-d)}{n} \\ &\geq n \log(2) + d \log \frac{d}{n} + (n-d) \frac{-d/n}{(n-d)/n} \\ &= n \log(2) + d \log \frac{d}{en} \ , \end{aligned}$$

and hence

$$\sum_{i=0}^{d} \binom{n}{i} = 2^n \mathbb{P}(N \leq d) \leq \exp\left(-d \log \frac{d}{en}\right) = \left(\frac{en}{d}\right)^d \ .$$

Besides, for the case $d \leq n \leq 2d$, the inequality is obvious since $(en/d)^d \geq 2^n$: indeed, function $f : x \mapsto -x \log(x/e)$ is increasing on $[0, 1]$, and hence for all $d \leq n \leq 2d$:

$$\frac{d}{n} \log \frac{en}{d} = f(d/n) \geq f(1/2) = \frac{1}{2} \log(2e) \geq \log(2) \ ,$$

which implies

$$\left(\frac{en}{d}\right)^d = \exp\left(d \log \frac{en}{d}\right) \geq \exp(n \log(2)) = 2^n \ .$$

Alternately, you may simply observe that for all $n \geq d$,

$$\left(\frac{d}{n}\right)^d \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{n}\right)^i \binom{n}{i} \leq \sum_{i=0}^{n} \left(\frac{d}{n}\right)^i \binom{n}{i} = \left(1 + \frac{d}{n}\right)^n \leq e^d \ .$$

Outline

Finite VC dimension implies Uniform Convergence

Theorem

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every distribution D dans for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of the sample $S \sim D^{\otimes n}$ we have

$$\sup_{h \in \mathcal{H}} \left| L_D(h) - L_S(h) \right| \leq \frac{1 + \sqrt{\log\left(2\tau_{\mathcal{H}}(2n)\right)}}{\delta\sqrt{n/2}} .$$

Note: this result is sufficient to prove that finite VC-dim $\implies$ learnable, but the dependency in δ is not correct at all: roughly speaking, the factor $1/\delta$ can be replaced by $\log(1/\delta)$.

Proof: symmetrization and Rademacher complexity (1/2)

We consider the 0-1 loss $\ell(h,(x,y)) = \mathbb{1}\{h(x) \neq y\}$, or any $[0,1]-$valued loss ℓ. We denote $Z_i = (X_i, Y_i)$, and observe that $L_D(h) = \mathbb{E}_{Z_i}[\ell(h, Z_i)] = \mathbb{E}_{S'}[L_{S'}(h)]$ if $S' = Z'_1, \dots, Z'_n$ denotes another iid sample of D. Hence,

$$
\begin{aligned}
\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|L_D(h) - L_S(h)\right|\right] &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'}[L_{S'}(h)] - L_S(h)\right|\right] = \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'}\left[L_{S'}(h) - L_S(h)\right]\right|\right] \\
&\leq \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}\left[\left|L_{S'}(h) - L_S(h)\right|\right]\right] \leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\right]\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \ell(h, Z'_i) - \ell(h, Z_i)\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\ell(h, Z'_i) - \ell(h, Z_i)\big)\right|\right] \quad \text{for all } \sigma \in \{\pm 1\}^n \\
&= \mathbb{E}_\Sigma\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\big(\ell(h, Z'_i) - \ell(h, Z_i)\big)\right|\right] \quad \text{if } \Sigma \sim \mathcal{U}\big(\{\pm 1\}^n\big) \\
&= \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\big(\ell(h, Z'_i) - \ell(h, Z_i)\big)\right|\right] .
\end{aligned}
$$

Now, for every fixed $S = \big\{(X_i, Y_i) : 1 \leq i \leq n\big\}$ and $S' = \big\{(X_i, Y_i) : 1 \leq i \leq n\big\}$, the number of different $\big(\ell(h, Z'_i) - \ell(h, Z_i)\big) \in [-1, 1]$ is bounded by $\tau_{\mathcal{H}}(2n)$. Indeed, let $C = C_{S,S'} = \big\{X_1, \dots, X_n\big\} \cup \{X'_1, \dots, X'_n\}$. Then $\forall \sigma \in \{-1, 1\}^n$,

$$
\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\ell(h, Z'_i) - \ell(h, Z_i)\big)\right| = \max_{h\in\mathcal{H}_C}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\ell(h, Z'_i) - \ell(h, Z_i)\big)\right| .
$$

Proof: symmetrization and Rademacher complexity (2/2)

Moreover, for every $h \in \mathcal{H}_C$ let $Z_h = \frac{1}{n}\sum_{i=1}^n \Sigma_i\big(\ell(h, Z_i') - \ell(h, Z_i)\big)$. Then $\mathbb{E}_\Sigma[Z_h] = 0$, each summand belongs to $[-1, 1]$ and by Hoeffding's inequality, for every $\epsilon > 0$:

$$\mathbb{P}_\Sigma\big[|Z_h| \geq \epsilon\big] \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right) .$$

Hence, by the union bound,

$$\mathbb{P}_\Sigma\big[\max_{h \in \mathcal{H}_C} |Z_h| \geq \epsilon\big] \leq 2\big|\mathcal{H}_C\big| \exp\left(-\frac{n\epsilon^2}{2}\right) .$$

The following lemma permits to deduce that

$$\mathbb{E}_\Sigma\left[\max_{h \in \mathcal{H}_C} |Z_h|\right] \leq \frac{1 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{n/2}} \leq \frac{1 + \sqrt{\log(2\tau_{\mathcal{H}}(2n))}}{\sqrt{n/2}} .$$

Hence,

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \big|L_D(h) - L_S(h)\big|\right] \leq \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \Sigma_i\big(\ell(h, Z_i') - \ell(h, Z_i)\big)\right|\right] \leq \frac{1 + \sqrt{\log(2\tau_{\mathcal{H}}(2n))}}{\sqrt{n/2}} ,$$

and we conclude by using Markov's inequality (poor idea! Better: McDiarmid's inequality).

Technical Lemma

Lemma

Let $a > 0$, $b \geq 1$, and let Z be a real-valued random variable such that for all $t \geq 0$, $\mathbb{P}(Z \geq t) \leq 2b \exp\left(-\frac{t^2}{a^2}\right)$. Then

$$\mathbb{E}[Z] \leq a\left(\sqrt{\log(2b)} + 1\right) .$$

Proof:

$$\begin{aligned}
\mathbb{E}[Z] &\leq \int_0^\infty \mathbb{P}(Z \geq t)dt \leq a\sqrt{\log(2b)} + \int_{a\sqrt{\log(2b)}}^\infty 2b \exp\left(-\frac{t^2}{a^2}\right) dt \\
&\leq a\sqrt{\log(2b)} + 2b \int_{a\sqrt{\log(2b)}}^\infty \frac{t}{a\sqrt{\log(2b)}} \exp\left(-\frac{t^2}{a^2}\right) dt \\
&= a\sqrt{\log(2b)} + \frac{2b}{a\sqrt{\log(2b)}} \times \frac{a^2}{2} \exp\left(-\frac{\left(a\sqrt{\log(2b)}\right)^2}{a^2}\right) \\
&= a\sqrt{\log(2b)} + \frac{a}{2\sqrt{\log(2b)}}
\end{aligned}$$

and for all $b \geq 1$, $2\sqrt{\log(2b)} \geq 2\sqrt{\log(2)} > 1$.

Outline

Application: Finite VC-dim classes are agnostically learnable

It suffices to prove that finite VC-dim implies the uniform convergence property. From Sauer's lemma, for all $n \geq d/2$ we have $\tau_{\mathcal{H}}(2n) \leq (2en/d)^d$. With the previous theorem, this yields that with probability at least $1-\delta$:

$$\sup_{h\in\mathcal{H}} \left|L_D(h) - L_S(h)\right| \leq \frac{1+\sqrt{d\log\left(4en/d\right)}}{\delta\sqrt{n/2}} \leq \frac{1}{\delta}\sqrt{\frac{8d\log(4en/d)}{n}}$$

as soon as $\sqrt{d\log\left(4en/d\right)} \geq 1$. To ensure that this is at most ϵ, one may choose

$$n \geq \frac{8d\log(n)}{(\delta\epsilon)^2} + \frac{8d\log(4e/d)}{(\delta\epsilon)^2} \ .$$

By the following lemma, it is sufficient that

$$n \geq \frac{32d\log\left(\frac{4d}{(\delta\epsilon)^2}\right)}{(\delta\epsilon)^2} + \frac{16d\log\left(\frac{4e}{d}\right)}{(\delta\epsilon)^2} \ .$$

Technical Lemma

Lemma

Let $a > 0$. Then

$$x \geq 2a\log(a) \quad \Longrightarrow \quad x \geq a\log(x)\ .$$

Proof: For $a \leq e$, true for every $x > 0$. Otherwise, for $a \geq \sqrt{e}$ we have $2a\log(a) \geq a$ and thus for every $t \geq 2a\log(a)$, as $f : t \mapsto t - a\log(t)$ is increasing on $[a, \infty)$, $f(t) \geq f(2a\log(a)) = a\log(a) - a\log(2\log(a)) \geq 0$, since for every $a > 0$ it holds that $a \geq 2\log(a)$.

Lemma

Let $a \geq 1, b > 0$. Then

$$x \geq 4a\log(2a) + 2b \quad \Longrightarrow \quad x \geq a\log(x) + b\ .$$

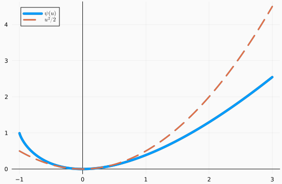
Proof: It suffices to check that $x \geq 2a\log(x)$ (given by the above lemma) and that $x \geq 2b$ (obvious since $4a\log(2a) \geq 0$).

More on Chernoff’s method

Outline

Chernoff's method for the Poisson distribution

Let $\mu > 0$ and let $X \sim \mathcal{P}(\mu)$. Then



$$\forall x \geq \mu, \quad \mathbb{P}(X \geq x) \leq \exp\left(-\left(x \ln \frac{x}{\mu} - (x-\mu)\right)\right)$$

$$\forall \epsilon \geq 0, \quad \mathbb{P}(X \geq \mu + \epsilon) \leq \exp\left(-\mu\varphi^*\left(\frac{\epsilon}{\mu}\right)\right)$$

and $\mathbb{P}(X \leq \mu - \epsilon) \leq \exp\left(-\mu\varphi^*\left(\frac{\epsilon}{\mu}\right)\right) \leq \exp\left(-\frac{\epsilon^2}{2\mu}\right)$,

where $\varphi^*(u) = (1+u)\ln(1+u) - u = \dfrac{u^2}{2}\displaystyle\int_0^1 \frac{1}{1+tu}\, 2(1-t)dt.$

Observe that $\mathrm{KL}(\mathcal{P}(x), \mathcal{P}(\mu)) = x \ln \frac{x}{\mu} - (x - \mu)$.

Left tail: For $u < 0$, $\varphi^*(u) \geq \frac{u^2}{2}$ and

$$\mathbb{P}\left(X \leq \mu - \sqrt{2\mu \ln \frac{1}{\delta}}\right) \leq \delta\,.$$

Proof

The two bounds are equivalent, for $x = \mu + \epsilon$. The first one is obtained by remarking that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda X}\right] = \sum_{k=0}^{\infty} e^{-\mu} \frac{\mu^k}{k} e^{\lambda k} = e^{-\mu} e^{\mu e^{\lambda}}$$

and hence

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[\left[e^{\lambda X}\right]}{e^{\lambda x}} = \exp\left(-\left(\lambda x - \mu\left(e^{\lambda} - 1\right)\right)\right) ,$$

which yields the result for $\lambda = \ln \frac{x}{\mu}$.

Alternatively, $\ln \mathbb{E}\left[e^{\lambda(X-\mu)}\right] = \mu\varphi(\lambda)$ with $\varphi(\lambda) \stackrel{\triangle}{=} e^{\lambda} - \lambda - 1$, and

$$\sup_{\lambda > 0} \lambda\epsilon - \mu\varphi(\lambda) = \mu\varphi^*\left(\frac{\epsilon}{\mu}\right)$$

with $\varphi^*(u) = (1+u)\ln(1+u) - u = \sup_{\lambda>0} \lambda u - \varphi(\lambda)$.

Poisson right tails

$\varphi(\lambda) = e^{\lambda} - \lambda - 1$	Taylor	$\varphi_1(\lambda) = \frac{\lambda^2}{2\left(1-\frac{\lambda}{3}\right)} = 9h\left(\frac{\lambda}{3}\right)$		$\varphi_2(\lambda) = 4h^*\left(-\frac{\lambda}{2}\right)$	use φ_1	$h(\lambda) = \frac{\lambda^2}{2(1-\lambda)}$
$\downarrow$ compute		$\downarrow$ use h^*				$\uparrow$ compute$(h^*)^*$
$\varphi^*(u) = (1+u)\ln(1+u) - u$		$\varphi_1^*(u) = 9h^*\left(\frac{u}{3}\right)$	$\ln(1+u) \geq \frac{u}{1+\frac{u}{2}}$	$\varphi_2^*(u) = \frac{u^2}{2+u} = 4h\left(-\frac{u}{2}\right)$	differentiate twice $\varphi^* - h^* \geq 0$	$h^*(u) = 1 + u - \sqrt{1+2u}$
		$\downarrow$ use $(h^*)^{-1}$				$\downarrow$ $h^*(\sqrt{2x}+x) = x$
$(\varphi^*)^{-1}(x)$		$(\varphi_1^*)^{-1}(x) = \sqrt{2x} + \frac{x}{3}$		$(\varphi_2^*)^{-1}(x)$	$\varphi_2^*(\sqrt{2x}+x) \geq x$	$(h^*)^{-1}(x)$ $\sqrt{2x} + x$

Optimal path:

1. $(h^*)^* = h$ and hence the formula for h^*
2. $h^*(x + \sqrt{2x}) = x$ and hence the formula for $(h^*)^{-1}$
3. $\varphi(\lambda) \leq \varphi_1(\lambda)$ since $e^{\lambda} - \lambda - 1 = \frac{\lambda^2}{2}\left(1 + \frac{\lambda}{3} + \frac{\lambda^2}{3\times 4} + \ldots\right) \leq \frac{\lambda^2}{2}\left(1 + \frac{\lambda}{3} + \frac{\lambda^2}{3^2} + \ldots\right) = \frac{\lambda^2/2}{1-\lambda/3}$
4. recognize $\varphi_1(\lambda) = 9h(\lambda/3)$ and hence $\varphi_1^*(u) = 9h^*(u/3)$
5. hence $(\varphi_1^*)^{-1}(x) = 3(h^*)^{-1}(x/9) = \sqrt{2x} + \frac{x}{3}$

hence

$$\mathbb{P}\left(X \geq \mu + \sqrt{2\mu \ln \frac{1}{\delta}} + \frac{\ln \frac{1}{\delta}}{3}\right) \leq \delta\ .$$

Simpler but weaker bound (two proofs):

- For $u > 0$, $\varphi^*(u) \geq h^*(u) \triangleq 1 + u - \sqrt{1+2u}$ and $(\varphi^*)^{-1}(x) \leq (h^*)^{-1}(x) = \sqrt{2x} + x$
- (personal) for $u > 0$, $\ln(1+u) \geq \frac{u}{1+u/2}$ and $\varphi^*(u) \geq \frac{(1+u)u - u(1+u/2)}{1+u/2} = \frac{u^2}{2+u} \triangleq \varphi_2^*(u)$. Since $\varphi_2^*(\sqrt{2x}+x) = x\frac{2+x+2\sqrt{2x}}{2+x+\sqrt{2x}} \geq x$, $(\varphi^*)^{-1}(x) \leq (\varphi_2^*)^{-1}(x) \leq \sqrt{2x} + x$.

Outline

The Log-Laplace function

Let X be a real-valued random variable with law P_X, expectation μ_X and variance σ_X^2. $\lambda \mapsto \mathbb{E}\left[e^{\lambda X}\right]$ is finite on an interval $(\underline{\lambda}, \overline{\lambda})$ and we assume that it is non-empty, ie $0 \in (\underline{\lambda}, \overline{\lambda})$. Chernoff's bound states that for $x \geq \mu$,

$$\mathbb{P}(X \geq x) \leq \exp\left(-\sup_{\lambda>0} \lambda x - \varphi_X(\lambda)\right) ,$$

where $\varphi_X(\lambda) \stackrel{\triangle}{=} \ln \mathbb{E}\left[e^{\lambda X}\right]$. For $\lambda \in (\underline{\lambda}, \overline{\lambda})$,

- $\varphi_X'(\lambda) = \mu_X(\lambda) \stackrel{\triangle}{=} \mathbb{E}^\lambda[X]$, where $\dfrac{d\mathbb{P}_X^\lambda}{d\mathbb{P}_X}(x) = \dfrac{e^\lambda x}{\mathbb{E}\left[e^{\lambda X}\right]}$.
- $\varphi_X''(\lambda) = \mu_X'(\lambda) = \sigma_X^2(\lambda) \stackrel{\triangle}{=} \mathbb{V}ar(P_X^\lambda) \geq 0$,
- $\mu_X : (\underline{\lambda}, \overline{\lambda}) \to (\underline{x}, \overline{x}) \subset \mathrm{Supp}(X)$ is increasing and $\mathcal{C}^\infty$, with $\mu_X(0) = \mu_X$.
- $\varphi_X(\lambda) = \displaystyle\int_0^\lambda \sigma_X^2(\ell)(\lambda - \ell)\, d\ell = \frac{\lambda^2}{2} \int_0^1 \sigma_X^2(\lambda t)\, 2(1-t)\, dt$.

Its Fenchel-Legendre transform

For all $x \in (\mu, \overline{x})$, since φ_X is smooth and convex $\varphi_X^\star(x) = \sup_{\lambda > 0} \lambda x - \varphi_X(\lambda) = \lambda_X(x)x - \varphi_X\big(\lambda_X(x)\big)$, where $\lambda_X(x) = \mu_X^{-1}(x)$.

- $\varphi_X^\star{}'(x) = \lambda(x) + x\lambda_X'(x) - \lambda_X'(x)\varphi_X'\big(\lambda(x)\big) = \lambda_X(x) = \mu_X^{-1}(x)$.
- $\varphi_X^\star{}''(x) = \frac{1}{\mu_X'\big(\lambda_X(x)\big)} = \frac{1}{\sigma_X^2\big(\mu_X^{-1}(x)\big)}$.

$$\varphi^\star{}_X(x) = \int_\mu^x \frac{x - u}{\sigma_X^2\big(\mu_X^{-1}(u)\big)}\,du = \frac{(x-\mu)^2}{2}\int_0^1 \frac{2(1-t)\,dt}{\sigma_X^2\Big(\mu_X^{-1}\big(\mu + t(x-\mu)\big)\Big)} \,.$$

Example

If $\mathbb{P}(0 \leq X \leq 1)$, $\varphi_X(\lambda) \leq \varphi_\mu(\lambda) = \ln\left(1 - \mu + \mu e^\lambda\right)$ with equality iff $X \sim \mathcal{B}(\mu)$, since by convexity of $u \mapsto e^{\lambda u}$, $\forall x \in [0,1]$, $e^{\lambda x} \leq (1-x) + x e^\lambda$.

Since for all λ, $\mathbb{P}_X^\lambda([0,1]) = 1$, $\sigma_X^2(\lambda) \leq 1/4$ and $\varphi_X(\lambda) \leq \lambda\mu + \frac{\lambda^2}{8}$ and

- The upper-bound on φ_X yields a lower bound on $\varphi_X\star$:

$$\varphi_X^\star(\mu+\epsilon) = \sup_{\lambda>0} \lambda(\mu+\epsilon) - \varphi_X(\lambda) \geq \sup_{\lambda>0} \lambda\epsilon - \frac{\lambda^2}{8} = 2\epsilon^2$$

- The expression for $\varphi_X^\star$ permits to re-derive it directly:

$$\varphi_X^\star(\mu+\epsilon) = \frac{\epsilon^2}{2} \int_0^1 \frac{2(1-t)\,dt}{\sigma_X^2\left(\mu_X^{-1}(\mu+t\epsilon)\right)} \geq 2\epsilon^2\ .$$

Connection to KL divergence

Observe that for all $\lambda \in (\underline{\lambda}, \overline{\lambda})$, $\mathrm{KL}\left(P_X^\lambda, P_X\right) = \lambda \mu_X(\lambda) - \varphi_X(\lambda)$. Hence,

$$\mathrm{KL}\left(P_X^{\lambda_X(x)}, P_X\right) = \lambda_X(x) \underbrace{\mu_X(\lambda_X(x))}_{=x} - \varphi_X(\lambda_X(x)) = \varphi_X^\star(x) \,.$$

Besides $\mathrm{KL}\left(P_X^{\lambda_X(x)}, P_X\right) = \inf \left\{ \mathrm{KL}(Q, P_X) : \mathbb{E}_Q[X] \geq x \right\}$. Indeed, For every $Q \ll P$ with $\mathbb{E}_Q[X] \geq x$,

$$\begin{aligned}
\mathrm{KL}(Q, P_X) &= \int_{\mathbb{R}} \log\left(\frac{dQ}{dP_X}(x)\right) dQ(x) \\
&= \int_{\mathbb{R}} \log\left(\frac{dQ}{dP_X^{\lambda_X(x)}}(x) \frac{dP_X^{\lambda_X(x)}}{dP}(x)\right) dQ(x) \\
&= \mathrm{KL}(Q, P_X^{\lambda_X(x)}) + \int_{\mathbb{R}} \log\left(\frac{e^{\lambda_X(x) x}}{\mathbb{E}[e^{\lambda_X(x)} X]}\right) dQ(x) \\
&= \mathrm{KL}(Q, P_{\lambda_X(x)}) + \lambda_X(x) \mathbb{E}_Q[X] - \log\left(\mathbb{E}[e^{\lambda_X(x)} X]\right) \\
&\geq \quad 0 \quad + \lambda_X(x) x - \quad \varphi_X(\lambda_X(x)) \quad = \mathrm{KL}(P_X^{\lambda_X(x)}, P) \,.
\end{aligned}$$

Case of a sum of independent variables

If $X = X_1 + \cdots + X_n$ where the $(X_i)_i$ are independent, then $\varphi_X = \sum_i \varphi_{X_i}$, $\mu_X = \sum_i \mu_{X_i} \stackrel{\triangle}{=} n\bar{\mu}$ and $\sigma_X^2 = \sum_i \sigma_{X_i}^2$.

Besides,

$$\varphi_X^\star(nx) = \int_{\mu_X}^{nx} \frac{nx - u}{\sigma_X^2\left(\mu_X^{-1}(u)\right)}\, du$$
$$= n \int_{\bar{\mu}}^{x} \frac{x - v}{\sigma_X^2\left(\mu_X^{-1}(u)\right)}\, dv\ .$$

Bennett's inequality for Bernoullis: if $\forall i, \forall \lambda > 0, \sigma_{X_i}^2(\lambda) \leq \mu_{X_i}(\lambda)$ then

$$\sigma_X^2\left(\mu_X^{-1}(v)\right) = \sum_i \sigma_{X_i}^2\left(\mu_X^{-1}(v)\right) \leq \sum_i \mu_{X_i}\left(\mu_X^{-1}(v)\right) = \mu_X\left(\mu_X^{-1}(v)\right) = v$$

and

$$\varphi_X^\star(nx) \geq n \int_{\bar{\mu}}^{x} \frac{x - v}{v}\, dv = n\ \left(x \ln \frac{x}{\bar{\mu}} - (x - \bar{\mu})\right)$$

or

$$\varphi_X^\star\left(n(\bar{\mu} + \epsilon)\right) \geq \bar{\mu}\varphi^*\left(\frac{\epsilon}{\bar{\mu}}\right)$$

$\varphi^*(u) = (1+u)\ln(1+u) - u = \frac{u^2}{2}\int_0^1 \frac{2(1-t)\ dt}{1+ut} \geq \frac{u^2}{2}\frac{1}{1+u\int_0^1 2(1-t)\ dt} = \frac{u^2}{2\left(1+\frac{u}{3}\right)}$.

Bennett's inequality: the other way

Bennett's inequality for bounded variables

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{P}(X_i \leq 1) = M$ and $\mathbb{E}[X_i^2] \leq \sigma^2$. Then, if $\bar{\mu} = (\mathbb{E}[X_1] + \mathbb{E}[X_n])/n$,

$$\mathbb{P}(\bar{X}_n \geq \bar{\mu} + \epsilon) \leq \exp\left(-\frac{n\sigma^2}{M^2}\varphi^*\left(\frac{M\,\epsilon}{\sigma^2}\right)\right) \leq \exp\left(-\frac{n\,\epsilon^2}{2\left(\sigma^2 + \frac{M\epsilon}{3}\right)}\right) .$$

Since for $x > 0$, $(\varphi^*)^{-1}(x) \leq \sqrt{2x} + \frac{x}{3}$,

$$\mathbb{P}\left(X \geq \mu + \sqrt{\frac{2\sigma^2 \ln \frac{1}{\delta}}{n}} + \frac{M \ln \frac{1}{\delta}}{3n}\right) \leq \delta .$$

Proof

We first prove the result for $M = 1$. Since $\frac{e^u - u - 1}{u^2/2} = \int_0^1 e^{ut} 2(1-t)dt$ increases with u, for all $x \leq 1$ $e^{\lambda x} - \lambda x - 1 \leq x^2 \varphi(\lambda)$ with $\varphi(\lambda) = e^\lambda - \lambda - 1$ and since $\mathbb{E}[X_i^2] \leq \sigma^2$:

$$\ln \mathbb{E}\left[e^{\lambda\left(X_i - \mathbb{E}[X_i]\right)} \right] \leq \ln\left(1 + \lambda \mathbb{E}[X_i] + \sigma^2 \varphi(\lambda)\right) - \lambda \mathbb{E}[X_i] \leq \ \sigma^2 \varphi(\lambda) \ .$$

Consequently, if $X = X_1 + \cdots + X_n$ then for all $\epsilon > 0$

$$I_{X - \mathbb{E}[X]}(n\epsilon) = \sup_{\lambda > 0} \lambda\epsilon - \sigma^2 \varphi(\lambda) = \sigma^2 \sup_{\lambda > 0} \lambda \frac{\epsilon}{\sigma^2} - \varphi(\lambda) = \sigma^2 \varphi^* \left(\frac{\epsilon}{\sigma^2}\right)$$

and

$$\mathbb{P}(X \geq \mathbb{E}[X] + \epsilon) \leq \exp\left(-\sigma^2 \varphi^* \left(\frac{\epsilon}{\sigma^2}\right)\right) \leq \exp\left(-\frac{\epsilon^2}{2\left(\sigma^2 + \frac{\epsilon}{3}\right)}\right) \ .$$

Finally, if $M \neq 1$ apply the result to the $Y_i = X_i / M$ which have a variance bounded by σ^2 / M^2.

Bernstein's inequality can be more general

Theorem

If for all $k \geq 3, \mathbb{E}[X^k] \leq 1/2k!\sigma^2 b^{k-2}$, then for all $\lambda \in (0, 1/b)$:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-\lambda b)}\right) .$$

Hence, if $X = X_1 + \cdots + X_n$ where the (X_i) are independent and $\forall k \geq 3, \mathbb{E}[X_i^k] \leq 1/2k!\sigma_i^2 b^{k-2}$, then for every $x > 0$,

$$\mathbb{P}(X > x) \leq \exp\left(-\frac{x^2}{2\left(\sigma^2 + xb\right)}\right)$$

with $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Proof: choose $\lambda = x/(\sigma^2 + xb)$

Remark: Bennett's condition is stronger since it implies $\mathbb{E}[X^k] \leq \mathbb{E}[X^2 b^{k-2}] \leq \sigma^2 b^{k-2}$.

Appl: fast rates in binary classification under margin condition

- Binary classification on $\mathcal{X}$ with $\eta(x) = \mathbb{P}(Y = 1|X = x)$, finite hypothesis class $\mathcal{H}$
- Bayes classifier h^* with Bayes risk L_D^*, empirical risk minimizer $\hat{h}_n \in \arg\min_{h\in\mathcal{H}} L_S(h)$
- Excess risk: $\forall h \in \mathcal{H}, L_D(h) - L_D^* = \mathbb{E}\big[|2\eta(X) - 1|\mathbb{1}\{h(X) \neq h^*(X)\}\big]$
- Massart's margin condition: $|2\eta(X) - 1| \geq 2\gamma > 0$ almost surely
- Then $L_D(h) - L_D^* \geq 2\gamma\mathbb{P}\big(h(X) \neq h^*(X)\big)$
- But for all $h \in \mathcal{H}$, since $L_D(h) = \mathbb{E}\big[L_S(h)\big]$:

$$\begin{aligned} L_D(h) - L_D^* &= L_D(h) - L_S(h) + L_S(h) - L_S(h^*) + L_S(h^*) - L_D(h^*) \\ &= \underbrace{L_S(h) - L_S(h^*)}_{\leq 0 \text{ for } h = \hat{h}_n} + \underbrace{L_S(h^*) - L_S(h) - \mathbb{E}\big[L_S(h^*) - L_S(h)\big]}_{\triangleq \frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]} \end{aligned}$$

where $Z_i = \mathbb{1}\{h^*(X_i) \neq Y_i\} - \mathbb{1}\{h(X_i) \neq Y_i\}$

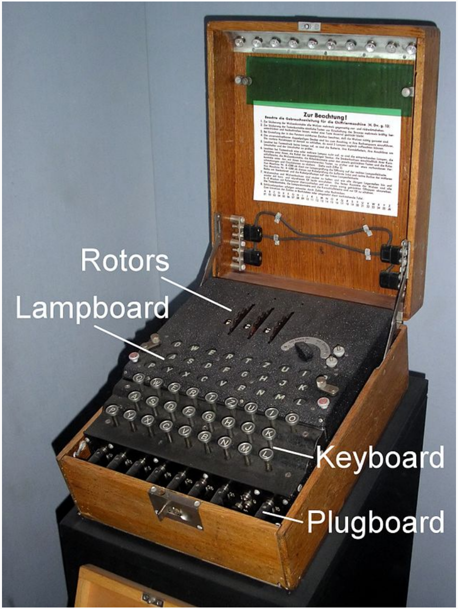
- $Z_i - \mathbb{E}[Z_i] \leq 2$ and $\mathbb{E}[Z_i^2] = \mathbb{P}\big(h(X_i) \neq h^*(X_i)\big) \leq \frac{L_D(h) - L_D^*}{2\gamma}$
- By Bernstein's inequality, with probability $\geq 1 - \delta/|\mathcal{H}|$ one has

$$L_D\big(\hat{h}_n\big) - L_D^* \leq \frac{2\log\frac{|\mathcal{H}|}{\delta}}{3n} + \sqrt{\frac{2\mathbb{E}[Z_1^2]\ln\frac{|\mathcal{H}|}{\delta}}{n}} \leq \frac{2\log\frac{|\mathcal{H}|}{\delta}}{3n} + \sqrt{\frac{\big(L_D\big(\hat{h}_n\big) - L_D^*\big)\ln\frac{|\mathcal{H}|}{\delta}}{\gamma n}}$$

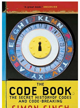
- Lemma: if $x \leq \frac{2\alpha}{3} + \sqrt{\frac{\alpha x}{\gamma}} \triangleq g(x)$ then $x \leq \frac{2\alpha}{\gamma}$, since $g(2\alpha/\gamma) \leq 2\alpha/\gamma$ for $\gamma \leq 1/2$
- Hence $\mathbb{P}\left(L_D\big(\hat{h}_n\big) - L_D^* \leq \frac{2\ln\frac{|\mathcal{H}|}{\delta}}{\gamma n}\right) \geq 1 - \delta$ and $n_{\mathcal{H}}(\epsilon, \delta) \leq \frac{2\ln\frac{|\mathcal{H}|}{\delta}}{\gamma\epsilon}$.

Application: estimating the missing mass

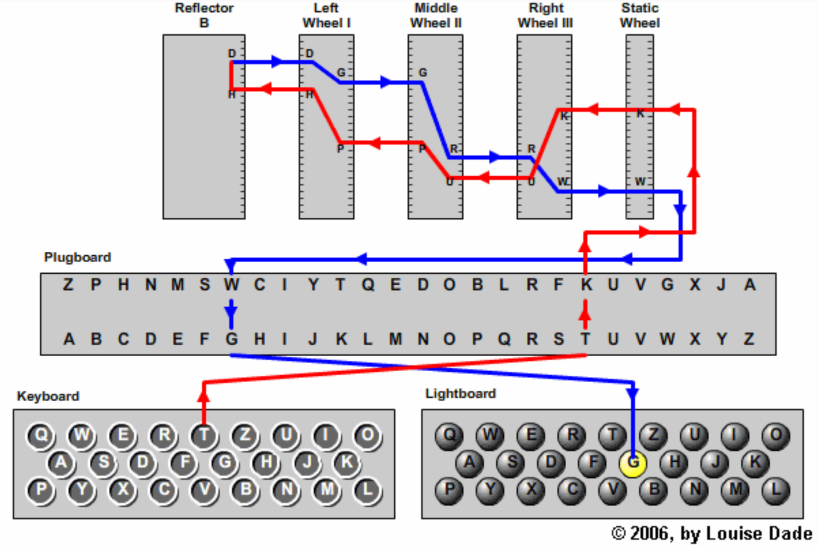
Enigma



- Electro-mechanical rotor cipher machines, 26 characters
- Invented at the end of WW1 by Arthur Scherbius
- Commercial use, then German Army during WW2
- First cracked by Marian Rejewski in the 1930s (Bomb), then improved to 3.10^{114} configurations
- Read Simon Singh, *The Code Book*

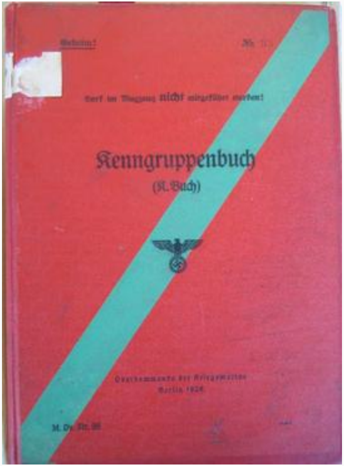


Enigma



Src: `http://enigma.louisedade.co.uk/`

Battle of the Atlantic



- Massively used by the German Kriegsmarine and Luftwaffe
- weakness: 3-letters setting to initiate communication, taken from the *Kenngruppenbuch*
- Government Code and Cypher School: Bletchley Park (on the train line between Cambridge and Oxford)
- Colossus (first programmable computers) in 1943

Estimating probabilities

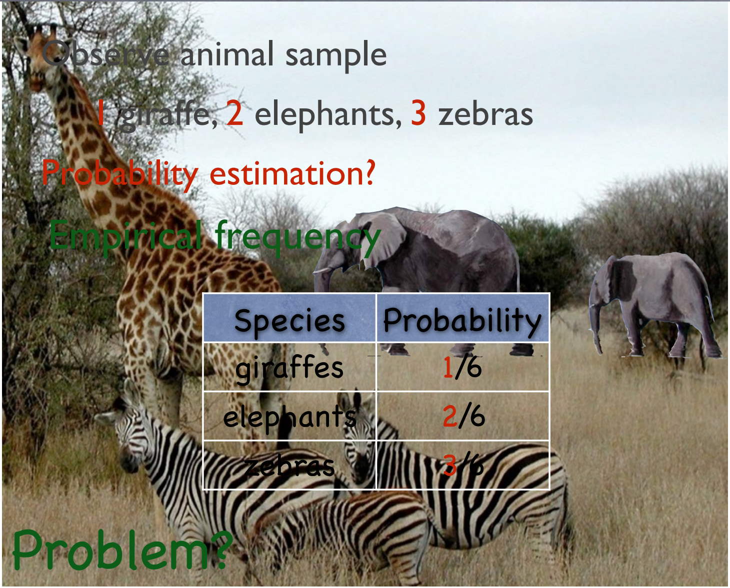
- Discrete alphabet A.
- Unknown probability P on A
- Sample $X_1, \dots, X_n$ of independent draws of P.
- Goal : use the sample estimate $P(a)$ for all $a \in A$.

Natural idea:

$$\hat{P}(a) = \frac{N(a)}{n}, \quad \text{where } N(a) = \#\{i : X_i = a\}$$

[Src: Alon Orlitsky, https://ic.epfl.ch/files/content/sites/ic/files/Inka/Orlitsky%20Talk%202016.pdf]

Safari preparation



Observe animal sample

1 giraffe, 2 elephants, 3 zebras

Probability estimation?

Empirical frequency

Species	Probability
giraffes	1/6
elephants	2/6
zebras	3/6

Problem?

[Src: Alon Orlitsky, https://ic.epfl.ch/files/content/sites/ic/files/Inka/Orlitsky%20Talk%202016.pdf]

Bigram Model for NLP [Src: https://nlp.stanford.edu/~wcmac]

Learning set:

john read moby dick

mary read a different book

she read a book by cher

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)}$$

$$P(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$$\begin{array}{cccccc}
P(& john & read & a & book &) \\
= & P(john|\cdot) & P(read|john) & P(a|read) & P(book|a) & P(\cdot|book) \\
= & \frac{c(\cdot\ john)}{\sum_w c(\cdot\ w)} & \frac{c(john\ read)}{\sum_w c(john\ w)} & \frac{c(reada)}{\sum_w c(read\ w)} & \frac{c(a\ book)}{\sum_w c(a\ w)} & \frac{c(book\ \cdot)}{\sum_w c(book\ w)} \\
= & \frac{1}{3} & \frac{1}{1} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} \\
\approx & 0.06 & & & &
\end{array}$$

Bigram Model for NLP

[Src: `https://nlp.stanford.edu/~wcmac`]

Learning set:

john read moby dick

mary read a different book

she read a book by cher

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)}$$

$$P(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$$\begin{array}{llllll}
P(& cher & read & a & book &) \\
= & P(cher|\cdot) & P(read|john) & P(a|read) & P(book|a) & P(\cdot|book) \\
= & \frac{c(\cdot\ cher)}{\sum_w c(\cdot\ w)} & \frac{c(cher\ read)}{\sum_w c(cher\ w)} & \frac{c(reada)}{\sum_w c(read\ w)} & \frac{c(a\ book)}{\sum_w c(a\ w)} & \frac{c(book\ \cdot)}{\sum_w c(book\ w)} \\
= & \frac{0}{3} & \frac{0}{1} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} \\
= & 0 & & & &
\end{array}$$

$\Longrightarrow$ useless, the unseen **must** be treated correctly.

Bayesian Approach: Laplace Estimator

Pierre-Simon de Laplace (1749-1827), Thomas Bayes (1702-1761)

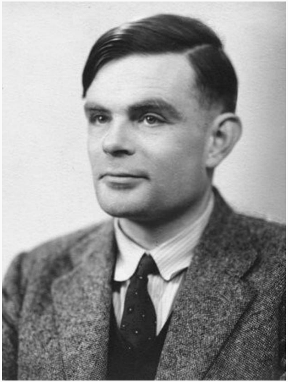
Will the sun rise tomorrow?

$$\hat{P}(a) = \frac{N(a) + 1}{n + |A|}$$

- good for small alphabets and many samples
- very bad when lots of items seen once (ex: DNA sequences)
- $|A|$ can be very large (or even infinite), but P concentrated on few items

$\Longrightarrow$ not a satisfying solution to the problem

Alan Turing

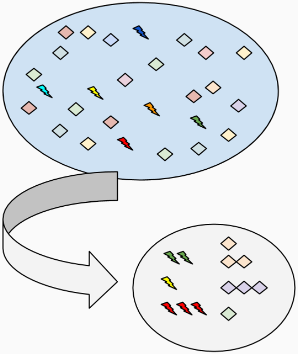


1912-1954
student of Godfrey Harold Hardy in Cambridge
PhD from Princeton with Alonzo Church

Irving John Good

1916-2009
Graduated in Cambridge
Academic carrer in Bayesian statistics in Manchester and then in the University of Virginia (USA)

Missing mass estimation



$X_1, \ldots, X_n$ independent draws of $P \in \mathfrak{M}_1(A)$.

$$O_n(x) = \sum_{m=1}^{n} \mathbb{1}\{X_m = x\}$$

How to 'estimate' the total mass of the *unseen* items

$$R_n = \sum_{x \in A} P(x)\, \mathbb{1}\{O_n(x) = 0\} \text{ ?}$$

The Good-Turing Estimator

See [I.J. Good, 1953], credits idea to A. Turing

Idea: in order to estimate the mass of the unseen

$$R_n = \sum_{x \in A} P(x) \, \mathbb{1}\{O_n(x) = 0\} \, ,$$

use the number of **hapaxes** = items seen only once (linguistic)

$$\hat{R}_n = \frac{U_n}{n}, \quad \text{where } U_n = \sum_{x \in A} \mathbb{1}\{O_n(x) = 1\}$$

Lemma [Good '53]: For *every* distribution P,

$$0 \leq \mathbb{E}\big[\hat{R}_n\big] - \mathbb{E}\big[R_n\big] \leq \frac{1}{n}$$

Completely non-parametric: no assumption on P

Bias of the Good-Turing Estimator

$$
\begin{aligned}
\mathbb{E}\big[\hat{R}_n\big] - \mathbb{E}\big[R_n\big] &= \frac{1}{n}\sum_{x\in A}\mathbb{P}\big(O_n(x)=1\big) - \sum_{x\in A} P(x)\,\mathbb{P}\big(O_n(x)=0\big) \\
&= \frac{1}{n}\sum_{x\in A} n\,P(x)\big(1-P(x)\big)^{n-1} - \sum_{x\in A} P(x)\big(1-P(x)\big)^{n} \\
&= \sum_{x\in A} P(x)\big(1-P(x)\big)^{n-1}\Big(1-\big(1-P(x)\big)\Big) \\
&= \frac{1}{n}\sum_{x\in A} P(x)\times n\,P(x)\big(1-P(x)\big)^{n-1} \\
&= \frac{1}{n}\sum_{x\in A} P(x)\,\mathbb{P}\big(O_n(x)=1\big) \\
&= \frac{1}{n}\,\mathbb{E}\left[\sum_{x\in A} P(x)\mathbb{1}\{O_n(x)=1\}\right] \quad \in \left[0,\frac{1}{n}\right]
\end{aligned}
$$

Jackknife interpretation

If we had additionnal samples, we would estimate R_n by the proportion of unseen elements in $X_{n+1}, X_{n+2}, \ldots$

We have no additionnal samples, **but** we keep every observation as a "test", pretending that the samples was made of everything else:

$$\begin{aligned}\hat{R}_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big\{x_i \notin \{x_j : j \neq i\}\big\} \\ &= \frac{1}{n}\sum_{i=1} \mathbb{1}\big\{O_n(x_i) = 1\big\} \\ &= \frac{1}{n}\sum_{x \in A} \mathbb{1}\big\{O_n(x) = 1\big\}\end{aligned}$$

Remark: jackknife is a **resampling method**, related to **bootstrap** and **crossvalidation** (of great use in Machine Learning).

Deviation Bounds

Proposition: With probability at least $1-\delta$ for *every* P,

$$\hat{R}_n - \frac{1}{n} - (1+\sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}} \leq R_n \leq \hat{R}_n + (1+\sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}}$$

See [McAllester and Schapire '00, McAllester and Ortiz '03]:

- deviations of $\hat{R}_n$: McDiarmid's inequality
- deviations of R_n: negative association

Other tool: Poissonization [see Optimal Probability Estimation with Applications to Prediction and Classification, by Acharya, Jafarpour, Orlitsky Suresh, Colt 2013]

References For Negative Association

Negative Association - Definition, Properties, and Applications, *by David Wajc* https://www.cs.cmu.edu/~dwajc/notes/Negative%20Association.pdf

Balls and Bins:A Study in Negative Dependence, *by Balls and Bins:A Study in Negative Dependence*, https://www.brics.dk/RS/96/25/BRICS-RS-96-25.pdf

Definition

Intuitively: $X_1, \dots, X_n$ are negatively associated when, if a subset I a variables is "high", a disjoint subset J has to be "low".

Definition

A set of real-valued random variables $X_1, X_2, ..., X_n$ is said to be negatively associated (NA) if for any two disjoint index sets $I, J \subset [n]$ and two functions f, g both monotone increasing or both monotone decreasing, it holds

$$\mathbb{E}\Big[f(X_i : i \in I)\, g\big(X_j : j \in J\big)\Big] \leq \mathbb{E}\big[f(X_i : i \in I)\big]\, \mathbb{E}\Big[g\big(X_j : j \in J\big)\Big]$$

NB: f is *monotone increasing* if $\forall i \in I, x_i \leq x'_i$ implies $f(x) \leq f(x')$.

First properties

Let $X_1, X_2, ..., X_n$ be NA.

- For all $i \neq j$, $\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ i.e. $\mathrm{Cov}(X_i, X_j) \leq 0$.
- For any disjoints subsets $I, J \subset [n]$ and all $x_1, \ldots, x_n$,

$$\mathbb{P}\big(X_i \geq x_i : i \in I \cup J\big) \leq \mathbb{P}\big(X_i \geq x_i : i \in I\big)\,\mathbb{P}\big(X_j \geq x_j : j \in J\big) \quad \text{and}$$
$$\mathbb{P}\big(X_i \leq x_i : i \in I \cup J\big) \leq \mathbb{P}\big(X_i \leq x_i : i \in I\big)\,\mathbb{P}\big(X_j \leq x_j : j \in J\big)$$

- For all monotone increasing functions $f_1, \ldots, f_k$ depending on disjoint subsets of the $(X_i)_i$,

$$\mathbb{E}\Big[\prod_j f_j(X)\Big] \leq \prod_j \mathbb{E}\big[f_j(X)\big]$$

- For all $x_1, \ldots, x_n$,

$$\mathbb{P}\left(\bigcap_i \{X_i \geq x_i\}\right) \leq \prod_i \mathbb{P}(X_i \geq x_i)$$
$$\text{and} \quad \mathbb{P}\left(\bigcap_i \{X_i \leq x_i\}\right) \leq \prod_i \mathbb{P}(X_i \leq x_i)$$

Consequence: NA concentrates better than independent

For Chernoff's method (which relies on exponential moments), NA variables can simply be treated as independent!

In particular:

Chernoff-Hoeffding bound

Let $X_1, \ldots, X_n$ be NA random variables with $X_i \in [a_i, b_i]$ a.s. Then $S = X_1 + \cdots + X_n$ satifies Hoeffding's tail bound: for all $t \geq 0$,

$$\mathbb{P}\Big[\big|S - E[S]\big| \geq t\Big] \leq 2 \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

Examples of NA variables

- Independent variables...
- **0-1 principle** If $X_1, \ldots X_n$ are Bernoulli variables and $\sum_i X_i \leq 1$ a.s., then they are NA.

  Let f and g are monotically increasing and depend on disjoint subsets of indices. $\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)] \iff \mathbb{E}[\tilde{f}(X)\tilde{g}(X)] \leq \mathbb{E}[\tilde{f}(X)]\mathbb{E}[\tilde{g}(X)]$, where $\tilde{f}(X) = f(X) - f(\vec{0})$ and $\tilde{g}(X) = g(X) - g(\vec{0})$. But $\tilde{f}(X)\tilde{g}(X) = 0$ always, while $\tilde{f}(X) \geq 0$ and $\tilde{g}(X) \geq 0$.
- **Permutation distributions** If $x_1 \leq \cdots \leq x_n$ and if $X_1, \ldots, X_n$ are random variables such that $\left\{X_1, \ldots, X_n\right\} = \left\{x_1, \ldots, x_n\right\}$ a.s., with all assignments equally likely, then they are NA.
- **Sampling without replacement** If $X_1, \ldots, X_n$ are sample without replacement from $\{x_1, \ldots, x_N\}$ (with $N \geq n$), then they are NA.

Closure properties

Union

If the $\{X_i : i \in I\}$ are NA, if $\{Y_j : j \in J\}$ are NA, and if the $\{X_i\}$ are independent from the $\{Y_j\}$, then the $\{X_i, Y_j : i \in I, j \in J\}$ are NA.

Concordant monotone

If the $\{X_i : i \in I\}$ are NA, if $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ are all monotonically increasing and depend on different subsets of $[n]$, then $\left\{f_j(X) : 1 \leq j \leq k\right\}$ are NA.
The same holds if $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ are all monotonically decreasing.

Bins and balls

The standard bins and balls process consists of m balls and n bins.

- each ball b is independently placed in bin i with probability $p_{b,i}$: $X_b \overset{indep}{\sim} \mathcal{M}ulti(p_{b,\cdot})$.
- *occupancy number* $B_i = \sum_{b=1}^{n} \mathbb{1}\{X_b = i\}$ number of balls in bin i.

In particular $\displaystyle\sum_{i=1}^{n} B_i = m$.

Prop: The B_i are NA.

Let $X_{b,i} = \mathbb{1}\{\text{ball b fell into bin i}\}$. By the $0-1$ principle, for all $1 \leq b \leq m$ the $\{X_{b,i} : 1 \leq i \leq n\}$ are NA. By independence and closure under union, so are the $\{X_{b,i} : 1 \leq b \leq m, 1 \leq i \leq n\}$. By closure under concordant monotone functions, the $B_i = \displaystyle\sum_{b=1}^{n} X_{b,i}$ are NA.

Consequence: Concentration of the number $N = \sum_i \mathbb{1}\{B_i = 0\}$ of empty bins, since the $(\mathbb{1}\{B_i = 0\})_i$ are NA.

If $p_{b,i} = 1/n$, then the number N of empty bins satisfies $N = n\,e^{-m/n} \pm O\big(\sqrt{n\,e^{-m/n}}\big)$.

Deviations of the missing mass R_n

R_n is better concentrated than $S_n = \sum_{x \in A} P(x)\, B_x$ where the $B_x \sim \mathcal{B}\Big(\big((1 - P(x)\big)^n\Big)$ are independent.

Hence

$$\mathbb{V}ar[R_n] \leq \sum_x P(x)^2 e^{-nP(x)} \leq \sum_x P(x) \max_{0 \leq u \leq 1} u\, e^{-nu} = \frac{1}{ne}$$

and

$$\mathbb{P}\big(R_n \geq \mathbb{E}[R_n] - \epsilon\big) \leq \exp\left(-\frac{n\epsilon^2}{2e}\right) .$$