Concentration: Lower bounds for deviations, and No Free Lunch theorem

Master 2 Mathematics and Computer Science

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

2020-2021

UNIVERSITÉ DE LYON | ENS DE LYON

Table of contents

Deviation Bound for Bernoulli Variables

Chernoff's Bound

Theorem (Chernoff-Hoeffding Deviation Bound)

Let $\mu \in (0,1)$. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{B}(\mu)$, *and let* $x \in (\mu, 1]$.

(i) *Chernoffs' bound for Bernoulli variables:*

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(-n\,\mathrm{kl}(x,\mu)\big)\ , \tag{1}$$

where $\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. *Same for left deviations.*

(ii) *If* $\phi(x) = \mathrm{kl}(x,\mu)$, *then* $\phi''(x) = 1/[x(1-x)]$ *and*

$$\begin{aligned}\mathrm{kl}(x,\mu) &= \frac{(x-\mu)^2}{2}\int_0^1 \phi''\big(\mu + s(x-\mu)\big)\ 2(1-s)ds \\ &\geq \frac{(x-\mu)^2}{2\tilde{x}(1-\tilde{x})} \quad \text{with } \tilde{x} = \frac{2\mu+x}{3} \text{ by Jensen, since } \phi'' \text{ is convex and } \int_0^1 s\,2(1-s)ds = \frac{1}{3} \\ &\geq \frac{1}{2\max_{x\leq u\leq p} u(1-u)}(x-\mu)^2 \quad \geq 2(x-\mu)^2\ .\end{aligned}$$

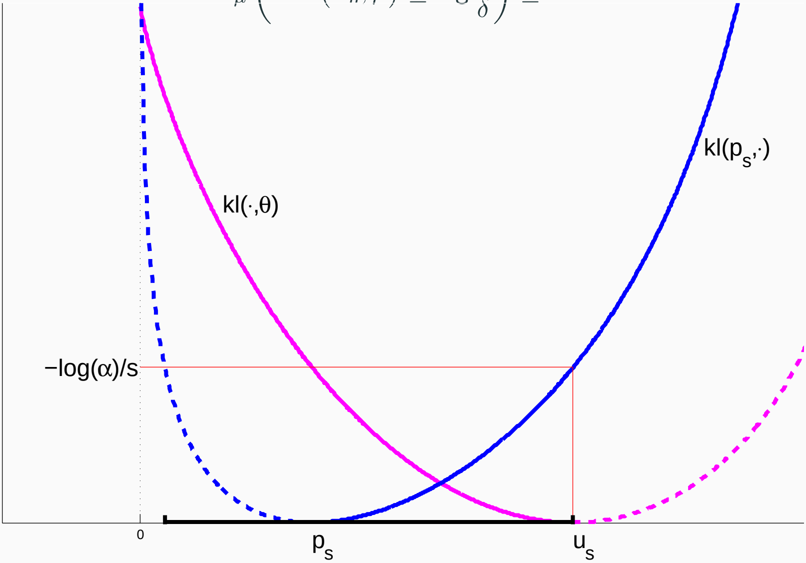
(iii) *Hoeffding's bound for Bernoulli variables:*

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(-2n(x-\mu)^2\big)\ . \tag{2}$$

(iv) *Inequalities* (1) *and* (2) *hold for arbitrary independent random variables with range* $[0,1]$ *and expectation* μ.

A Divergence on the Set of Possible Means

$$\mathbb{P}_\mu\left(n\ \mathrm{kl}\left(\bar{X}_n, \mu\right) \geq \log\frac{1}{\delta}\right) \leq 2\delta$$



Examples

- If $\mu < 1/2$,

$$\mathbb{P}\left(\bar{X}_n > \frac{1}{2}\right) \leq \exp\left(-\frac{n}{2}(1-2\mu)^2\right) \ .$$

(Consequence of Chernoff or direct computation with $(1-u)^n \leq exp(-n\,u)$, or of Hoeffding).

- For all $\mu \in [0,1]$, Chernoff's bound with $\log(u) \geq (u-1)/u$ yields

$$\mathbb{P}\left(\bar{X}_n < \frac{\mu}{2}\right) \leq \exp\left(-\frac{1-\log(2)}{2}\,n\mu\right) \approx \exp\left(-0.153\,n\mu\right) \leq \exp\left(-\frac{n\mu}{7}\right) \ .$$

Hoeffding yields a very poor result, but (ii) gives:

$$\mathbb{P}\left(\bar{X}_n < \frac{\mu}{2}\right) \leq \exp\left(-\frac{3}{20}n\mu\right) = \exp\left(-0.15\,n\mu\right) \leq \exp\left(-\frac{n\mu}{8}\right) \ .$$

Sub-Gaussian inequalities

Bennett's and Bernstein's inequalities

Let $(X_i)_{1\leq i\leq n}$ be independent random variables upper-bounded by 1, let $\bar{\mu} = (\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n])/n$, let σ^2 be such that $\mathbb{E}[X_i^2] \leq \sigma^2$ for all i and let $\phi(u) = (1+u)\log(1+u) - u$. Then, for all $x > 0$,

$$\mathbb{P}\big(\bar{X} \geq \bar{\mu} + x\big) \leq \exp\left(-n\,\sigma^2\phi\left(\frac{x}{\sigma^2}\right)\right) \leq \exp\left(-\frac{n\,x^2/2}{\sigma^2 + x/3}\right) .$$

Bernstein from Bennett: $\phi(x) \geq \dfrac{x^2}{2\left(1+\frac{x}{3}\right)}$ since $\psi(x) = 2\left(1+\frac{x}{3}\right)\phi(x) - x^2 \geq 0$.

Extension: if $X_i \leq b$ with $b > 0$,

$$\mathbb{P}\big(\bar{X}_n \geq \bar{\mu} + x\big) \leq \exp\left(-\frac{n\sigma^2}{b^2}\phi\left(\frac{bx}{\sigma^2}\right)\right) \leq \exp\left(-\frac{n\,x^2/2}{\sigma^2 + bx/3}\right) .$$

Example: for X with range in $[0,1]$,

$$\mathbb{P}\left(\bar{X}_n < \frac{\mu}{2}\right) \leq \exp\left(-n\left(\frac{3}{2}\log\frac{3}{2} - \frac{1}{2}\right)\mu\right) \leq \exp\left(-\frac{3n\mu}{28}\right) .$$

Parenthesis: a nice proof for the technicalities of Bernstein

From [Pollard, MiniEmpirical ex.14, http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf]

For any sufficiently smooth real-valued function g defined at least in a neighborhood of 0 let

$$G(x) = \frac{g(x) - g(0) - xg'(0)}{x^2/2} \text{ if } x \neq 0, \text{ and } G(0) = g''(0) \ .$$

By Taylor's integral formula

$$g(x) - g(0) - xg'(0) = \int_0^x g''(u)(x-u)du = x^2 \int_0^1 g''(sx)(1-s)ds \ .$$

Thus, $G(x) = \int g''(sx)d\nu(s)$, where $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$.

Hence, if g is convex then $g'' \geq 0$ and $G \geq 0$. Moreover, if g'' is increasing then the functions $x \mapsto g''(sx)$ for $s \in [0,1]$ are all increasing and G is also increasing as an average of increasing functions. For $g(u) = \exp(u)$, this yields that $(\exp(u) - u - 1)/u^2$ is increasing, as required for the proof of Bernstein's inequality.

Similarly, if g'' is convex then G is also convex as an average of convex functions $\Big(x \mapsto g''(sx)\Big)_s$. Moreover, by Jensen's inequality applied to convex function $\psi(s) = g''(xs)$ with the probability measure $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$

$$G(x) = \int_0^1 g''(xs)\, 2(1-s)ds \geq g''\left(x \int_0^1 s \times 2(1-s)ds\right) = g''\left(\frac{x}{3}\right) \ .$$

For $g(u) = (1+u)\log(1+u) - u$, $g''(u) = 1/(1+u)$ and this yields:

$$\frac{g(u)}{u^2/2} \geq g''\left(\frac{u}{3}\right) = \frac{1}{1+u/3} \ .$$

Exercise: for $X_i \overset{iid}{\sim} \mathcal{B}(\mu)$, $\mathbb{P}(\bar{X}_n \geq 2\mu) \leq \exp(-n\times?)$

Chernoff + Taylor: since $\log(u) \geq (u-1)/u$,

$$\mathrm{kl}(2\mu, \mu) = 2\mu \log(2) + (1-2\mu)\log\frac{1-2\mu}{1-2\mu} \geq 2\mu\log(2) - \mu = \mu(2\log(2)-1) \approx 0.386\,\mu\ .$$

Chernoff with convexity:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{4/3\mu} = \frac{3}{8}\,\mu = 0.375\mu\ .$$

Improved Hoeffding:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{\max_{\mu\leq u\leq 2\mu} u(1-u)} \geq \frac{\mu^2/2}{2\mu} = \frac{1}{4}\,\mu = 0.25\mu\ .$$

Bennett:

$$2\mu\log\frac{2\mu}{\mu} - (2\mu-\mu) = \mu(2\log(2)-1) \approx 0.386\,\mu\ .$$

Bernstein:

$$\frac{(2\mu-\mu)^2/2}{\mu(1-\mu)+(2\mu-\mu)/3} \geq \frac{\mu^2/2}{\mu+\mu/3}\,\frac{3}{8}\,\mu = 0.375\mu\ .$$

Hoeffding: $2(2\mu-\mu)^2 = 2\mu^2$, very poor (as expected) when μ is small.

Kullback-Leibler divergence

Kullback-Leibler divergence

Definition

Let P and Q be two probability distributions on a measurable set Ω. The Kullback-Leibler divergence from Q to P is defined as follows:

- if P is not absolutely continuous with respect to Q, then $\mathrm{KL}(P, Q) = +\infty$;
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

$$\mathrm{KL}(P, Q) = \int_\Omega \log \frac{dP}{dQ}\, dP = \int_\Omega \frac{dP}{dQ} \log \frac{dP}{dQ}\, dQ\ .$$

Property: $0 \leq \mathrm{KL}(P, Q) \leq +\infty$, $\mathrm{KL}(P, Q) = 0$ iff $P = Q$.

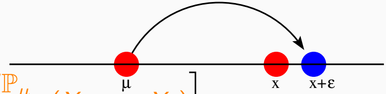
If $P \ll Q$ and $f = \frac{dP}{dQ}$, $\int_\Omega f \log(f)\, dQ = \int_\Omega \left[f \log(f)\right]_+ dQ - \int_\Omega \left[f \log(f)\right]_- dQ$, the later is finite since $\left[f \log(f)\right]_- \leq 1/e$.

Examples:

$\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = \mathrm{kl}(p, q)$, $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

Lower Bound: Change of Measure

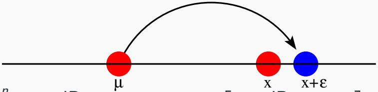
For all $\epsilon > 0$ and all $\alpha > 0$,



$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times \frac{d\mathbb{P}_\mu}{d\mathbb{P}_{x+\epsilon}}(X_1, \dots, X_n)\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad\qquad \times e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon, \mu) + \alpha\right\}}\left(1 - o_n(1)\right).
\end{aligned}
$$

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,



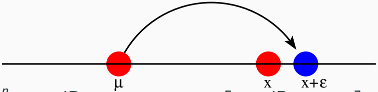
$$\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad \times e^{-\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right).
\end{aligned}$$

Asymptotic Optimality (Large Deviation Lower Bound)

$$\liminf_n \frac{1}{n}\log\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \geq -\mathrm{kl}(x,\mu).$$

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,



$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^{n} \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad \times e^{-\sum_{i=1}^{n} \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right) .
\end{aligned}
$$

Asymptotic Optimality (Large Deviation Principle)

$$
\frac{1}{n}\log \mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \underset{n\to\infty}{\longrightarrow} -\mathrm{kl}(x,\mu) .
$$

Properties of KL divergence

Tensorization of entropy:

If $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, then

$$\mathrm{KL}(P, Q) = \mathrm{KL}(P_1, Q_1) + \mathrm{KL}(P_2, Q_2)\,.$$

Contraction of entropy data-processing inequality:

Let $(\Omega, \mathcal{A})$ be a measurable space, and let P and Q be two probability measures on $(\Omega, \mathcal{A})$. Let $X : \Omega \to (\mathcal{X}, \mathcal{B})$ be a random variable, and let P^X (resp. Q^X) be the push-forward measures, ie the laws of X wrt P (resp. Q). Then

$$\mathrm{KL}\left(P^X, Q^X\right) \leq \mathrm{KL}(P, Q)\,.$$

Pinsker's inequality:

Let $P, Q \in \mathfrak{M}_1(\Omega, \mathcal{A})$. Then

$$\|P - Q\|_{TV} \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \leq \sqrt{\frac{\mathrm{KL}(P, Q)}{2}}\,.$$

Proof: contraction

Contraction: if $\mathrm{KL}(P, Q) = +\infty$, the result is obvious. Otherwise, $P \ll Q$ and there exists $\frac{dP}{dQ} : \Omega \to \mathbb{R}$ such that for all measurable $f : \Omega \to \mathbb{R}$, $\int_\Omega f\, dP = \int_\Omega f\, \frac{dP}{dQ}\, dQ$.

- We first prove that $P^X \ll Q^X$ and, if $\gamma(x) := \mathbb{E}_Q\left[\frac{dP}{dQ}\middle| X = x\right]$ is the Q-a.s. unique function such that $\mathbb{E}_Q\left[\frac{dP}{dQ}\middle| X\right] = \gamma(X)$, then $\gamma = \frac{dP^X}{dQ^X}$. Indeed, for all $B \in \mathcal{B}$,

$$
\begin{aligned}
P^X(B) = P(X \in B) &= \int_{X \in B} \frac{dP}{dQ} dQ = \mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X \in B\}\right] \\
&= \mathbb{E}_Q\left[\mathbb{E}_Q\left[\frac{dP}{dQ}\mathbb{1}\{X \in B\}\middle| X\right]\right] = \mathbb{E}_Q\left[\mathbb{1}\{X \in B\}\mathbb{E}_Q\left[\frac{dP}{dQ}\middle| X\right]\right] \\
&= \mathbb{E}_Q\left[\mathbb{1}\{X \in B\}\gamma(X)\right] = \int_{X \in B} \gamma(X) dQ = \int_B \gamma dQ^X
\end{aligned}
$$

  and hence $P^X \ll Q^X$ and $\frac{dP^X}{dQ^X} = \gamma$.
- Now,

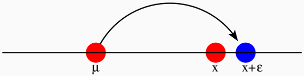
$$
\begin{aligned}
\mathrm{KL}\left(P^X, Q^X\right) &= \int_{\mathcal{X}} \gamma \log \gamma \; dQ^X = \int_\Omega \gamma(X) \log \gamma(X)\, dQ \\
&= \mathbb{E}_Q\left[\phi\left(E_Q\left[\frac{dP}{dQ}\middle| X\right]\right)\right] \quad \text{where } \phi := x \mapsto x\log(x) \text{ is convex} \\
&\leq \mathbb{E}_Q\left[\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\middle| X\right]\right] \qquad \text{by (conditional) Jensen's inequality} \\
&= \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right] = \mathrm{KL}(P, Q)\ .
\end{aligned}
$$

Proof: Pinsker

Let $A \in \mathcal{A}$, $p = P(A)$ and $q = Q(A)$. By contraction,

$$\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^{\mathbb{1}_A}, Q^{\mathbb{1}_A}) = \mathrm{KL}\left(\mathcal{B}(P(A)), \mathcal{B}(Q(A))\right) = \mathrm{kl}\left(P(A), Q(A)\right) \geq 2\big(P(A) - Q(A)\big)^2 .$$

Lower Bound: the Entropic Way



Let $\Omega = \{0,1\}^n$, $X_i(\omega) = \omega_i$

Probability laws on Ω: $\mathbb{P}_p = \mathcal{B}(p)^{\otimes n}$.

For all $\epsilon > 0$,

$$n\,\mathrm{kl}(x+\epsilon, \mu) = \mathrm{KL}\left(\mathbb{P}_{x+\epsilon}, \mathbb{P}_\mu\right) \qquad \mathrm{KL}(P \otimes P', Q \otimes Q') = \mathrm{KL}(P, Q) + \mathrm{KL}(P', Q')$$

$$\geq \mathrm{KL}\left(\mathbb{P}_{x+\epsilon}^{\mathbb{1}\{\bar{X}_n \geq x\}}, \mathbb{P}_\mu^{\mathbb{1}\{\bar{X}_n \geq x\}}\right)$$

$\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^X, Q^X)$
contraction of entropy
= data-processing inequality

$$= \mathrm{kl}\left(\mathbb{P}_{x+\epsilon}\left(\bar{X}_n \geq x\right), \mathbb{P}_\mu\left(\bar{X}_n \geq x\right)\right)$$

$$\geq \mathbb{P}_{x+\epsilon}\left(\bar{X}_n \geq x\right) \log \frac{1}{\mathbb{P}_\mu\left(\bar{X}_n \geq x\right)} - \log(2) \qquad \mathrm{kl}(p, q) \geq p \log \frac{1}{q} - \log 2$$

A non-asymptotic lower bound

$$\forall \epsilon > 0, \qquad \mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \geq e^{-\frac{n\,\mathrm{kl}(x+\epsilon,\mu) + \log(2)}{1 - e^{-2n\epsilon^2}}}.$$

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

The No-Free-Lunch theorem

A learning algorithm A for binary classification maps a sample $S \sim \mathcal{D}^{\otimes n}$ to a decision rule $\hat{h}_n$.

Theorem

Let A be any learning algorithm for binary classification over a domain $\mathcal{X}$. If the training set size is $n \leq |\mathcal{X}|/2$, 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 $n \geq 8 \log(7|\mathcal{H}|/6)$, is a successful learner in that setting.

Proof

Take $C \subset \mathcal{X}$ of cardinality $2n$, and $\{0,1\}^C = \{f_1, \ldots, f_T\}$ where $T = 2^{2n}$. For each $1 \le i \le T$, we denote by D_i the probability distribution on $C \times \{0,1\}$ defined by

$$D_i(\{x,y\}) = \begin{cases} \frac{1}{2n} \text{ if } y = f_i(x) , \\ 0 \text{ otherwise.} \end{cases}$$

We will show that $\max_{1 \le i \le T} \mathbb{E}\left[L_{D_i}(A(S))\right] \ge 1/4$, which entails the result thanks to the small lemma: if $P(0 \le Z \le 1) = 1$ and $\mathbb{E}[Z] \ge 1/4$, then $\mathbb{P}(Z \ge 1/8) \ge 1/7$. Indeed, $1/4 \le \mathbb{E}[Z] \le \mathbb{P}(Z < 1/8)/8 + \mathbb{P}(Z \ge 1/8) = 1/8 - 7\,\mathbb{P}(Z \ge 1/8)/8$.

All the X-samples $S_1^X, \ldots, S_k^X$, for $k = (2n)^n$, are equaly likely. For $1 \le j \le k$, if $S_j^X = (x_1, \ldots, x_n)$ we denote by $S_j^i = \big((x_1, f_i(x_1)), \ldots, (x_n, f_i(x_n)\big)$, and $\hat{f}_j^i = A(S_j^i)$.

$$\max_{1 \le i \le T} \mathbb{E}\left[L_{D_i}(A(S))\right] = \max_{1 \le i \le T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(\hat{f}_j^i\right) \ge \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(\hat{f}_j^i\right)$$

$$= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) \ge \min_{1 \le j \le k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) .$$

Fix $1 \le j \le k$, denote $S_j^X = (x_1, \ldots, x_n)$ and define $\{v_1, \ldots, v_p\} = C \setminus \{x_1, \ldots, x_n\}$, where $p \ge n$. Then

$$L_{D_i}\left(\hat{f}_j^i\right) = \frac{1}{2n} \sum_{x \in C} \mathbb{1}\{\hat{f}_j^i(x) \neq f_i(x)\} \ge \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\}$$

and hence

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(\hat{f}_j^i\right) \ge \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} \ge \frac{1}{2} \min_{1 \le r \le p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} .$$

Fix $1 \le r \le p$. Then the functions $\{f_i : 1 \le i \le T\}$ can be grouped into $T/2$ pairs of functions $\left(\tilde{f}_i^0, \tilde{f}_i^1\right)$, $1 \le i \le T/2$ which agree on all $x \in C$ except on v_r, and for all $1 \le i \le T/2$ it holds that $\mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^1(v_r)\} = 1$. Hence, $\sum_{i=1}^{T} \mathbb{1}\{\hat{f}_j^i(v_r) \neq f_i(v_r)\} = \sum_{i=1}^{T/2} \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{\hat{f}_j^i(v_r) \neq \tilde{f}_i^1(v_r)\} = T/2$, which concludes the proof.

Consequence: infinite VC-dimension $\implies$ no learnability

Recall that 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 $S = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \overset{iid}{\sim} D$,

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

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

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 $C \subset \mathcal{X}$ 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 $\mathcal{X} \times \{0,1\}$ and $h : \mathcal{X} \to \{0,1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.

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 $n \leq (c+1)^d/2$, then there exists a distribution $\mathcal{D}$ over $[0,1]^d \times \{0,1\}$ such that:

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

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