Butlletí de la Societat Catalana de Matemàtiques
Vol. 24, núm. 2, 2009. Pàg. 97–136. DOI: 10.2436/20.2002.01.23

Desigualtats de concentració*

Gábor Lugosi

Resum Les lleis dels grans nombres en la teoria clàssica de probabilitats asseguren que la suma de variables aleatòries independents es troba, sota certes condicions febles, molt a prop del seu valor esperat amb alta probabilitat. Aquestes sumes són l'exemple més senzill de variables aleatòries concentrades al voltant de la seva mitjana. Alguns resultats més recents revelen que aquest comportament és compartit per una immensa classe de funcions de variables aleatòries independents. Aquests resultats es coneixen generalment com a *desigualtats de concentració*.

El propòsit d'aquest article és oferir una introducció a algunes d'aquestes desigualtats.

Paraules clau: desigualtats de concentració.

Classificació MSC2010: 60E15, 60F10.

1 Introducció

Tal com assegura la llei dels grans nombres, potser el més fonamental dels principis de la teoria de probabilitats, el valor mitjà de variables aleatòries independents es troba, amb alta probabilitat, a prop del seu valor esperat, sempre que es compleixin algunes condicions tècniques naturals. Una gran part de la recerca que s'ha dut a terme en la teoria de probabilitats clàssica ha estat dedicada a quantificar aquesta llei.

A les últimes dècades, la teoria clàssica s'ha estès d'una manera inesperadament general i potent. D'acord amb aquests descobriments recents, anomenats *fenòmens de concentració de mesura*, qualsevol funció de diverses variables aleatòries independents és propera al seu valor esperat, sempre que cap de les variables no tingui una influència massa gran sobre el valor de la funció. Aquesta elegant teoria ha tingut un gran impacte en diverses àrees de

* L'autor reconeix el suport de la beca MTM2006-05650 del Ministeri de Ciència i Tecnologia espanyol i de la PASCAL Network of Excellence amb la beca EC núm. 506778.

Traducció de l'anglès de Guillem Perarnau i Oriol Serra.

les matemàtiques, incloent-hi l'anàlisi, la geometria en grans dimensions, la combinatòria i les matemàtiques discretes, l'estadística matemàtica i altres.

El propòsit d'aquest article és oferir una introducció senzilla a alguns dels resultats i les idees que rauen darrere aquest fascinant camp en constant creixement.

Primer de tot, recordarem l'essencial d'algunes coses bàsiques. Per a qualsevol variable aleatòria no negativa X, l'esperança de X es pot expressar com,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq t\}dt.$$

Això implica la *desigualtat de Markov*, que diu que, per a qualsevol variable aleatòria no negativa X i $t > 0$,

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}.$$

Resulta de la desigualtat de Markov que si ϕ és una funció no negativa estrictament creixent, aleshores, per a qualsevol variable aleatòria X i nombre real t,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}.$$

Una aplicació d'això amb $\phi(x) = x^2$ és la *desigualtat de Txebixev*: si X és una variable aleatòria arbitrària i $t > 0$, aleshores

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\left\{|X - \mathbb{E}X|^2 \geq t^2\right\} \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^2\right]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}.$$

Més en general, prenent $\phi(x) = x^q$ $(x \geq 0)$ per a qualsevol $q > 0$, tenim que

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^q\right]}{t^q}.$$

En exemples específics es pot triar el valor de q per tal d'optimitzar la fita superior que en resulta. Aquestes fites sobre els moments proporcionen sovint una estimació molt acurada de les cues de probabilitat.[1] Una idea relacionada és a la base del *mètode de fitació de Chernoff*. Prenent $\phi(x) = e^{\lambda x}$ on λ es un nombre positiu arbitrari, per a qualsevol variable aleatòria X i qualsevol $t > 0$, tenim que

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{\lambda X} \geq e^{\lambda t}\} \leq \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda t}}.$$

Les fites superiors per la probabilitat de les cues de X es poden obtenir estimant la *funció generadora de moments* $F(\lambda) = \mathbb{E}e^{\lambda X}$. En el mètode de Chernoff, es

1 En tot l'article les expressions $\mathbb{P}\{|X - \mathbb{E}X|\} \geq t$ s'anomenen *cues de probabilitat*, traducció del terme anglès d'ús estès *tail probabilities*. L'objectiu general de tota la teoria que s'exposa és veure que aquestes cues tenen valors petits i que tota la probabilitat es *concentra* al voltant del valor mitjà. (N. del t.)

troba una $\lambda > 0$ que minimitza la fita superior. Això sovint és més convenient que optimitzar la fita basada en moments presentada anteriorment.

El mètode de Chernoff resulta ideal per tractar sumes de variables aleatòries independents. Per tal d'il·lustrar-ho, siguin $X_1, \ldots, X_n$ variables aleatòries independents que prenen valors reals i sigui $S_n = \sum_{i=1}^n X_i$. A causa de la independència, $\mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_i)$ i utilitzant la desigualtat de Txebixev,

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\} \leq \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{t^2}.$$

En altres paraules, escrivint $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i)$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

La raó per la qual el mètode de fitació de Chernoff és tan convenient per a les sumes de variables aleatòries independents és que, ja que el valor esperat per al producte de variables aleatòries independents és igual al producte dels seus valors esperats, la fita de Chernoff esdevé

$$\begin{aligned}\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right)\right] \\ &= e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}X_i)}\right] \quad \text{(per la independència).} \qquad (1)\end{aligned}$$

Ara, el problema de trobar fites ajustades es redueix a trobar bones fites superiors per a la funció generadora de moments de les variables aleatòries $X_i - \mathbb{E}X_i$. Hi ha moltes maneres de fer-ho. Per a variables aleatòries fitades, potser la versió més elegant és la *desigualtat de Hoeffding*, que assegura que, si X és una variable aleatòria amb $\mathbb{E}X = 0$ que pren valors en l'interval $[a, b]$, aleshores, per a qualsevol $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2(b-a)^2/8}.$$

Aquest simple fet, combinat amb (1) implica immediatament la desigualtat de cues de Hoeffding [47]:

1 Teorema *Siguin $X_1, \ldots, X_n$ variables aleatòries independents i fitades tals que X_i pren valors en $[a_i, b_i]$ amb probabilitat 1. Aleshores, per a qualsevol $t > 0$, tenim que*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

i

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

El teorema anterior és generalment conegut com a *desigualtat de Hoeffding*. Per apreciar la seva potència, considerem una distribució binomial simètrica, és a dir, quan la X_i pren valors 0 i 1 amb probabilitat $1/2$. En aquest cas especial tenim que

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \leq e^{-2t^2/n}.$$

El teorema central del límit mostra que no es poden esperar fites superiors significativament millors.

Notes bibliogràfiques

Es coneixen diversos mètodes per provar desigualtats de concentració, incloent-hi mètodes basats en martingales (vegeu els articles panoràmics de McDiarmid [73, 74]), mètodes de teoria de la informació (vegeu Ahlswede, Gács i Körner [1]; Marton [65, 66, 67]; Dembo [32]; Massart [68], i Rio [88]), el mètode d'inducció de Talagrand [96, 94, 95] vegeu també Luczak i McDiarmid [64]; McDiarmid [75]; Panchenko [80, 81, 82], el mètode de desacoblament que es troba explicat a De La Peña i Giné [83] i l'anomenat *mètode de l'entropia*, basat en desigualtats logarítmiques de Sobolev, desenvolupat per Ledoux, [57, 56], (vegeu també Bobkov i Ledoux [13]; Massart [69]; Rio [88]; Klein [52]; Boucheron, Lugosi i Massart [17, 18]; Bousquet [20], i Boucheron, *et al.* [16]). Alguns mètodes específics per al problema també s'han desenvolupat en teoria de grafs aleatoris (vegeu Janson, Łuczak i Ruciński [50] per a un article panoràmic).

Les desigualtats exponencials de cues per a sumes de variables aleatòries independents van ser provades des dels inicis de la teoria de probabilitats. Entre els pioners, mencionem Bernstein [10], Craig [27], Uspensky [100], Chernoff [22], Okamoto [79], Bennett [9] i Hoeffding [47]. La desigualtat de Hoeffding és deguda a Hoeffding [47]. Per a variables aleatòries binomials, va ser provada per Chernoff [22] i Okamoto [79].

2 Fitant la variància: la desigualtat d'Efron-Stein

El principal missatge de les desigualtats de concentració que descriurem en aquesta secció és que moltes de les desigualtats de cues per a sumes de variables aleatòries independents poden ser esteses a funcions generals de variables aleatòries independents. La desigualtat més simple i potent d'aquest tipus és coneguda com a *desigualtat d'Efron-Stein*. Aquesta desigualtat fita la variància d'una funció general.

Per descriure aquesta desigualtat, sigui $\mathcal{X}$ un cert conjunt i sigui $f : \mathcal{X}^n \to \mathbb{R}$ una funció mesurable de n variables. Estem interessats en desigualtats sobre la diferència entre la variable aleatòria $Z = f(X_1, \ldots, X_n)$ i el seu valor esperat $\mathbb{E}Z$ quan $X_1, \ldots, X_n$ són variables aleatòries independents arbitràries que prenen valors en $\mathcal{X}$.

Recalquem que les variables X_i poden tenir distribucions diferents; l'única hipòtesi essencial és la independència. La desigualtat d'Efron-Stein proporciona una fita en termes de variacions locals de la funció f. Aquesta desigualtat és potser la més simple entre nombrosos resultats que mostren que les propietats de concentració podrien ser controlades estudiant el comportament local de la funció que ens interessa. A causa de la seva simplicitat, en presentem la prova a continuació.

Les desigualtats de concentració, com la desigualtat d'Efron-Stein, mostren que, en un cert sentit, la suma de variables aleatòries independents és la menys concentrada de totes les funcions generals sobre aquestes variables que són *regulars* en un sentit apropiat.

Abans d'obtenir una fita per a la variància d'una funció general Z de les variables aleatòries independents, podem adquirir una certa intuïció considerant primer el cas especial en què les variables $X_1, \ldots, X_n$ prenen valors reals i $Z = X_1 + \cdots + X_n$. En aquest cas podem utilitzar la fórmula exacta

$$\operatorname{Var}(Z) = \sum_{i=1}^{n} \operatorname{Var}(X_i) .$$

És clar que la clau d'aquesta fórmula és que la independència implica l'ortogonalitat dos a dos de les variables $X_i - \mathbb{E}X_i$. Ara, una idea natural seria fitar la variància de la funció general expressant $Z - \mathbb{E}Z$ com una suma de variables aleatòries ortogonals. Per tal de fer-ho, denotem per $\mathbb{E}_i$ l'operador esperança condicionada a $(X_1, \ldots, X_i)$ i fem servir la convenció $\mathbb{E}_0 = \mathbb{E}$. Aleshores, definim

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$$

per a cada $i = 1, \ldots, n$. Començant des de la descomposició

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i,$$

es té

$$\operatorname{Var}(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right] + 2\sum_{j>i} \mathbb{E}\left[\Delta_i \Delta_j\right].$$

Ara, si $j > i$, la igualtat $\mathbb{E}_i \Delta_j = 0$ implica que

$$\mathbb{E}_i\left[\Delta_j \Delta_i\right] = \Delta_i \mathbb{E}_i \Delta_j = 0 ,$$

i, per tant, $\mathbb{E}\left[\Delta_j \Delta_i\right] = 0$. Així doncs, obtenim

$$\operatorname{Var}(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right].$$

Fins ara no hem utilitzat el fet que Z és una funció de les variables independents $X_1, \ldots, X_n$. A causa de la independència, però,

$$\mathbb{E}_i Z = \int_{\mathcal{X}^{n-i}} f\left(X_1, \ldots, X_i, x_{i+1}, \ldots, x_n\right) d\mu_{i+1}\left(x_{i+1}\right) \ldots d\mu_n\left(x_n\right) ,$$

on, per a cada $j = 1, \ldots, n$, μ_j denota la distribució de probabilitat de X_j. Si denotem per $\mathbb{E}^{(i)}$ l'operador esperança condicionada a $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, tenim que

$$\mathbb{E}^{(i)} Z = \int_{\mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)\, d\mu_i(x_i).$$

Aleshores, pel teorema de Fubini,

$$\mathbb{E}_i\left[\mathbb{E}^{(i)} Z\right] = \mathbb{E}_{i-1} Z.$$

Havent fet aquesta observació, és fàcil deduir el següent:

2 Teorema (desigualtat d'Efron-Stein) *Siguin $X_1, \ldots, X_n$ variables aleatòries independents i sigui $Z = f(X)$ una funció de quadrat integrable de $X = (X_1, \ldots, X_n)$. Aleshores*

$$\mathrm{Var}(Z) \le \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] \stackrel{\text{def}}{=} v.$$

A més, si $X_1', \ldots, X_n'$ són còpies independents de $X_1, \ldots, X_n$ i si definim per a cada $i = 1, \ldots, n$

$$Z_i' = f\left(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\right),$$

tenim que

$$v = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)_+^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)_-^2\right],$$

on $x_+ = \max(x, 0)$ i $x_- = \max(-x, 0)$ denoten la part positiva i la negativa d'un nombre real x. També,

$$v = \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i)^2\right],$$

on l'ínfim és pres entre la classe de totes les variables Z_i, $i = 1, \ldots, n$, que són $X^{(i)}$-mesurables i de quadrat integrable.

Prova Notem que, fent servir l'observació anterior, podem escriure

$$\Delta_i = \mathbb{E}_i\left[Z - \mathbb{E}^{(i)} Z\right].$$

Per la desigualtat de Jensen, usada condicionalment,

$$\Delta_i^2 \le \mathbb{E}_i\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right].$$

Fent servir que $\mathrm{Var}(Z) \le \sum_{i=1}^n \mathbb{E}\left[\Delta_i^2\right]$, obtenim la desigualtat desitjada.

Per provar les identitats per a v, denotem per $\mathrm{Var}^{(i)}$ l'operador variància condicionada a $X^{(i)}$. Aleshores, podem escriure v com a

$$v = \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}^{(i)}(Z)\right].$$

Notem ara que es pot utilitzar simplement el fet elemental que si X i Y són variables aleatòries independents i idènticament distribuïdes que prenen valors reals, aleshores $\mathrm{Var}(X) = (1/2)\mathbb{E}[(X-Y)^2]$. Com que, condicionant $X^{(i)}$, Z_i' és una còpia independent de Z, podem escriure

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}\left[\left(Z - Z_i'\right)^2\right] = \mathbb{E}^{(i)}\left[\left(Z - Z_i'\right)_+^2\right] = \mathbb{E}^{(i)}\left[\left(Z - Z_i'\right)_-^2\right],$$

on s'ha utilitzat el fet que les distribucions condicionades de Z i Z_i' són idèntiques. L'última identitat s'obté notant que, per a qualsevol variable aleatòria real X, $\mathrm{Var}(X) = \inf_{a\in\mathbb{R}} \mathbb{E}[(X-a)^2]$. Usant aquest fet condicionalment, tenim que, per a cada $i = 1, \ldots, n$,

$$\mathrm{Var}^{(i)}(Z) = \inf_{Z_i} \mathbb{E}^{(i)}\left[(Z - Z_i)^2\right].$$

Notem que aquest ínfim és assolit sempre que $Z_i = \mathbb{E}^{(i)}Z$. □

Observem que en el cas que $Z = \sum_{i=1}^n X_i$ sigui la suma de variables aleatòries independents, aleshores la desigualtat d'Efron-Stein esdevé igualtat.

2.1 Funcions amb diferències fitades

A continuació, il·lustrarem l'ús de la desigualtat d'Efron-Stein amb alguns exemples. La més simple de les aplicacions és obtinguda si la funció és regular en el sentit que si canvia el valor d'una sola variable, el valor de la funció no pot canviar gaire. Més concretament, diem que la funció $f : \mathcal{X}^n \to \mathbb{R}$ té la *propietat de les diferències fitades* si, per algunes constants no negatives $c_1, \ldots, c_n$,

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i'\in\mathcal{X}}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \le c_i,\ 1 \le i \le n.$$

En altres paraules, si canviem la i-èsima variable de f i mantenim les altres fixades, el valor de la funció no pot variar en més de c_i. Aleshores, la desigualtat d'Efron-Stein implica el següent:

3 COROL·LARI *Si $X_1, \ldots, X_n$ són variables aleatòries independents i f té la propietat de les diferències fitades amb constants $c_1, \ldots, c_n$, aleshores $Z = f(X_1, \ldots, X_n)$ satisfà*

$$\mathrm{Var}(Z) \le \frac{1}{4}\sum_{i=1}^n c_i^2.$$

PROVA De la desigualtat d'Efron-Stein,

$$\mathrm{Var}(Z) \leq \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i)^2\right],$$

on l'ínfim es pren entre la classe de totes les Z_i, $i = 1, \ldots, n$, que són $X^{(i)}$-mesurables i de quadrat integrable. Aquí escollim

$$Z_i = \frac{1}{2}\left(\sup_{x_i' \in \mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i', X_{i+1}, \ldots, X_n) + \inf_{x_i' \in \mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i', X_{i+1}, \ldots, X_n)\right).$$

Per tant,

$$(Z - Z_i)^2 \leq \frac{c_i^2}{4},$$

d'on l'enunciat. □

Donem ara algunes aplicacions interessants d'aquest corol·lari. En tots els casos, la fita per a la variància es pot obtenir de manera senzilla, mentre que una estimació directa de la variància és força més complicada.

1 EXEMPLE (BIN PACKING) El problema del *bin packing* és un dels problemes clàssics de la investigació operativa. Donats n nombres $x_1, \ldots, x_n \in [0, 1]$, es planteja el problema següent: quin és el nombre mínim de pots (*bins*) necessaris perquè aquests nombres hi puguin ser empaquetats de tal manera que la suma de nombres a cada pot no excedeixi d'1. Sigui $f(x_1, \ldots, x_n)$ aquest nombre. Aleshores, canviant un dels x_i, el valor de $f(x_1, \ldots, x_n)$ no pot canviar en més d'una unitat; per tant, sempre que $X_1, \ldots, X_n$ siguin independents, $Z = f(X_1, \ldots, X_n)$ satisfà

$$\mathrm{Var}(Z) \leq \frac{n}{4}.$$

Aquesta fita superior no és millorable, ja que si X_i són variables aleatòries simètriques, que segueixen una distribució de Bernoulli (és a dir, $\mathbb{P}\{X_i = 0\} = \mathbb{P}\{X_i = 1\} = 1/2$), aleshores Z segueix una distribució binomial amb paràmetres n i $1/2$ i $\mathrm{Var}(Z) = n/4$. D'altra banda, algunes fites més ajustades, que depenen de la distribució dels X_i, es poden demostrar amb arguments més sofisticats (vegeu la secció 4.2).

2 EXEMPLE (SUBSEQÜÈNCIA COMUNA MÉS LLARGA) La versió més senzilla del problema de la subseqüència més llarga és la següent: siguin $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_n$ dues seqüències de llançaments de moneda. Definim Z com la llargada de la subseqüència comuna més llarga que apareix a les dues seqüències, és a dir,

$$Z = \max\{k : X_{i_1} = Y_{j_1}, \ldots, X_{i_k} = Y_{j_k}, \text{ on } 1 \leq i_1 < \cdots < i_k \leq n \text{ i } 1 \leq j_1 < \cdots < j_k \leq n\}.$$

El comportament de la $\mathbb{E}Z$ ha estat investigat en molts articles. És conegut que $\mathbb{E}Z/n$ convergeix en algun nombre γ, el valor del qual no es coneix. Està conjecturat que és $2/(1+\sqrt{2})$ i és conegut que està contingut en l'interval $[0,75796, 0,83763]$. Tot i això, nosaltres estem interessats en la concentració de la variable Z.

És senzill veure que canviar el resultat d'una sola tirada no pot canviar la llargada de la subseqüència més llarga en més d'una unitat; per tant, Z satisfà la propietat de les diferències fitades amb $c_i = 1$. Conseqüentment,

$$\mathrm{Var}(Z) \le \frac{n}{4}.$$

Per la desigualtat de Txebixev, amb alta probabilitat, Z està desviada del seu valor esperat $\sqrt{n}$, llevat de constants. En altres paraules, està força concentrada entorn de la seva mitjana, cosa que significa que els resultats de $\mathbb{E}Z$ ens revelen el comportament essencial de la llargada màxima d'una subseqüència comuna entre dues cadenes aleatòries.

3 Exemple (estimació de densitat del nucli) Una qüestió bàsica en l'estadística no paramètrica és estimar la densitat d'una variable aleatòria real. Més precisament, donada una mostra de n variables aleatòries independents i idènticament distribuïdes $X_1, \dots, X_n$, distribuïdes segons alguna densitat de probabilitat desconeguda ϕ, es vol estimar ϕ.

Un dels estimadors més estudiats és l'*estimador de densitat del nucli* definit per

$$\phi_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

on $h_n > 0$ és un paràmetre suavitzador i K és una funció no negativa amb $\int K = 1$. El comportament de l'estimador es mesura típicament amb la norma L_1:

$$Z(n) = f(X_1, \dots, X_n) = \int |\phi(x) - \phi_n(x)| dx.$$

És senzill veure que

$$\begin{aligned} |f(x_1, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)| &\le \frac{1}{nh_n} \int \left| K\left(\frac{x - x_i}{h_n}\right) - K\left(\frac{x - x_i'}{h_n}\right) \right| dx \\ &\le \frac{2}{n}, \end{aligned}$$

de manera que sense cap altra feina obtenim

$$\mathrm{Var}(Z(n)) \le \frac{2}{n}.$$

És conegut que, per a qualsevol ϕ, $\sqrt{n}\,\mathbb{E}Z(n) \to \infty$ (vegeu Devroye i Györfi [36]), la qual cosa juntament amb la desigualtat de Txebixev, que implica per a qualsevol $\epsilon > 0$

$$\mathbb{P}\left\{\left|\frac{Z(n)}{\mathbb{E}Z(n)} - 1\right| \geq \epsilon\right\} = \mathbb{P}\left\{|Z(n) - \mathbb{E}Z(n)| \geq \epsilon\mathbb{E}Z(n)\right\} \leq \frac{\mathrm{Var}(Z(n))}{\epsilon^2(\mathbb{E}Z(n))^2} \to 0$$

quan $n \to \infty$. Per tant, $Z(n)/\mathbb{E}Z(n) \to 1$ en probabilitat o, en altres paraules, $Z(n)$ és «relativament estable». Això significa que l'error L_1 aleatori es comporta com el seu valor esperat.

2.2 Funcions autofitades

Una altra propietat senzilla que se satisfà en molts exemples importants és la coneguda propietat d'*autofitació*. Diem que una funció no negativa $f : \mathcal{X}^n \to \mathbb{R}$ té la propietat d'autofitació si existeixen funcions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ tals que, per a qualsevol $x_1, \ldots, x_n \in \mathcal{X}$ i tota $i = 1, \ldots, n$,

$$0 \leq f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$$

i també

$$\sum_{i=1}^{n} (f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)) \leq f(x_1, \ldots, x_n).$$

Per a funcions autofitades, clarament tenim que

$$\sum_{i=1}^{n} (f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n))^2 \leq f(x_1, \ldots, x_n).$$

I aleshores l'última expressió de v en el teorema 2 implica el següent:

4 COROL·LARI *Si f té la propietat d'autofitació, aleshores*

$$\mathrm{Var}(Z) \leq \mathbb{E}Z.$$

A continuació, mencionem algunes aplicacions d'aquest simple corol·lari. Resulta que en molts casos la fita obtinguda millora significativament la que hauríem obtingut utilitzant simplement el corol·lari 3.

Una classe important de funcions que satisfan la propietat d'autofitació consisteix en les anomenades «funcions de configuració».

Suposem que tenim una propietat P definida sobre la unió de productes finits d'un conjunt $\mathcal{X}$, és a dir, una seqüència de conjunts $P_1 \subset \mathcal{X}, P_2 \subset \mathcal{X} \times \mathcal{X}, \ldots, P_n \subset \mathcal{X}^n$. Diem que $(x_1, \ldots, x_m) \in \mathcal{X}^m$ satisfà la propietat P si $(x_1, \ldots, x_m) \in P_m$. Suposem que P és *hereditària* en el sentit que si $(x_1, \ldots, x_m)$ satisfà P, aleshores també ho fa qualsevol subseqüència $(x_{i_1}, \ldots, x_{i_k})$ de $(x_1, \ldots, x_m)$. La funció f que a un vector $x = (x_1, \ldots, x_n)$ li assigna la mida de la subseqüència més llarga que satisfà P és la *funció de configuració* associada amb la propietat P.

El corol·lari 4 implica el resultat següent:

5 COROL·LARI *Sigui f una funció de configuració i sigui $Z = f(X_1, \dots, X_n)$, on $X_1, \dots, X_n$ són variables aleatòries independents. Aleshores,*

$$\operatorname{Var}(Z) \leq \mathbb{E}Z.$$

PROVA Pel corol·lari 4 n'hi ha prou amb demostrar que qualsevol funció de configuració és autofitada. Sigui $Z_i = f(X^{(i)}) = f(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. La condició $0 \leq Z - Z_i \leq 1$ se satisfà trivialment. D'altra banda, suposem que $Z = k$ i sigui $\{X_{i_1}, \dots, X_{i_k}\} \subset \{X_1, \dots, X_n\}$ una subseqüència de llargada k tal que $f_k(X_{i_1}, \dots, X_{i_k}) = k$. (Noteu que, segons la definició de funció de configuració, aquesta subseqüència existeix.) Clarament, si l'índex i és tal que $i \notin \{i_1, \dots, i_k\}$, llavors $Z = Z_i$, i per tant,

$$\sum_{i=1}^{n} (Z - Z_i) \leq Z$$

també se satisfà, cosa que completa la demostració. □

Per il·lustrar el fet que les funcions de configuració apareixen de manera natural en diferents aplicacions, presentem alguns exemples procedents de camps diversos.

4 EXEMPLE (NOMBRE DE VALORS DIFERENTS EN UNA MOSTRA DISCRETA) Siguin $X_1, \dots, X_n$ variables aleatòries independents, idènticament distribuïdes, que prenen valors en el conjunt d'enters positius amb $\mathbb{P}\{X_i = k\} = p_k$. Denotem per $Z(n)$ el nombre de valors diferents presos per aquestes n variables. Aleshores, podem escriure

$$Z(n) = \sum_{i=1}^{n} \mathbb{1}_{\{\{X_i \neq X_1, \dots, X_i \neq X_{i-1}\}\}},$$

de manera que el valor esperat de $Z(n)$ pot ser calculat de manera senzilla:

$$\mathbb{E}Z(n) = \sum_{i=1}^{n} \sum_{j=1}^{\infty} (1 - p_j)^{i-1} p_j.$$

És clar que $\mathbb{E}[Z(n)]/n \to 0$ quan $n \to \infty$. Però la pregunta és: com n'està de concentrada la distribució $Z(n)$? Com que $Z(n)$ satisfà la propietat de les diferències fitades amb $c_i = 1$, el corol·lari 3 implica $\operatorname{Var}(Z(n)) \leq n/4$ i, per tant, $Z(n)/n \to 0$ amb alta probabilitat per la desigualtat de Txebixev. D'altra banda, és obvi que $Z(n)$ és una funció de configuració associada a la propietat de *distintivitat*, i pel corol·lari 5 tenim que

$$\operatorname{Var}(Z(n)) \leq \mathbb{E}Z(n)$$

que és una millora important, ja que $\mathbb{E}Z(n) = o(n)$.

5 Exemple (dimensió VC) Una de les quantitats combinatòries centrals en la teoria de processos empírics i en la teoria d'aprenentatge estadístic és la *dimensió de Vapnik-Chervonenkis*. Sigui $\mathcal{A}$ una col·lecció arbitrària de subconjunts de $\mathcal{X}$ i sigui $x = (x_1, \dots, x_n)$ un vector de n punts de $\mathcal{X}$. Definim la *traça* de $\mathcal{A}$ en x com

$$\mathrm{tr}(x) = \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{A}\}\,.$$

El *coeficient d'esmicolament* (o *funció de creixement de Vapnik-Chervonenkis*) de $\mathcal{A}$ en x és $T(x) = |\mathrm{tr}(x)|$, la mida de la traça. $T(x)$ és el nombre de subconjunts diferents de conjunts de punts $\{x_1, \dots, x_n\}$ generats per la intersecció d'aquest conjunt amb els elements de $\mathcal{A}$. Un subconjunt $\{x_{i_1}, \dots, x_{i_k}\}$ de $\{x_1, \dots, x_n\}$ es diu que està *esmicolat* si $2^k = T(x_{i_1}, \dots, x_{i_k})$. La *dimensió VC* $D(x)$ de $\mathcal{A}$ (respecte de x) és la mida k del subconjunt esmicolat més gran de x. De la definició és obvi que $f(x) = D(x)$ és una funció de configuració (associada a la propietat d'*esmicolament*) i, aleshores, si $X_1, \dots, X_n$ són variables aleatòries independents, tenim que

$$\mathrm{Var}(D(X)) \le \mathbb{E}D(X).$$

Es pot provar també que $\log_2 T(X)$ és autofitada, cosa que implica que $\mathrm{Var}(\log_2 T(X)) \le \mathbb{E}\log_2 T(X)$.

6 Exemple (subseqüències creixents) Considerem un vector $x = (x_1, \dots, x_n)$ de n nombres diferents en l'interval $[0,1]$. Els enters positius $i_1 < i_2 < \cdots < i_m$ formen una *subseqüència creixent* si $x_{i_1} < x_{i_2} < \cdots < x_{i_m}$ (on $i_1 \ge 1$ i $i_m \le n$). Denotem amb $L(x)$ la llargada de la subseqüència creixent més llarga. Clarament, $L(x)$ és una funció de configuració (associada amb la propietat de la *seqüència creixent*) i, aleshores, si $X_1, \dots, X_n$ són variables aleatòries independents tals que són diferents amb probabilitat 1 (això està justificat si cada X_i té una distribució absolutament contínua), aleshores $\mathrm{Var}(L(X)) \le \mathbb{E}L(X)$. Si els X_i estan distribuïts uniformement per $[0,1]$, aleshores és conegut que $\mathbb{E}L(X) \sim 2\sqrt{n}$. La fita obtinguda per la variància resulta massa feble, l'ordre correcte és $\mathrm{Var}(L(X)) = O(n^{1/3})$, un resultat difícil.

En una variació del problema, $X_1, \dots, X_n$ prenen els seus valors en un conjunt finit $\{1, \dots, m\}$. Definim ara $L^{(m)}(X)$ com la longitud de la subseqüència creixent més llarga de $X = (X_1, \dots, X_n)$, és a dir, l'enter positiu més gran k pel qual existeixen $1 \le i_1 < \cdots < i_k \le n$ tals que $X_{i_1} \le X_{i_2} \le \ldots \le X_{i_k}$. L'anàlisi de la variància segueix sense canvis i, com abans, tenim que $\mathrm{Var}(L^{(m)}(X)) \le \mathbb{E}L^{(m)}(X)$. Aquesta estimació té ara l'ordre de magnitud correcte, ja que és conegut que si les X_i estan uniformement distribuïdes, aleshores $(L^{(m)}(X) - n/m)/\sqrt{2n/m}$ convergeix en distribució a una variable aleatòria la distribució de la qual depèn de m.

2.3 Més exemples

7 Exemple (valor propi més gran d'una matriu simètrica aleatòria) Sigui A una matriu real simètrica en la qual els coeficients $X_{i,j}$, $1 \leq i \leq j \leq n$ són variables aleatòries independents amb valor absolut fitat per 1. Sigui $Z = \lambda_1$ el valor propi més gran d'A. La propietat d'aquest valor propi que necessitem per fitar-ne la variància és que si $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ és el vector propi associat a λ_1 amb $\|v\| = 1$, aleshores

$$\lambda_1 = v^T A v = \sup_{u:\|u\|=1} u^T A u.$$

Per utilitzar el teorema 2, considerem la matriu simètrica $A'_{i,j}$, obtinguda substituint $X_{i,j}$ en A per una còpia independent $X'_{i,j}$, mantenint les altres variables fixades. Sigui $Z'_{i,j}$ el valor propi més gran d'aquesta matriu. Per la propietat del valor propi més gran anteriorment esmentada,

$$\begin{aligned}
(Z - Z'_{i,j})_+ &\leq \left(v^T A v - v^T A'_{i,j} v\right) \mathbb{1}_{\{Z > Z'_{i,j}\}} \\
&= \left(v^T (A - A'_{i,j}) v\right) \mathbb{1}_{\{Z > Z'_{i,j}\}} = \left(v_i v_j (X_{i,j} - X'_{i,j})\right)_+ \\
&\leq 2|v_i v_j|.
\end{aligned}$$

Aleshores,

$$\sum_{1 \leq i \leq j \leq n} (Z - Z'_{i,j})_+^2 \leq \sum_{1 \leq i \leq j \leq n} 4|v_i v_j|^2 \leq 4 \left(\sum_{i=1}^n v_i^2\right)^2 = 4.$$

Prenent esperances a banda i banda, i utilitzant la desigualtat d'Efron-Stein, tenim que $\mathrm{Var}(Z) \leq 4$. Així doncs, la variancia està fitada per una constant independentment de la mida de la matriu i de la distribució dels coeficients. L'única condició que necessitem és la independència i la fitació de les variables $X_{i,j}$. De fet, no cal que tinguin la mateixa distribució. Aquesta mateixa prova funciona també per al valor propi més petit.

8 Exemple (arbre generador de pes mínim) Considerem la variable aleatòria T_m, definida com la suma dels pesos d'un arbre generador mínim del graf complet K_m, amb pesos $X_{i,j}$ ($1 \leq i < j \leq m$) sobre les arestes, distribuïts uniformement i independent en l'interval $[0, 1]$. (Així doncs, T_m és una funció de $n = \binom{m}{2}$ variables aleatòries independents.) És conegut que el valor esperat de T_m convergeix a la constant $\zeta(3) = \sum_{i=1}^{\infty} i^{-3}$. Ara, fitem la variància de T_m. Utilitzant les desigualtats d'Efron-Stein directament obtenim una fita no gaire bona, però un simple truc ens portarà a un valor més proper al real. Aquesta idea és que el pes més gran d'una aresta en l'arbre generador mínim no pot ser gaire gran, com a molt, d'ordre $\log m/m$, amb alta probabilitat. Aquesta observació ens permet canviar T_m per la variable aleatòria $\overline{T}_m$ obtinguda quan

els $X_{i,j}$ són reemplaçats per $\min(X_{i,j}, \delta_m)$, on $\delta_m > 0$ és un nombre positiu petit. Notem que si $\delta_m = c \log m/m$ per a alguna $c > 1$, aleshores $T_m = \overline{T}_m$ amb alta probabilitat. Per tal de veure-ho, observem que $T_m \neq \overline{T}_m$ implica que el pes més gran d'una aresta de l'arbre generador mínim sigui més gran que δ_m. Però això és la probabilitat que un graf aleatori amb el model d'Erdős-Rényi $G(m, \delta_m)$ (és a dir, un graf amb m vèrtexs en el qual cada una de les $\binom{m}{2}$ arestes existeix amb probabilitat δ_m de manera independent) no sigui connex. En la teoria de grafs aleatoris és ben conegut que aquesta probabilitat és, com a molt, de $2\left(e^{m(1-c)/2} - 1\right) + 2^{m+1} m^{-(c-1)m/4}$, que està fitat per $4m^{-c/4}$, si $c \geq 2$. Com que

$$\begin{aligned}
\mathrm{Var}(T_m) &= \mathbb{E}\left[T_m^2\right] - (\mathbb{E}T_m)^2 \\
&= \mathbb{E}\left[T_m^2 \mathbb{1}_{\{T_m = \overline{T}_m\}}\right] + \mathbb{E}\left[T_m^2 \mathbb{1}_{\{T_m \neq \overline{T}_m\}}\right] - (\mathbb{E}\overline{T}_m)^2 \\
&\leq \mathrm{Var}(\overline{T}_m) + m^2 \mathbb{P}\left\{T_m \neq \overline{T}_m\right\} \\
&\leq \mathrm{Var}(\overline{T}_m) + 4m^{2-c/4},
\end{aligned}$$

en tenim prou fitant la variància de $\overline{T}_m$. Utilitzem aquí la variant següent de la desigualtat d'Efron-Stein:

$$\mathrm{Var}(\overline{T}_m) \leq \mathbb{E} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \left(\overline{T}_m - \overline{T}'_{m,(i,j)}\right)_-^2,$$

on $\overline{T}'_{m,(i,j)}$ s'obté reemplaçant $X_{i,j}$ per una còpia independent. Clarament, si es canvia el pes $X_{i,j}$, aleshores $\overline{T}_m$ només pot decréixer si l'aresta (i,j) forma part de l'arbre generador mínim. Com que tenim $m-1$ d'aquestes arestes i aquest canvi no pot ser superior a δ_m,

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \left(\overline{T}_m - \overline{T}'_{m,(i,j)}\right)_-^2 \leq m\delta_m^2.$$

En resum,

$$\mathrm{Var}(T_m) \leq m\delta_m^2 + 4m^{2-c/4} = \frac{144 \log^2 m}{m} + \frac{4}{m},$$

on triem $c = 12$. Aquesta fita no dóna l'ordre correcte, ja que és conegut que, asimptòticament , $\mathrm{Var}(T_m) \sim (6\zeta(4) - 4\zeta(3))/m$. (De fet, $\sqrt{m}(T_m - \zeta(3))$ convergeix, en distribució, a una variable aleatòria normal centrada amb variància $6\zeta(4) - 4\zeta(3)$.) Tot i això, aquest argument il·lustra com la desigualtat d'Efron-Stein pot ser utilitzada d'una manera molt senzilla per obtenir desigualtats no asimptòtiques molt potents.

Notes bibliogràfiques

La desigualtat d'Efron-Stein rep el nom de l'article d'Efron i Stein [39]. Steele [92] i Rhee i Talagrand [86] van obtenir-ne versions millorades i la versió presentada en el teorema 2. La demostració desenvolupada aquí apareix a [86].

El comportament de $Z = f(X_1, \dots, X_n)$ en el problema de: *bin packing*, on $X_1, \dots, X_n$ són variables aleatòries independents, ha estat extensament estudiat vegeu, per exemple, Rhee i Talagrand [87], Rhee [85] i Talagrand [94].

El problema de la subseqüència més llarga ha estat estudiat intensament durant els últims trenta anys: vegeu Chvátal i Sankoff [24], Deken [31], Dančík i Paterson [29] i Steele [91, 93]. Aquesta va ser una de les primeres aplicacions de la desigualtat d'Efron-Stein (vegeu Steele [92]), on la potència i la simplicitat de la desigualtat es demostren clarament.

Les funcions de configuració van ser definides per Talagrand [94, Secció 7]. La nostra definició està presa de [17] i és una lleu modificació de la donada per Talagrand.

L'estabilitat relativa de l'error L_1 de l'estimació de densitat del nucli és deguda a Devroye [33, 34]. Per a més informació sobre el comportament de l'error L_1 de l'estimació de densitat del nucli, recomanem al lector Devroye i Györfi [36] i Devroye i Lugosi [38].

Les propietats de concentració de les funcions autofitades han estat estudiades per Boucheron, Lugosi i Massart [17]; Rio [88]; Bousquet [20]; Maurer [71], i McDiarmid i Reed [76].

La dimensió de Vapnik-Chervonenkis i la funció de creixement s'anomenen així en honor de Vapnik i Chervonenkis [102, 103] (vegeu també Blumer *et al.* [12]; Devroye, Györfi i Lugosi [37]; Anthony i Bartlett [6], i Vapnik [101].

El fet que la subseqüència creixent més llarga en una permutació aleatòria de n nombres satisfà $\mathbb{E}L(X) \sim 2\sqrt{n}$ és degut a Logan i Shepp [63] (vegeu també Hammersley [45], Aldous i Diaconis [3] i Groeneboom [43]. El famós article de Baik, Deift i Johansson [7] estableix la distribució límit de $L(X)$. Aquest resultat implica que $\mathrm{Var}(L(X)) = O(n^{1/3})$. Ledoux [60] obté una desigualtat no asimptòtica i exponencial per les cues que té la forma correcta. Per als resultats anteriors sobre la concentració de $L(X)$, vegeu Frieze [41], Bollobás i Brightwell [14] i Talagrand [94].

El teorema del límit per a la subseqüència creixent més llarga en una paraula aleatòria sobre un alfabet finit mencionat a l'exemple 6 és degut a Tracy i Widom [98] i Johansson [51] (vegeu també Its, Tracy i Widom [48]).

L'argument per fitar la variància del valor propi més gran d'una matriu simètrica aleatòria està basat en Alon, Krivelevich i Vu [4], que proven una fita exponencial per la cua que reproduirem més endavant. El fenomen que la variància sigui fitada asimptòticament ja havia estat descobert per Füredi i Komlós [42], que van provar un teorema per al límit del valor propi més gran.

El valor asimptòtic $\lim_{m\to\infty} \mathbb{E}T_m = \zeta(3)$ del pes esperat de l'arbre generador mínim va ser determinat per Frieze [40]. El teorema del límit mencionat a l'exemple 8 és degut a Janson [49] i Wästlund [104].

3 El mètode de l'entropia

En la secció anterior, hem vist que la desigualtat d'Efron-Stein és una eina potent per fitar la variància de funcions de variables aleatòries independents. Aleshores, via la desigualtat de Txebixev, es poden obtenir fàcilment fites per a les cues de les distribucions d'aquestes funcions. Com en el cas de les sumes de variables aleatòries independents, però, les fites que s'obtenen basant-se en la variància són sovint poc satisfactòries i es poden obtenir millores substancials. L'objectiu d'aquesta secció és exposar una metodologia que permet obtenir fites exponencials en molts casos. La cerca d'aquest tipus de desigualtats ha estat un tema cabdal en teoria de probabilitat a les últimes dècades. En aquesta secció, donarem les idees principals que hi ha darrere del que s'anomena *mètode de l'entropia*, una tècnica natural però potent. Aquest mètode és potser la manera més simple d'obtenir anàlegs exponencials de la desigualtat d'Efron-Stein.

Per tal d'introduir les idees principals, recordem primer algunes nocions bàsiques de la teoria de la informació. Per mantenir el material a un nivell elemental, ens limitem només a variables aleatòries discretes. L'extensió a distribucions arbitràries és immediata però més aviat tècnica.

3.1 Teoria de la informació bàsica

En aquesta secció, resumim algunes propietats bàsiques de l'entropia d'una variable aleatòria discreta. Una bona referència d'introducció a la teoria de la informació és el llibre de Cover i Thomas [26].

Sigui X una variable aleatòria que pren valors en un conjunt numerable $\mathcal{X}$ amb distribució $\mathbb{P}\{X = x\} = p(x)$, $x \in \mathcal{X}$. L'*entropia* (de Shannon) de X es defineix com

$$H(X) = \mathbb{E}[-\log p(X)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

(on log denota el logaritme natural i $0\log 0 = 0$). Si (X, Y) és un parell de variables aleatòries discretes que pren valors a $\mathcal{X}\times\mathcal{Y}$ aleshores l'*entropia conjunta* $H(X, Y)$ de X i Y és l'entropia del parell (X, Y). L'*entropia condicionada* $H(X|Y)$ es defineix com

$$H(X|Y) = H(X, Y) - H(Y).$$

Observem que si escrivim $p(x, y) = \mathbb{P}\{X = x, Y = y\}$ i $p(x|y) = \mathbb{P}\{X = x|Y = y\}$, aleshores

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x, y)\log p(x|y),$$

d'on veiem que $H(X|Y) \geq 0$. També és fàcil veure que la relació que defineix l'entropia condicionada es manté com a certa quan es condiciona condicionar, és a dir, per a qualsevol de les tres variables aleatòries discretes X, Y, Z,

$$H(X, Y|Z) = H(Y|Z) + H(X|\, Y, Z).$$

(Simplement afegim $H(Z)$ a les dues bandes de la igualtat i fem servir la definició de l'entropia condicional.) Iterant la darrera identitat obtenim la *regla de la cadena* per a l'entropia: per a variables aleatòries discretes arbitràries $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}).$$

Siguin P i Q dues distribucions de probabilitat sobre un conjunt numerable $\mathcal{X}$ amb funcions de probabilitat p i q. La *divergència de Kullback-Leibler*, o *entropia relativa* de P i Q és

$$D(P\|Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Com que $\log x \leq x - 1$,

$$D(P\|Q) = -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x \in \mathcal{X}} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) = 0,$$

de manera que l'entropia relativa és sempre no negativa, i és zero si i només si $P = Q$. Aquest simple fet té algunes conseqüències interessants. Per exemple, si $\mathcal{X}$ és un conjunt finit amb N elements, X és una variable aleatòria amb distribució P i prenem Q com la distribució uniforme sobre $\mathcal{X}$, aleshores $D(P\|Q) = \log N - H(X)$ i, per tant, l'entropia de X no és mai més gran que el logaritme del cardinal del seu domini.

Considerem un parell de variables aleatòries X, Y amb distribució conjunta $P_{X,Y}$ i distribucions marginals P_X i P_Y. Observant que $D(P_{X,Y}\|P_X \times P_Y) = H(X) - H(X|Y)$, la no-negativitat de l'entropia relativa implica que $H(X) \geq H(X|Y)$, és a dir, condicionant es redueix l'entropia. És igualment fàcil de veure que aquest fet segueix sent cert per a entropies condicionades, és a dir,

$$H(X|Y) \geq H(X|Y, Z).$$

Ara, podem provar la desigualtat següent de Han [46]:

6 Teorema (desigualtat de Han) *Siguin $X_1, \ldots, X_n$ variables aleatòries discretes. Aleshores,*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Prova Per a cada $i = 1, \ldots, n$, d'acord amb la definició d'entropia condicionada i amb el fet que condicionar redueix l'entropia,

$$\begin{aligned} &H(X_1, \ldots, X_n) \\ &\quad = H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\quad \leq H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}) \quad i = 1, \ldots, n. \end{aligned}$$

Sumant aquestes n desigualtats i fent servir la regla de la cadena per a l'entropia, obtenim

$$nH(X_1,\ldots,X_n) \le \sum_{i=1}^{n} H(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) + H(X_1,\ldots,X_n),$$

que és el que volíem provar. □

Acabem aquesta secció amb una desigualtat que es pot veure com una versió de la desigualtat de Han per a entropies relatives. Com va observar Massart [70], aquesta desigualtat es pot fer servir per provar la desigualtat clau de *tensorialització*, crucial per als nostres propòsits.

Amb aquest objectiu, considerem un conjunt numerable $\mathcal{X}$ i siguin P i Q distribucions de probabilitat a $\mathcal{X}^n$, on $P = P_1 \times \cdots \times P_n$ és una mesura producte. Denotem els elements de $\mathcal{X}^n$ per $\boldsymbol{x} = (x_1,\ldots,x_n)$ i escrivim $\boldsymbol{x}^{(i)} = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ per denotar el vector $(n-1)$-dimensional obtingut per la supressió de la component i-èsima de $\boldsymbol{x}$. Denotem per $Q^{(i)}$ i $P^{(i)}$ les distribucions marginals de $\boldsymbol{x}$ d'acord amb Q i P, és a dir,

$$Q^{(i)}(\boldsymbol{x}) = \sum_{x\in\mathcal{X}} Q(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_n)$$

i

$$\begin{aligned}P^{(i)}(\boldsymbol{x}) &= \sum_{x\in\mathcal{X}} P(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_n)\\ &= \sum_{x\in\mathcal{X}} P_1(x_1)\cdots P_{i-1}(x_{i-1})P_i(x)P_{i+1}(x_{i+1})\cdots P_n(x_n).\end{aligned}$$

Aleshores, basant-se en la desigualtat de Han, es pot provar fàcilment el següent:

7 Teorema (desigualtat de Han per a entropies relatives)

$$D(Q\|P) \ge \frac{1}{n-1}\sum_{i=1}^{n} D(Q^{(i)}\|P^{(i)})$$

o, de manera equivalent,

$$D(Q\|P) \le \sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right).$$

3.2 Tensorialització de l'entropia

Ara estem preparats per presentar algunes desigualtats de concentració exponencials. Com en la secció 2, considerem variables aleatòries independents $X_1,\ldots,X_n$ i investiguem propietats de concentració de $Z = f(X_1,\ldots,X_n)$. La base del mètode d'entropia és una extensió molt potent de la desigualtat d'Efron-Stein. Observem que la desigualtat d'Efron-Stein es pot reescriure com

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i(Z^2) - (\mathbb{E}_i(Z))^2\right]$$

o, escrivint $\phi(x) = x^2$,

$$\mathbb{E}\phi(Z) - \phi(\mathbb{E}Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Z) - \phi(\mathbb{E}_i(Z))\right].$$

Resulta que aquesta desigualtat segueix sent certa per a una àmplia classe de funcions convexes ϕ. El cas que considerem aquí és $\phi(x) = x \log x$. En aquest cas, com es veu en la prova que segueix, la part esquerra de la desigualtat es pot escriure com l'entropia relativa entre la distribució induïda per Z en $\mathcal{X}^n$ i la distribució de X_1^n. D'aquí, el nom *tensorialització de la desigualtat de l'entropia* (vegeu, per exemple, Ledoux [57]).

8 TEOREMA *Sigui $\phi(x) = x \log x$ per a $x > 0$. Siguin $X_1, \ldots, X_n$ variables aleatòries independents que prenen valors en $\mathcal{X}$ i sigui f una funció positiva en $\mathcal{X}^n$. Amb $Y = f(X_1, \ldots, X_n)$, tenim*

$$\mathbb{E}\phi(Y) - \phi(\mathbb{E}Y) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Y) - \phi(\mathbb{E}_i(Y))\right].$$

PROVA Només provem l'enunciat per a variables aleatòries discretes $X_1, \ldots, X_n$. L'extensió al cas general és tècnica però immediata. El teorema és una conseqüència directa de la desigualtat de Han per a entropies relatives. Observem primer que si la desigualtat és certa per a una variable aleatòria Y, aleshores també ho és per a cY, on c és una constant positiva. Per tant, podem suposar que $\mathbb{E}Y = 1$. Definim una mesura de probabilitat Q a $\mathcal{X}^n$ com

$$Q(\boldsymbol{x}) = f(\boldsymbol{x})P(\boldsymbol{x}),$$

on P denota la distribució de $X_1^n = X_1, \ldots, X_n$. Aleshores, és clar que

$$\mathbb{E}\phi(Y) - \phi(\mathbb{E}Y) = \mathbb{E}[Y \log Y] = D(Q\|P),$$

que, pel teorema 7, no és més gran que $\sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right)$. Un càlcul directe, però, mostra que

$$\sum_{i=1}^{n}\left(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\right) = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_i\phi(Y) - \phi(\mathbb{E}_i(Y))\right]$$

i en segueix l'enunciat. □

La quantitat de l'esquerra en la desigualtat del teorema 8 s'anomena sovint l'*entropia* de la variable aleatòria no negativa Y. Es denota per

$$\mathrm{Ent}(Y) = \mathbb{E}[Y \log Y] - (\mathbb{E}Y) \log \mathbb{E}Y.$$

Aquesta quantitat està estretament relacionada amb l'entropia de Shannon, però no s'hi ha de confondre. La desigualtat tensorialitzada de l'entropia es pot reescriure com

$$\mathrm{Ent}(Y) \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Ent}^{(i)}(Y)\right],$$

on $\mathrm{Ent}^{(i)}(Y) = \mathbb{E}^{(i)}[Y \log Y] - \left(\mathbb{E}^{(i)}Y\right) \log\left(\mathbb{E}^{(i)}Y\right)$. (Recordem que $\mathbb{E}^{(i)}$ denota l'esperança condicionada a $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.)

3.3 L'argument de Herbst i concentració gaussiana

La idea principal del mètode per provar desigualtats de concentració és aplicar el teorema 8 a la variable aleatòria positiva $Y = e^{\lambda Z}$. Aleshores, denotant per $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$ la funció generadora de moments de Z, la part esquerra de la desigualtat del teorema 8 esdevé

$$\mathrm{Ent}(e^{\lambda Z}) = \lambda \mathbb{E}\left[Z e^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] = \lambda F'(\lambda) - F(\lambda) \log F(\lambda).$$

La nostra estratègia és obtenir fites superiors de $\mathrm{Ent}(e^{\lambda Z})$, que per la identitat anterior, es poden convertir en desigualtats diferencials per a la funció generadora de moments $F(\lambda)$. En molts casos, aquestes desigualtats es poden transformar en fites superiors per a $F(\lambda)$ i aleshores és fàcil obtenir fites per a les cues de distribució via les tècniques de fitació de Chernoff. Aquesta és la idea principal del que es coneix a la literatura com a *argument de Herbst*.

Originalment, l'argument de Herbst va ser aplicat per provar desigualtats de concentració per a funcions de variables aleatòries independents gaussianes. Això era possible gràcies a una elegant fita superior de l'entropia del quadrat d'una tal variable aleatòria coneguda com a *desigualtat logarítmica gaussiana de Sobolev*:

9 Teorema (desigualtat logarítmica gaussiana de Sobolev) *Sigui* $X = (X_1, \ldots, X_n)$ *un vector de* n *variables aleatòries independents gaussianes normalitzades. Sigui* $f : R^n \to \mathbb{R}$ *una funció de classe* C^1. *Aleshores,*

$$\mathrm{Ent}\left(f^2\right) \leq 2\, \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

Aquest resultat, via l'argument de Herbst, porta a desigualtats exponencials de cues de distribució per a funcions suaus de variables aleatòries gaussianes independents. El resultat bàsic és la clàssica desigualtat de concentració gaussiana següent:

10 Teorema (desigualtat de Tsirelson) *Sigui* $X = (X_1, \ldots, X_n)$ *un vector de* n *variables aleatòries independents gaussianes normalitzades. Siguin* $L > 0$ *i* $f : \mathbb{R}^n \to \mathbb{R}$ *una funció diferenciable amb gradient uniformement fitat per* L*. Aleshores, per a tot* $\lambda \in \mathbb{R}$*,*

$$\log \mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}f(X))}\right] \leq \frac{\lambda^2}{2}L^2.$$

A més, per a tot $t > 0$*,*

$$\mathbb{P}\left\{f(X) - \mathbb{E}f(X) \geq t\right\} \leq e^{-\frac{t^2}{2L^2}}.$$

Prova Podem suposar sense pèrdua de generalitat que $\mathbb{E}f(X) = 0$. Fent servir la desigualtat logarítmica gaussiana de Sobolev per a la funció $e^{\lambda f/2}$, obtenim

$$\begin{aligned} \mathrm{Ent}\left(e^{\lambda f}\right) &\leq 2\mathbb{E}\left\|\nabla e^{\lambda f(X)/2}\right\|^2 \\ &= \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda f(X)}\|\nabla f(X)\|^2\right] \\ &\leq \frac{\lambda^2 L^2}{2}\mathbb{E}\left[e^{\lambda f(X)}\right]. \end{aligned}$$

Si escrivim $F(\lambda) = \mathbb{E}\left[e^{\lambda f(X)}\right]$, obtenim la desigualtat diferencial

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \frac{\lambda^2 L^2}{2}F(\lambda).$$

Per resoldre-la, dividim les dues bandes pel nombre positiu $\lambda^2 F(\lambda)$. Definint $G(\lambda) = \log F(\lambda)$, observem que la banda esquerra és precisament la derivada de $G(\lambda)/\lambda$. Així doncs, obtenim la desigualtat

$$\left(\frac{G(\lambda)}{\lambda}\right)' \leq \frac{L^2}{2}.$$

Per la regla de L'Hôpital, obtenim que $\lim_{\lambda\to 0} G(\lambda)/\lambda = F'(0)/F(0) = \mathbb{E}Z$. Si $\lambda > 0$, integrant de 0 a λ, obtenim $G(\lambda)/\lambda \leq \mathbb{E}Z + \lambda L^2/2$ o, en altres paraules,

$$F(\lambda) \leq e^{\lambda\mathbb{E}Z + \lambda^2 L^2/2}.$$

com havíem enunciat. Finalment, per la desigualtat de Markov,

$$\mathbb{P}\left\{Z > \mathbb{E}Z + t\right\} \leq \inf_{\lambda>0} F(\lambda)e^{-\lambda\mathbb{E}Z-\lambda t} \leq \inf_{\lambda>0} e^{\lambda^2 L^2/2-\lambda t} = e^{-t^2/(2L^2)}.$$

□

Una característica important del teorema de Tsirelson és que la banda dreta de la desigualtat no depèn de la dimensió n. Aquesta desigualtat ha servit de referència per al desenvolupament de desigualtats de concentració en les tres últimes dècades. El següent exemple n'il·lustra una aplicació prototípica i important:

9 Exemple (norma d'un vector gaussià) Sigui $X = (X_1, \dots, X_n)$ un vector gaussià amb esperança zero i matriu de covariància Γ. Sigui $p \geq 1$ i considerem la variable aleatòria real definida per la p-norma de X, és a dir,

$$Z = \|X\|_p = \left(\sum_{i=1}^{n} |X_i|^p \right)^{1/p}.$$

Com que Γ és definida com a no negativa, hi ha una matriu, A, $n \times n$, que satisfà $A^T A = \Gamma$. Aleshores, el vector gaussià X es distribueix com a AY, on $Y = (Y_1, \dots, Y_n)$ i les components de Y són variables aleatòries gaussianes normalitzades i independents. Ara, $f(y) = \|Ay\|_p$ és una funció de Lipschitz de $\mathbb{R}^n$ a $\mathbb{R}$ amb constant de Lipschitz L igual a l'operador norma de A que envia ℓ_2 a ℓ_p, és a dir,

$$L = \|A\|_{\ell_2 \to \ell_p} \stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^n : \|y\|_2 = 1} \|Ay\|_p.$$

Aleshores, per la desigualtat de Tsirelson, tenim, per a tot $t > 0$,

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geq t\} \leq 2e^{-t^2/(2L^2)}.$$

3.4 Concentració de funcions de variables aleatòries independents

L'argument esbossat en la secció anterior per provar la desigualtat de concentració gaussiana en el teorema 10 es pot generalitzar per incloure funcions de variables aleatòries independents arbitràries. L'únic pas en el qual es feia servir que la distribució fos gaussiana era en la desigualtat logarítmica de Sobolev del teorema 9. Provant *desigualtats logarítmiques de Sobolev modificades* convenientment, però, es poden obtenir resultats més generals. En aquesta secció, presentem una mostra d'aquestes desigualtats. El resultat següent, que es dedueix fàcilment del teorema 8, és un bon punt per començar.

Com en tota la secció, considerem variables aleatòries independents $X_1, \dots, X_n$ que prenen valors en un cert espai $\mathcal{X}$, una funció $f : \mathcal{X}^n \to \mathbb{R}$ i la variable aleatòria $Z = f(X_1, \dots, X_n)$. Recordem la notació

$$Z_i = f_i(X^{(i)}) = f_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n),$$

on $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ és una funció arbitrària.

11 Teorema (una desigualtat logarítmica de Sobolev modificada) *Sigui $\phi(x) = e^x - x - 1$. Per a qualsevol $\lambda \in \mathbb{R}$, se satisfà*

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \phi\left(-\lambda(Z - Z_i)\right)\right].$$

La desigualtat de concentració exponencial que segueix és fàcil d'obtenir amb la desigualtat logarítmica de Sobolev modificada i, tot i això, resulta tenir nombroses aplicacions interessants. La seva prova és essencialment idèntica a la de la desigualtat de Tsirelson, però, gràcies a la generalitat del teorema 11, no cal que ens limitem a funcions de variables aleatòries de Bernoulli.

12 TEOREMA *Considerem una funció real de n variables aleatòries independents* $Z = f(X_1, \ldots, X_n)$ *i sigui* $Z_i = \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n)$.

Suposem que hi ha una constant positiva v tal que, gairebé segur,

$$\sum_{i=1}^{n} (Z - Z_i)^2 \leq v.$$

Aleshores, per a tot $t > 0$,

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} \leq e^{-t^2/(2v)}.$$

PROVA El resultat es dedueix fàcilment de la desigualtat logarítmica de Sobolev modificada i de l'argument de Herbst. Observem que, per a $x > 0$, $\phi(-x) \leq \frac{x^2}{2}$, i, per tant, per a tot $\lambda > 0$, el teorema 11 implica

$$\begin{aligned}\lambda\mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log\mathbb{E}\left[e^{\lambda Z}\right] &\leq \mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}\frac{\lambda^2}{2}(Z - Z_i)^2\right] \\ &\leq \frac{\lambda^2 v}{2}\mathbb{E}\left[e^{\lambda Z}\right],\end{aligned}$$

on hem fet servir la hipòtesi del teorema. La desigualtat que obtenim té exactament la mateixa forma que ens hem trobat en la prova del teorema 10 i la prova es pot acabar de manera idèntica. □

La manera més simple d'aplicar el teorema 12 es troba considerant funcions de diferències fitades, que hem introduït en la secció 2.1:

13 TEOREMA (DESIGUALTAT DE DIFERÈNCIES FITADES) *Suposem que la funció f satisfà la condició de diferències fitades amb constants $c_1, \ldots, c_n$ i denotem per*

$$v = \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

Sigui $Z = f(X_1, \ldots, X_n)$, on les variables X_i són independents. Aleshores,

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} \leq e^{-t^2/(2v)}.$$

La constant a l'exponent és millor que la que obtindríem amb una aplicació directa del teorema 12. Es pot provar amb un argument una mica més acurat. Observem que, com que la condició de diferències fitades és simètrica, Z també satisfà la desigualtat inferior

$$\mathbb{P}\{Z - \mathbb{E}Z < -t\} \leq e^{-t^2/(2v)}.$$

La desigualtat de diferències fitades estén el corol·lari 3 a una desigualtat de concentració exponencial. Així doncs, les aplicacions del corol·lari 3 en tots els exemples de funcions amb diferències fitades de la secció 2.1 (com el *bin packing*, la llargada de la subseqüència comuna més llarga i l'error L_1 de l'estimador de densitat) es troben ara substancialment millorades sense cap feina addicional.

A continuació, descrivim una altra aplicació, que és l'exemple més simple de desigualtat de concentració per a sumes de vectors aleatoris independents.

10 Exemple (una desigualtat de tipus Hoeffding en espais de Hilbert) Com a il·lustració de la potència de la desigualtat de diferències fitades, deduïm una desigualtat de tipus Hoeffding per a sumes de variables aleatòries que prenen valors en un espai de Hilbert. En particular, siguin $X_1, \ldots, X_n$ variables aleatòries amb esperança zero que prenen valors en un espai de Hilbert separable tal que $\|X_i\| \le c_i/2$ amb probabilitat 1 i denotem per $v = (1/4)\sum_{i=1}^n c_i^2$. Aleshores, per a tot $t \ge \sqrt{v}$,

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n X_i\right\| > t\right\} \le e^{-(t-\sqrt{v})^2/(2v)}.$$

Això es dedueix, simplement, observant que, per la desigualtat triangular, $Z = \|\sum_{i=1}^n X_i\|$ satisfà la propietat de diferències fitades amb constants c_i i, per tant,

$$\begin{aligned}\mathbb{P}\left\{\left\|\sum_{i=1}^n X_i\right\| > t\right\} &= \mathbb{P}\left\{\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|\right\} \\ &\le \exp\left(-\frac{\left(t - \mathbb{E}\|\sum_{i=1}^n X_i\|\right)^2}{2v}\right).\end{aligned}$$

La prova es pot completar observant que, a causa de la independència,

$$\mathbb{E}\left\|\sum_{i=1}^n X_i\right\| \le \sqrt{\mathbb{E}\left\|\sum_{i=1}^n X_i\right\|^2} = \sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2} \le \sqrt{v}.$$

La desigualtat de diferències fitades és un resultat clàssic de concentració i se'n coneixen diverses proves. Ha tingut incomptables aplicacions gràcies a la simplicitat de les hipòtesis i la generalitat del context. El resultat diu que si una funció no és massa sensible a pertorbacions individuals d'una qualsevol de les seves variables, aleshores satisfà una desigualtat de concentració exponencial. Amb tot, cal ressaltar que el teorema 12 és substancialment més potent. Per entendre per què aquesta desigualtat és un pas endavant significatiu en relació amb el teorema 13, observem simplement que en les hipòtesis del teorema 12 no s'exigeix que f tingui diferències fitades. L'únic que cal és que

$$\sup_{\substack{x_1,\ldots,x_n,\\ x'_1,\ldots,x'_n\in\mathcal{X}}} \sum_{i=1}^{n} \Big(f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_n)\Big)^2 \le v,$$

una condició òbviament molt més feble. Per il·lustrar per què el teorema 12 és una millora essencial, recordem el cas del valor propi més gran d'una matriu aleatòria simètrica, descrit anteriorment. Per a aquest exemple, el teorema 13 no arriba a donar cap desigualtat significativa.

11 Exemple (valor propi més gran d'una matriu aleatòria simètrica) Com abans, considerem una matriu aleatòria simètrica real A amb coeficients $X_{i,j}$, $1 \le i \le j \le n$, on $X_{i,j}$ són variables aleatòries independents amb valor absolut fitat per 1. Sigui $Z = \lambda_1$ el valor propi més gran de A. En la secció 2.3, ja hem vist que, gairebé segur,

$$\sum_{1\le i\le j\le n} (Z - Z_{i,j})^2 \le 4.$$

Hem fet servir aquesta estimació i la desigualtat de Efron-Stein per concloure que $\mathrm{Var}(Z) \le 4$. Fent servir el teorema 12, obtenim, sense cap feina addicional, l'estimació subgaussiana

$$\mathbb{P}\{Z > \mathbb{E}Z + t\} \le e^{-t^2/8}.$$

És clar que la desigualtat de diferències fitades no és útil aquí, ja que no es poden tractar les diferències individuals $Z - Z'_{i,j}$ d'una manera sensible, mentre que la suma dels seus quadrats està fitada per 4.

A continuació, descrivim una altra aplicació interessant del teorema 12. En la secció 3.3 hem provat el resultat fonamental que qualsevol funció de Lipschitz d'un vector gaussià normalitzat té cues subgaussianes. El mètode de l'entropia ens permet estendre aquest resultat a distribucions producte molt més generals, tot i que ens caldrà una condició addicional de convexitat sobre la funció de Lipschitz.

Una funció $f : [0,1]^n \to \mathbb{R}$ és *separadament convexa* si, per a cada $i = 1, \ldots, n$, f és una funció convexa de la i-èsima variable si la resta de variables estan fixades.

14 Teorema *Siguin $X_1,\ldots,X_n$ variables aleatòries independents que prenen valors en l'interval $[0,1]$ i sigui $f : [0,1]^n \to \mathbb{R}$ una funció separadament convexa tal que $|f(x) - f(y)| \le L\|x - y\|$ per a tot parell $x, y \in [0,1]^n$, on $L > 0$ és una constant fixada. Suposem que existeixen totes les derivades parcials de f. Aleshores, per a tot $t > 0$, la variable $Z = f(X_1,\ldots,X_n)$ satisfà*

$$\mathbb{P}\{Z > \mathbb{E}Z + t\} \le e^{-t^2/(2L^2)}.$$

Prova Pel teorema 12, n'hi ha prou amb fitar la variable aleatòria $\sum_{i=1}^n (Z - Z_i)^2$, on $Z_i = \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n)$. Denotem per X_i' el valor de x_i' pel qual el mínim s'assoleix (això queda garantit per la continuïtat de f i la compacitat del seu domini). Aleshores, escrivint $\overline{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$, tenim

$$\begin{aligned}
\sum_{i=1}^n (Z - Z_i)^2 &= \sum_{i=1}^n (f(X) - f(\overline{X}^{(i)})^2 \\
&\le \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(X)\right)^2 (X_i - X_i')^2 \\
&\qquad \text{(per la convexitat separada)} \\
&\le \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(X)\right)^2 \\
&= \|\nabla f(X)\|^2 \le L^2.
\end{aligned}$$

on, en el darrer pas, s'ha fet servir la propietat de Lipschitz de f. Per tant, podem aplicar el teorema 12 amb $v = L^2$. □

Observem que una fita naïf, fent servir la condició de Lipschitz, només ens donaria $\sum_{i=1}^n (f(X) - f(\overline{X}^{(i)})^2 \le 4nL^2$. La hipòtesi de convexitat dóna una millora immensa sobre aquesta simple fita.

12 Exemple (valor singular més gran d'una matriu aleatòria) Sigui A una matriu $m \times n$ amb coeficients $X_{i,j}$ ($i = 1, \ldots, m$, $j = 1, \ldots, n$) que són variables aleatòries independents que prenen valors en $[0, 1]$. Estem interessats en la concentració del valor singular més gran Z de A, definit com l'arrel quadrada del valor propi més gran de la matriu simètrica $A^T A$. Així,

$$Z = \sqrt{\lambda_1(A^T A)} = \sqrt{\sup_{u \in \mathbb{R}^n : \|u\| = 1} u^T A^T A u} = \sup_{u \in \mathbb{R}^n : \|u\| = 1} \|Au\|.$$

Per a cada vector fixat u, $\|Au\|$ és una funció convexa del vector de dimensió mn format per les variables $X_{i,j}$. Com que el suprem de funcions convexes és convex, Z és una funció convexa de $X_{i,j}$. Per tal d'aplicar el teorema 14, podem fer servir la desigualtat de Lidskii, un resultat clàssic d'àlgebra lineal, que diu que si $A = (x_{i,j})_{m \times n}$ i $B = (y_{i,j})_{m \times n}$ són dues matrius, aleshores, denotant per $s_1(M) \ge \ldots, s_n(M)$ els valors singulars d'una matriu M de mida $m \times n$,

$$\begin{aligned}
(s_1(A) - s_1(B))^2 &\le \sum_{i=1}^n (s_i(A) - s_i(B))^2 \le \sum_{i=1}^n s_i(A - B)^2 \\
&= \operatorname{tr}((A - B)^T (A - B)) = \sum_{i=1}^m \sum_{j=1}^n (x_{i,j} - y_{i,j})^2.
\end{aligned}$$

Per tant, el valor singular més gran és una funció de Lipschitz amb constant $L = 1$ i, pel teorema 14,

$$\mathbb{P}\{Z > \mathbb{E}Z + t\} \le e^{-t^2/2}.$$

3.5 Funcions autofitades

Ara, tornarem a veure les funcions autofitades que hem introduït en la secció 2.2. Recordem que una funció $f : \mathcal{X}^n \to \mathbb{R}$ té la propietat de ser autofitada si hi ha funcions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ tals que, per a tot $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ i cada $i = 1, \ldots, n$,

$$0 \le f(\boldsymbol{x}) - f_i(\boldsymbol{x}^{(i)}) \le 1$$

i

$$\sum_{i=1}^{n} \left(f(\boldsymbol{x}) - f_i(\boldsymbol{x}^{(i)}) \right), \le f(\boldsymbol{x})$$

on $\boldsymbol{x}^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Si $X_1, \ldots, X_n$ són variables aleatòries independents que prenen valors en $\mathcal{X}$ i $Z = f(X_1, \ldots, X_n)$ per a una funció f autofitada, aleshores la desigualtat d'Efron-Stein implica que $\mathrm{Var}(Z) \le \mathbb{E}Z$. Hem vist diversos exemples interessants de funcions autofitades, incloent-hi diverses funcions de configuració. Fent servir el mètode de l'entropia no és difícil obtenir fites de concentració exponencials per a funcions autofitades. Aquí n'enunciem una sense demostració. Considerem la funció

$$\phi(v) = \sup_{u \ge -1} (uv - h(u)) = e^v - v - 1.$$

15 Teorema *Suposem que Z satisfà la propietat de ser autofitada. Aleshores, per a tot $\lambda \in \mathbb{R}$,*

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \le \phi(\lambda)\mathbb{E}Z.$$

A més, per a tot $t > 0$,

$$\mathbb{P}\{Z \ge \mathbb{E}Z + t\} \le \exp\left(-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right)$$

i, per a cada $0 < t \le \mathbb{E}Z$,

$$\mathbb{P}\{Z \le \mathbb{E}Z - t\} \le \exp\left(-\frac{t^2}{2\mathbb{E}Z}\right).$$

El teorema 15 proporciona desigualtats de concentració per a qualsevol funció autofitada. En la secció 2.2, ja hem discutit diversos exemples d'aquesta mena de funcions.

Notes bibliogràfiques

Els principis clau del mètode d'entropia es basen en les idees que permeten provar les desigualtats de concentració gaussiana fent ús de les desigualtats logarítmiques de Sobolev. Va ser Ledoux [57] qui es va adonar que aquestes idees es poden fer servir per obtenir un camí alternatiu a algunes de les desigualtats de concentració exponencials de Talagrand per processos empírics i caos de Rademacher. Les idees de Ledoux van ser desenvolupades per Massart [69], Bousquet [20], Klein [52], Rio [88], Klein i Rio [53], Ledoux [58, 59] i Boucheron, Lugosi i Massart [17, 18, 19].

Es poden veure variacions de la desigualtat de tensorialització de l'entropia a Beckner [8], Latała i Oleszkiewicz [54], Ledoux [57], Boucheron *et al.* [16] i Chafaï [21].

Les desigualtats logarítmiques de Sobolev per a distribucions gaussianes van ser obtingudes inicialment per Gross [44]. Adrecem el lector interessat a l'excel·lent text d'Ané *et al.* [5] per a una extensa panoràmica sobre desigualtats logarítmiques de Sobolev amb connexions amb altres desigualtats funcionals, cadenes de Markov, teoria de la informació, etc.

L'argument atribuït a Herbst per deduir desigualtats de concentració basant-se en desigualtats logarítmiques de Sobolev apareix a Davies i Simon [30], (vegeu també Aida, Masuda i Shigekawa [2]). El mètode va ser extensament generalitzat i popularitzat per Ledoux [57, 56, 58, 59].

La desigualtat de concentració gaussiana del teorema 10 va ser provada originalment per Tsirelson, Ibragimov i Sudakov [99] fent servir arguments diferents dels que hem donat aquí, basats en càlcul estocàstic. Una forma més afinada d'aquesta desigualtat es dedueix del teorema isoperimètric gaussià, establert per primer cop de manera independent per Borell [15] i per Tsirelson i Sudakov [25].

La desigualtat de les diferències fitades és potser la desigualtat de concentració més simple i més àmpliament usada. La idea bàsica d'escriure una funció de variables aleatòries independents com una suma de diferències de martingales i de fer servir desigualtats exponencials per a martingales va ser introduïda inicialment en diverses aplicacions, incloent-hi Yurinskii [105], Maurey [72], Milman i Schechtman [77] i Shamir i Spencer [90]. La desigualtat va ser escrita explícitament per primer cop i il·lustrada amb una varietat d'aplicacions en un article panoràmic excel·lent de McDiarmid [73], i el mateix resultat s'ha referenciat sovint com a *desigualtat de McDiarmid*. Els mètodes de martingales han servit com una eina flexible i versàtil per provar desigualtats de concentració, vegeu també els articles panoràmics més recents de McDiarmid [74] i de Chung i Lu [23].

Mencionem un altre mètode que ha resultat útil per provar desigualtats de concentració exponencials, incloent-hi la desigualtat de diferències fitades. La idea principal es deu a Marton [66], qui va refinar resultats anteriors de la teoria de la informació (vegeu Ahlswede, Gács i Körner [1]; Marton [65], i Csiszár i Körner [28]).

El mètode es basa en l'acoblament, que es referencia sovint com a *mètode de transport*. Té l'avantatge que es pot estendre fàcilment a algunes funcions de variables aleatòries dependents (vegeu Marton [66] i Samson [89]). Notem que el mètode del transport també es pot fer servir per deduir concentració gaussiana, com la desigualtat de Tsirelson (vegeu Talagrand [97]). Per a més informació sobre aquest tema, adrecem el lector a Dembo [32] i Ledoux [59].

Les desigualtats per a cues de sumes de variables aleatòries independents sobre espais de Hilbert de l'exemple de la secció anterior són només un exemple senzill. Hi ha una vasta literatura que tracta sobre cues per a sumes de vectors aleatoris. De fet, algunes d'aquestes aplicacions han estat les motivacions més importants per alguns dels avenços més significatius en desigualtats de concentració. El lector interessat pot consultar Yurinskii [105, 106], Ledoux i Talagrand [61] i Pinelis [84].

La desigualtat exponencial pel valor propi més gran d'una matriu aleatòria va ser provada per Alon, Krivelevich i Vu [4]. Aquests autors van fer servir la desigualtat convexa de Talagrand que es presenta en la secció següent. Maurer [71] obté un exponent millor amb una anàlisi una mica més acurada.

El teorema 14 va ser establert per primer cop per Talagrand [96], que també va donar la desigualtat inferior corresponent. Més precisament, Talagrand prova que

$$\mathbb{P}\{|f(X) - \mathbb{M}f(X)| > t\} \leq 4e^{-t^2/(4L^2)},$$

on $\mathbb{M}f(X)$ denota la mediana de la variable aleatòria $f(X)$ i L és la constant de Lipschitz de f. De fet, aquesta desigualtat se satisfà amb la condició més feble que els conjunts de nivell $\{x : g(x) \leq t\}$ siguin convexos. La demostració que hem donat aquí es deu a Ledoux [57].

Les funcions autofitades van ser introduïdes per Boucheron, Lugosi i Massart [17], on es prova el teorema 15 basant-se en tècniques desenvolupades per Massart [69]. Diverses generalitzacions de funcions autofitades han estat considerades per Boucheron, Lugosi i Massart [18, 19]; Boucheron *et al.* [16]; Devroye [35]; Maurer [71], i McDiarmid i Reed [76].

4 Concentració i isoperimetria

Les desigualtats de concentració que hem discutit en aquest article estan íntimament relacionades amb problemes isoperimètrics. En aquesta secció final, tractem algunes d'aquestes relacions.

El teorema isoperimètric clàssic diu que entre tots els subconjunts de $\mathbb{R}^n$ amb un volum donat, les boles euclidianes minimitzen l'àrea de la seva superfície. Una formulació equivalent és que, per a tot $t > 0$, entre tots els conjunts (mesurables) $A \subset \mathbb{R}^n$ d'un volum donat, aquells pels quals el volum dels conjunts «inflats» de A

$$A_t = \{x \in \mathbb{R}^n : d(x, A) \leq t\}$$

té un volum mínim són les boles euclidianes. L'avantatge d'aquesta formulació és que evita la noció d'àrea d'una superfície i el problema es pot generalitzar a qualsevol espai mètric. De fet, s'han estudiat incomptables versions del problema isoperimètric clàssic.

En particular, donat un espai mètric $\mathcal{X}$ amb la seva mètrica d, podem considerar l'espai mesurable format per $\mathcal{X}$, la σ-àlgebra de tots els conjunts de Borel de $\mathcal{X}$ i una mesura de probabilitat $\mathbb{P}$. Sigui X una variable aleatòria que pren valors en $\mathcal{X}$ amb la distribució donada per $\mathbb{P}$. El problema isoperimètric en aquest cas és el següent: donada $p \in (0,1)$ i $t > 0$, es tracta de determinar els conjunts A amb $\mathbb{P}\{X \in A\} \geq p$ pels quals la mesura $\mathbb{P}\{d(X,A) \geq t\}$ és maximal. Tot i que la solució exacta només es coneix en pocs casos especials, es poden obtenir fites útils per a $\mathbb{P}\{d(X,A) \geq t\}$ en circumstàncies d'una generalitat remarcable. Aquestes fites es coneixen com *desigualtats isoperimètriques*.

En lloc de revisar la rica literatura sobre desigualtats isoperimètriques, veurem com les desigualtats de concentració i les desigualtats isoperimètriques estan relacionades. Descriurem les idees principals en un context senzill, però aquestes idees funcionen igualment en circumstàncies més generals. Com que estem interessats sobretot en desigualtats de concentració per a funcions de variables aleatòries independents, ens centrarem en el cas d'espais producte.

4.1 Perspectiva isoperimètrica de la desigualtat de diferències fitades

Considerem variables aleatòries independents $X_1, \ldots, X_n$ que prenen valors en un conjunt mesurable $\mathcal{X}$ i denotem per $X = (X_1, \ldots, X_n)$ el vector d'aquestes variables, que pren valors en $\mathcal{X}^n$.

Sigui $A \subset \mathcal{X}^n$ un conjunt mesurable arbitrari i escrivim $\mathbb{P}\{A\} = \mathbb{P}\{X \in A\}$. La *distància de Hamming* $d_H(x,y)$ entre els vectors $x, y \in \mathcal{X}^n$ es defineix com el nombre de coordenades en les quals x i y difereixen. Sigui

$$d_H(x,A) = \min_{y \in A} d_H(x,y),$$

la distància de Hamming entre el conjunt A i el punt x. Aleshores, obtenim la desigualtat isoperimètrica següent de la desigualtat de diferències fitades (teorema 13):

16 Teorema *Per a qualsevol $t > 0$,*

$$\mathbb{P}\left\{ d_H(X,A) \geq t + \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}\{A\}}} \right\} \leq e^{-2t^2/n}.$$

Observem que a la banda dreta de la desigualtat tenim la mesura del complement del conjunt «inflat» per $t + \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}\{A\}}}$ del conjunt A, és a dir, la mesura del conjunt de punts tals que la seva distància de Hamming des d'A és d'almenys $t + \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}\{A\}}}$. Si considerem un conjunt, amb $\mathbb{P}\{A\} = 1/10^6$, per exemple, veiem quelcom sorprenent: la mesura del conjunt de punts tals que la distància de Hamming a A és més gran que $10\sqrt{n}$ és menor que e^{-108}. En altres paraules, les mesures producte estan concentrades en conjunts extremadament petits (d'aquí, el nom *concentració de mesura*).

PROVA Observem que la funció $f(x) = d_H(x, A)$ no pot canviar més que 1 si es modifica una component de x, és a dir, té la propietat de diferències fitades amb constants $c_1 = \cdots = c_n = 1$. Així doncs, per la desigualtat de diferències fitades (teorema 13),

$$\mathbb{P}\{\mathbb{E}d_H(X, A) - d_H(X, A) \geq t\} \leq e^{-2t^2/n}.$$

Prenent $t = \mathbb{E}d_H(X, A)$, però, la banda esquerra esdevé $\mathbb{P}\{d_H(X, A) \leq 0\} = \mathbb{P}\{A\}$, de manera que la desigualtat anterior implica

$$\mathbb{E}d_H(X, A) \leq \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}\{A\}}}.$$

Aleshores, fent servir la desigualtat de les diferències fitades una altra vegada obtenim

$$\mathbb{P}\left\{d_H(X, A) \geq t + \sqrt{\frac{n}{2} \log \frac{1}{\mathbb{P}\{A\}}}\right\} \leq e^{-2t^2/n},$$

tal com volíem. □

Observem que la desigualtat de diferències fitades també es pot deduir del teorema anterior. Efectivament, si considerem una funció f en $\mathcal{X}^n$ que tingui la propietat de les diferències fitades amb constants $c_i = 1$ (per simplicitat), aleshores podem considerar el conjunt $A = \{x \in \mathcal{X}^n : f(x) \leq \mathbb{M}Z\}$, on $\mathbb{M}Z$ denota la mediana de la variable aleatòria $Z = f(X_1, \ldots, X_n)$. Aleshores, tenim clarament $\mathbb{P}\{A\} \geq 1/2$, de manera que el teorema anterior implica

$$\mathbb{P}\left\{Z - \mathbb{M}Z \geq t + \sqrt{\frac{n}{2} \log 2}\right\} \leq e^{-2t^2/n}.$$

Aquesta desigualtat té la mateixa forma que la desigualtat de diferències fitades llevat que el valor esperat de Z és substituït per la seva mediana. Es dedueix fàcilment de la desigualtat de Txebixev, però, que, per a qualsevol variable aleatòria Z amb variància finita, $|\mathbb{E}Z - \mathbb{M}Z| \leq \sqrt{\mathrm{Var}(Z)}$ i, per tant, la diferència entre tots dos paràmetres és, normalment, negligible.

4.2 Desigualtat de la distància convexa de Talagrand

El punt de vista isoperimètric va conduir Talagrand a desenvolupar, en una notable sèrie d'articles, un mètode d'inducció per provar resultats potents de concentració en molts casos en els quals la desigualtat de diferències fitades falla. Potser la més àmpliament usada d'aquestes desigualtats és la que s'anomena *desigualtat de la distància convexa*.

Per comprendre la desigualtat de Talagrand, observem primer que el teorema 16 es pot generalitzar fàcilment al cas en el qual la distància del punt x des del conjunt A es mesura per una *distància de Hamming amb pesos*

$$d_\alpha(x, A) = \inf_{y \in A} d_\alpha(x, y) = \inf_{y \in A} \sum_{i: x_i \neq y_i} |\alpha_i|,$$

on $\alpha = (\alpha_1, \ldots, \alpha_n)$ és un vector de nombres no negatius. Repetint l'argument de la demostració del teorema 16, obtenim, per a tot α,

$$\mathbb{P}\left\{d_\alpha(X, A) \geq t + \sqrt{\frac{\|\alpha\|^2}{2} \log \frac{1}{\mathbb{P}\{A\}}}\right\} \leq e^{-2t^2/\|\alpha\|^2},$$

on $\|\alpha\| = \sqrt{\sum_{i=1}^n \alpha_i^2}$ denota la norma euclidiana de α. Així, per exemple, per a tots els vectors α amb norma unitat $\|\alpha\| = 1$,

$$\mathbb{P}\left\{d_\alpha(X, A) \geq t + \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}\{A\}}}\right\} \leq e^{-2t^2}.$$

D'aquesta manera, denotant per $u = \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}\{A\}}}$, per a tot $t \geq u$,

$$\mathbb{P}\left\{d_\alpha(X, A) \geq t\right\} \leq e^{-2(t-u)^2}.$$

D'una banda, si $t \leq \sqrt{-2 \log \mathbb{P}\{A\}}$, aleshores $\mathbb{P}\{A\} \leq e^{-t^2/2}$. D'altra banda, com que $(t-u)^2 \geq t^2/4$ si $t \geq 2u$, per a qualsevol $t \geq \sqrt{2 \log \frac{1}{\mathbb{P}\{A\}}}$, la desigualtat anterior implica $\mathbb{P}\left\{d_\alpha(X, A) \geq t\right\} \leq e^{-t^2/2}$. Així doncs, per a tot $t > 0$, tenim

$$\sup_{\alpha: \|\alpha\|=1} \mathbb{P}\{A\} \cdot \mathbb{P}\left\{d_\alpha(X, A) \geq t\right\} \leq \sup_{\alpha: \|\alpha\|=1} \min\left(\mathbb{P}\{A\}, \mathbb{P}\left\{d_\alpha(X, A) \geq t\right\}\right) \leq e^{-t^2/2}.$$

El missatge principal de la desigualtat de Talagrand és que la desigualtat anterior es manté certa fins i tot si el suprem es pren dins la probabilitat. Per fer aquest enunciat precís, introduïm, per a cada $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, la *distància convexa* de x des d'A per

$$d_T(x, A) = \sup_{\alpha \in [0,\infty)^n: \|\alpha\|=1} d_\alpha(x, A).$$

Ara estem a punt per enunciar la desigualtat de la distància convexa de Talagrand.

17 Teorema (desigualtat de la distància convexa) *Per a qualsevol subconjunt $A \subseteq \mathcal{X}^n$ amb $\mathbb{P}\{X \in A\} \geq 1/2$ i $t > 0$,*

$$\min\left(\mathbb{P}\{A\}, \mathbb{P}\left\{d_T(X, A) \geq t\right\}\right) \leq e^{-t^2/4}.$$

Encara que a primera vista no és obvi com es pot fer servir el resultat de Talagrand per provar resultats de concentració per a funcions generals f de X, amb relativament poc esforç el teorema es pot convertir en desigualtats molt útils. La prova original de Talagrand es basa en un mètode d'inducció *ad hoc*, però el teorema es pot provar també pel mètode de l'entropia, fent servir el teorema 12. Adrecem el lector interessat a [18].

13 EXEMPLE (UNA ALTRA VISITA A BIN PACKING) Aquí descrivim una aplicació de la desigualtat de la distància convexa per al problema del *bin packing*, que hem tractat en la secció 2.1, tal com apareix a Talagrand [94]. Sigui $f(x)$ el mínim nombre de pots de cabuda 1 en els quals es poden empaquetar els nombres $x_1, \ldots, x_n \in [0,1]$. Considerem la variable aleatòria $Z = f(X)$, on $X_1, \ldots, X_n$ són variables aleatòries independents que prenen valors en $[0,1]$.

18 COROL·LARI *Sigui* $\Sigma = \sqrt{\mathbb{E}\sum_{i=1}^n X_i^2}$. *Aleshores, per a cada* $t > 0$,

$$\mathbb{P}\{|Z - \mathbb{M}Z| \geq t + 1\} \leq 8e^{-t^2/(16(2\Sigma^2+t))}.$$

PROVA Observem primer (i aquesta és l'única propietat específica de f que es fa servir a la prova) que, per a cada $x, y \in [0,1]^n$,

$$f(x) \leq f(y) + 2 \sum_{i:x_i \neq y_i} x_i + 1.$$

Per veure això, n'hi ha prou amb veure que els x_i pels quals $x_i \neq y_i$ es poden empaquetar en, com a molt, $\left\lfloor 2\sum_{i:x_i\neq y_i} x_i \right\rfloor + 1$ pots. Per això, només cal trobar un empaquetament en el qual hi hagi, com a molt, un pot que estigui ple a menys de la meitat. Un empaquetament així ha d'existir perquè sempre podem empaquetar el contingut de dos pots mig plens en un.

Denotant per $\alpha = \alpha(x) \in [0,\infty)^n$ el vector unitat $x/\|x\|$, tenim

$$\sum_{i:x_i \neq y_i} x_i = \|x\| \sum_{i:x_i \neq y_i} \alpha_i = \|x\| d_\alpha(x,y).$$

Sigui a un nombre positiu i definim el conjunt $A_a = \{y : f(y) \leq a\}$. Aleshores, per l'argument anterior i per la definició de distància convexa, per a cada $x \in [0,1]^n$ hi ha un $y \in A_a$ tal que

$$f(x) \leq f(y) + 2 \sum_{i:x_i \neq y_i} x_i + 1 \leq a + 2\|x\| d_T(x, A_a) + 1,$$

d'on deduïm que, per a cada $a > 0$, $Z \leq a + 2\|X\| d_T(X, A_a) + 1$. Així doncs, si escrivim $\Sigma = \sqrt{\mathbb{E}\sum_{i=1}^n X_i^2}$, per a cada $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\{Z \geq a + 1 + t\} \\
&\quad \leq \mathbb{P}\left\{Z \geq a + 1 + t\frac{2\|X\|}{2\sqrt{2\Sigma^2 + t}}\right\} + \mathbb{P}\left\{\|X\| \geq \sqrt{2\Sigma^2 + t}\right\} \\
&\quad \leq \mathbb{P}\left\{d_T(X, A_a) \geq \frac{t}{2\sqrt{2\Sigma^2 + t}}\right\} + e^{-(3/8)(\Sigma^2+t)}
\end{aligned}$$

on la fita per al segon terme prové d'una simple aplicació d'una desigualtat clàssica per a la suma de variables aleatòries independents anomenada *desigualtat de Bernstein*.

Per obtenir la desigualtat buscada, fem servir la fita que hem obtingut amb dues eleccions diferents de a. Per tal d'obtenir una fita per a la cua superior de Z, prenem $a = \mathbb{M}Z$. Aleshores, $\mathbb{P}\{A_a\} \geq 1/2$ i la desigualtat de la distància convexa dóna

$$\mathbb{P}\{Z \geq \mathbb{M}Z + 1 + t\} \leq 2\left(e^{-t^2/(16(2\Sigma^2+t))} + e^{-(3/8)(\Sigma^2+t)}\right) \leq 4e^{-t^2/(16(2\Sigma^2+t))}.$$

Obtenim una desigualtat similar de la mateixa manera per a $\mathbb{P}\{Z \leq \mathbb{M}Z - 1 - t\}$ prenent $a = \mathbb{M}Z - t - 1$. □

Notes bibliogràfiques

La connexió entre desigualtats de concentració i propietats isoperimètriques ve ja de Lévy [62]. La importància de la concentració de mesura en la teoria asimptòtica dels espais de Banach està resumida en el text que avui ha esdevingut clàssic de Milman i Schechtman [77] on molts dels resultats inicials estan recollits. El llibre més recent de Ledoux [59] és un resum excel·lent sobre concentració de mesura i les seves connexions amb problemes isoperimètrics i altres nocions geomètriques relacionades. Per a panoràmiques sobre desigualtats isoperimètriques discretes, ens referim a Leader [55] i Bezrukov [11].

L'aproximació *isoperimètrica* a desigualtats de concentració va ser promoguda, en gran part, per una notable sèrie d'articles de Talagrand [94, 95, 96]. El mètode d'inducció de Talagrand va provocar un avenç important en la teoria i les aplicacions de les desigualtats de concentració exponencials.

Talagrand [94], Steele [93] i Molloy i Reed [78] resumeixen una gran varietat d'aplicacions de la desigualtat de la distància convexa.

Referències

[1] Ahlswede, R.; Gács, P.; Körner, J. «Bounds on conditional probabilities with applications in multi-user communication». *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 34 (1976), 157–177. [Correccions a 39 (1977), 353–354]

[2] Aida, S.; Masuda, T.; Shigekawa, I. «Logarithmic Sobolev inequalities and exponential integrability». *J. Funct. Anal.*, 126 (1994), 83–101.

[3] Aldous, D.; Diaconis, P. «Hammersley's interacting particle process and longest increasing subsequences». *Probab. Theory Related Fields*, 103 (2) (1995), 199–213.

[4] Alon, N.; Krivelevich, M.; Vu, V. H. «On the concentration of eigenvalues of random symmetric matrices». *Israel J. Math.*, 131 (2002), 259–267.

[5] ANÉ, C.; BLACHÈRE, S.; CHAFAÏ, D.; FOUGÈRES, P.; GENTIL, I.; MALRIEU, F.; ROBERTO, C.; SCHEFFER, G. *Sur les inégalités de Sobolev logarithmiques*. París: Société Mathématique de France, 2000. (Panoramas et Synthèses, 10.)

[6] ANTHONY, M.; BARTLETT, P. L. *Neural network learning: Theoretical foundations*. Cambridge: Cambridge University Press, 1999.

[7] BAIK, J.; DEIFT, P.; JOHANSSON, K. «On the distribution of the length of the second row of a Young diagram under Plancherel measure». *Geom. Funct. Anal.*, 10 (2000), 702–731.

[8] BECKNER, W. «A generalized Poincaré inequality for Gaussian measures». *Proc. Amer. Math. Soc.*, 105 (1989), 397–400.

[9] BENNETT, G. «Probability inequalities for the sum of independent random variables». *J. Amer. Statist. Assoc.*, 57 (1962), 33–45.

[10] BERNSTEIN, S. N. *The theory of probabilities*. Moscou: Gastehizdat, 1946.

[11] BEZRUKOV, S. L. «Isoperimetric problems in discrete spaces». A: *Extremal problems for finite sets (Visegrád, 1991)*. Budapest: János Bolyai Math. Soc., 1994. (Bolyai Soc. Math. Stud., 3, 59–91).

[12] BLUMER, A.; EHRENFEUCHT, A.; HAUSSLER, D.; WARMUTH, M. K. «Learnability and the Vapnik-Chervonenkis dimension». *J. ACM*, 36 (1989), 929–965.

[13] BOBKOV, S.; LEDOUX, M. «Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution». *Probab. Theory Related Fields*, 107 (1997), 383–400.

[14] BOLLOBÁS, B.; BRIGHTWELL, G. «The height of a random partial order: Concentration of measure». *Ann. Appl. Probab.*, 2 (1992), 1009–1018.

[15] BORELL, C. «The Brunn-Minkowski inequality in Gauss space». *Invent. Math.*, 30 (1975), 207–216.

[16] BOUCHERON, S.; BOUSQUET, O.; LUGOSI, G.; MASSART, P. «Moment inequalities for functions of independent random variables». *Ann. Probab.*, 33 (2005), 514–560.

[17] BOUCHERON, S.; LUGOSI, G.; MASSART, P. «A sharp concentration inequality with applications». *Random Structures Algorithms*, 16 (2000), 277–292.

[18] BOUCHERON, S.; LUGOSI, G.; MASSART, P. «Concentration inequalities using the entropy method». *Ann. Probab.*, 31 (2003), 1583–1614.

[19] BOUCHERON, S.; LUGOSI, G.; MASSART, P. «On concentration of self-bounding functions». 2009. [Manuscrit]

[20] BOUSQUET, O. «A Bennett concentration inequality and its application to suprema of empirical processes». *C. R. Acad. Sci. Paris, Ser. I*, 334 (6) (2002), 495–500.

[21] CHAFAÏ, D. *On ϕ-entropies and ϕ-Sobolev inequalities*. 2002. arXiv.math. PR/0211103. [Informe tècnic]

[22] Chernoff, H. «A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations». *Ann. Math. Stat.*, 23 (1952), 493–507.

[23] Chung, F.; Lu, L. «Concentration inequalities and martingale inequalities: a survey». *Internet Math.*, 3 (2006), 79–127.

[24] Chvátal, V.; Sankoff, D. «Longest common subsequences of two random sequences». *J. Appl. Probab.*, 12 (1975), 306–315.

[25] Cirel'son, B. S.; Sudakov, V. N. «Extremal properties of half spaces for spherically invariant measures». *J. Soviet. Math.*, 9 (1978), 9–18.

[26] Cover, T. M.; Thomas, J. A. *Elements of information theory*. Nova York: John Wiley, 1991.

[27] Craig, C. «On the Tchebychef inequality of Bernstein». *Ann. Math. Stat.*, 4 (1933), 94–102.

[28] Csiszár, I.; Körner, J. *Information theory: Coding theorems for discrete memoryless systems*. Nova York: Academic Press, 1981.

[29] Dančík, V.; Paterson, M. «Upper bound for the expected». A: *Proceedings of STACS'94*. Nova York: Springer 1994 , 669–678. (Lecture Notes in Comput. Sci., 775).

[30] Davies, E. B.; Simon, B. «Ultracontractivity and the heat kernel for Schrödinger operators and Dirichlet Laplacians». *J. Funct. Anal.*, 59 (1984), 335–395.

[31] Deken, J. P. «Some limit results for longest common subsequences». *Discrete Math.*, 26 (1979), 17–31.

[32] Dembo, A. «Information inequalities and concentration of measure». *Ann. Probab.*, 25 (1997), 927–939.

[33] Devroye, L. «The kernel estimate is relatively stable». *Probab. Theory Related Fields*, 77 (1988), 521–536.

[34] Devroye, L. «Exponential inequalities in nonparametric estimation». A: Roussas, G. [ed.]. *Nonparametric functional estimation and related topics*. Dordrecht: Kluwer Academic Publishers: NATO ASI Series, 1991, 31–44

[35] Devroye, L. «Laws of large numbers and tail inequalities for random tries and Patricia trees». *J. Comput. Appl. Math.*, 142 (2002), 27–37.

[36] Devroye, L.; Györfi, L. *Nonparametric density estimation: The L_1 view*. Nova York: John Wiley, 1985.

[37] Devroye, L.; Györfi, L.; Lugosi, G. *A probabilistic theory of pattern recognition*. Nova York: Springer, 1996.

[38] Devroye, L.; Lugosi, G. *Combinatorial methods in density estimation*. Nova York: Springer-Verlag, 2000.

[39] Efron, B.; Stein, C. «The jackknife estimate of variance». *Ann. Stat.*, 9 (1981), 586–596.

[40] Frieze, A. M. «On the value of a random minimum spanning tree problem». *Discrete Appl. Math.*, 10 (1985), 47–56.

[41] Frieze, A. M. «On the length of the longest monotone subsequence in a random permutation». *Ann. Appl. Probab.*, 1 (1991), 301–305.

[42] Füredi, Z.; Komlós, J. «The eigenvalues of random symmetric matrices». *Combinatorica*, 1 (1981), 233–241.

[43] Groeneboom, P. «Hydrodynamical methods for analyzing longest increasing subsequences: Probabilistic methods in combinatorics and combinatorial optimization». *J. Comput. Appl. Math.*, 142 (2002), 83–105.

[44] Gross, L. «Logarithmic Sobolev inequalities». *Amer. J. Math.*, 97 (1975), 1061–1083.

[45] Hammersley, J. M. «A few seedlings of research». A: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971).* Vol. I: *Theory of statistics* Berkeley, Calif,: Univ. California Press, 1972, 345–394.

[46] Han, T. S. «Nonnegative entropy measures of multivariate symmetric correlations». *Information and Control*, 36 (1978), 133–156.

[47] Hoeffding, W. «Probability inequalities for sums of bounded random variables». *J. Amer. Statist. Assoc.*, 58 (1963), 13–30.

[48] Its, A. R.; Tracy, C. A.; Widom, H. «Random words, toeplitz determinants and integrable systems I». A: Bleher, P.; Its, A. R. [ed.]. *Random matrix models and their applications.* Cambridge: Cambridge University Press, 2001, 245–258. (Math. Sci. Res. Inst. Publ.), 40.

[49] Janson, S. «The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph». *Random Structures Algorithms*, 7 (1995), 337–355.

[50] Janson, S.; Łuczak, T.; Ruciński. A. *Random graphs.* Nova York: John Wiley, 2000.

[51] Johansson, K. «Discrete orthogonal polynomial ensembles and the Plancherel measure». *Ann. Math.*, 153 (2001), 259–296.

[52] Klein, T. «Une inégalité de concentration à gauche pour les processus empiriques». *C. R. Acad. Sci. Paris, Ser. I*, 334 (6) (2002), 501–504.

[53] Klein, T.; Rio, E. «Concentration around the mean for maxima of empirical processes». *Ann. Probab.*, 33 (2005), 1060–1077.

[54] Latała, R.; Oleszkiewicz, C. «Between Sobolev and Poincaré». A: *Geometric aspects of functional analysis, Israel Seminar (GAFA), 1996–2000.* Berlín: Springer, 2000, 147–168. (Lecture Notes in Math., 1745).

[55] Leader, I. «Discrete isoperimetric inequalities». A: *Probabilistic combinatorics and its applications (San Francisco, CA, 1991)*, 44, 57–80. Providence, R. I.: Amer. Math. Soc., 1991.

[56] Ledoux, M. «Isoperimetry and gaussian analysis». A: Bernard, P. [ed.]. *Ecole d'Eté de Probabilités de St-Flour XXIV-1994*, 1996, 165–294. (Lecture Notes in Math., 1648).

[57] LEDOUX, M. «On Talagrand's deviation inequalities for product measures». *ESAIM Probab. Stat.*, 1 (1997), 63–87. http://www.emath.fr/ps/

[58] LEDOUX, M. «Concentration of measure and logarithmic Sobolev inequalities». A: *Séminaire de Probabilités XXXIII*. Berlín: Springer, 1999, 120–216. (Lecture Notes in Math. 1709)

[59] LEDOUX, M. *The concentration of measure phenomenon*. Providence, R. I.: Amer. Math. Soc., 2001.

[60] LEDOUX, M. «Distributions of invariant ensembles from the classical orthogonal polynomials: the discrete case». *Electron. J. Probab.*, 10 (2005), 1116–1146.

[61] LEDOUX, M.; TALAGRAND, M. *Probability in Banach space*. Nova York: Springer, 1991.

[62] LÉVY, P. *Problèmes concrets d'analyse fonctionelle*. París: Gauthier-Villars, 1951.

[63] LOGAN, B. F.; SHEPP, L. A. «A variational problem for Young tableaux». *Adv. Math.*, 26 (1977), 206–222.

[64] LUCZAK, M. J.; MCDIARMID, C. «Concentration for locally acting permutations». *Discrete Math.*, 265 (2003), 159–171.

[65] MARTON, K. «A simple proof of the blowing-up lemma». *IEEE Trans. Inform. Theory*, 32 (1986), 445–446.

[66] MARTON, K. «Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration». *Ann. Probab.*, 24 (1996), 857–866.

[67] MARTON, K. «A measure concentration inequality for contracting Markov chains». *Geom. Funct. Anal.*, 6 (1996), 556–571. [Correccions: 7 (1997), 609–613]

[68] MASSART, P. *Optimal constants for Hoeffding type inequalities*. Université de Paris-Sud, 1998. [Informe tècnic de matemàtiques 98.86]

[69] MASSART, P. «About the constants in Talagrand's concentration inequalities for empirical processes». *Ann. Probab.*, 28 (2000), 863–884.

[70] MASSART, P. «Some applications of concentration inequalities to statistics». *Annales de la Faculté des Sciencies de Toulouse*, IX (2000), 245–303.

[71] MAURER, A. «Concentration inequalities for functions of independent variables». *Random Structures Algorithms*, 29 (2006), 121–138.

[72] MAUREY, B. «Construction de suites symétriques». *C. R. Acad. Sci. Paris, A* , 288 (1979), 679–681.

[73] MCDIARMID, C. «On the method of bounded differences». A: *Surveys in combinatorics 1989*. Cambridge: Cambridge University Press, 1989, 148–188.

[74] MCDIARMID, C. «Concentration». A: HABIB, M.; MCDIARMID, C.; RAMIREZ-ALFONSIN, J.; REED, B. [ed.]. *Probabilistic methods for algorithmic discrete mathematics*. Nova York: Springer, 1998, 195–248.

[75] McDiarmid, C. «Concentration for independent permutations». *Combin. Probab. Comput.*, 2 (2002), 163–178.

[76] McDiarmid, C.; Reed, B. «Concentration for self-bounding functions and an inequality of talagrand». *Random Structures Algorithms*, 29 (2006), 549–557.

[77] Milman, V.; Schechtman, G. *Asymptotic theory of finite-dimensional normed spaces*. Nova York: Springer, 1986.

[78] Molloy, M.; Reed, B. *Graph colouring and the probabilistic method*. Berlín: Springer, 2002. (Algorithms Combin., 23).

[79] Okamoto, M. «Some inequalities relating to the partial sum of binomial probabilities». *Ann. Inst. Stat. Math.*, 10 (1958), 29–35.

[80] Panchenko, D. «A note on Talagrand's concentration inequality». *Electron. Commun. Probab.*, 6 (2001), 55–65.

[81] Panchenko, D. «Some extensions of an inequality of Vapnik and Chervonenkis». *Electron. Commun. Probab.*, 7 (2002), 55–65.

[82] Panchenko, D. «Symmetrization approach to concentration inequalities for empirical processes». *Ann. Probab.*, 31 (2003), 2068–2081.

[83] Peña, V. H. de la ; Giné, E. *Decoupling: From dependence to independence*. Nova York: Springer 1999.

[84] Pinelis, I. «Optimum bounds on moments of sums of independent random vectors». *Siberian Adv. Math.*, 5 (1995), 141–150.

[85] Rhee, W. «A matching problem and subadditive Euclidean functionals». *Ann. Appl. Probab.*, 3 (1993), 794–801.

[86] Rhee, W. T.; Talagrand, M. «Martingale inequalities and the jackknife estimate of variance». *Stat. Probab. Lett.*, 4 (1986), 5–6.

[87] Rhee, W. T.; Talagrand, M. «Martingales, inequalities; NP-complete problems». *Math. Oper. Res.*, 12 (1987), 177–181.

[88] Rio, E. «Inégalités de concentration pour les processus empiriques de classes de parties». *Probab. Theory Related Fields*, 119 (2001), 163–175.

[89] Samson, P.-M. «Concentration of measure inequalities for Markov chains and ϕ-mixing processes». *Ann. Probab.*, 28 (2000), 416–461.

[90] Shamir, E.; Spencer, J. «Sharp concentration of the chromatic number on random graphs $g_{n,p}$». *Combinatorica*, 7 (1987), 374–384.

[91] Steele, J. M. «Long common subsequences and the proximity of two random strings». *SIAM J. Appl. Math.*, 42 (1982), 731–737.

[92] Steele, J. M. «An Efron-Stein inequality for nonsymmetric statistics». *Ann. Stat.*, 14 (1986), 753–758.

[93] Steele, J. M. *Probability theory and combinatorial optimization*. SIAM, 3600 University City Science Center, Phila, PA 19104, 1996. (CBMS-NSF Regional Conference Series in Applied Mathematics, 69).

[94] Talagrand, M. «Concentration of measure and isoperimetric inequalities in product spaces». *Publ. Math. Inst. Hautes Études Sci.*, 81 (1995), 73–205.

[95] Talagrand, M. «New concentration inequalities in product spaces». *Invent. Math.*, 126 (1996), 505–563.

[96] Talagrand, M. «A new look at independence». *Ann. Probab.*, 24 (1996), 1–34. [Col·laboració especial]

[97] Talagrand, M. «Transportation cost for Gaussian and other product measures». *Geom. Funct. Anal.*, 6 (1996), 587–600.

[98] Tracy, C. A.; Widom, H. «On the distributions of the lengths of the longest monotone subsequences in random words». *Probab. Theory Related Fields*, 119 (2001), 350–380.

[99] Tsirelson, B. S.; Ibragimov, I. A.; Sudakov, V. N. «Norm of gaussian sample function». A: *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*. Berlín: Springer, 1976, 20–41. (Lecture Notes in Math., 550).

[100] Uspensky, J. V. *Introduction to mathematical probability*. Nova York: McGraw-Hill, 1937.

[101] Vapnik, V. N. *Statistical learning theory*. Nova York: John Wiley, 1998.

[102] Vapnik, V. N.; Chervonenkis, A. Y. «On the uniform convergence of relative frequencies of events to their probabilities.» *Theory Probab. Appl.*, 16 (1971), 264–280.

[103] Vapnik, V. N.; Chervonenkis, A. Y. *Theory of pattern recognition*. Moscou: Nauka, 1974. [En rus. Traducció alemanya: *Theorie der Zeichenerkennung*. Berlín: Akademie, 1979]

[104] Wästlund, J. «Evaluation of Janson's constant for the variance in the random minimum spanning tree problem». *Linköping Studies in Math.*, 7 (2005).

[105] Yurinskii, V. V. «Exponential inequalities for sums of random vectors». *J. Multivariate Anal.*, 6 (1976), 472–499.

[106] Yurinsky, V. *Sums and gaussian vectors*. Nova York: Springer, 1995.

ICREA i Departament d'Economia i Empresa
Universitat Pompeu Fabra
C/ Ramon Trias Fargas, 15-27
08005 Barcelona
lugosi@upf.edu