Introducción a las desigualdades de concentración.

Guiomar Paniagua Pelaz

Trabajo Fin de Grado
dirigido por Eustasio Del Barrio Tellado

Grado en Matemáticas
Universidad de Valladolid
Julio 2016

Universidad de Valladolid

Índice general

Introducción

Este trabajo explora alguno de los principales métodos para obtener desigualdades de concentración, es decir, cotas para la probabilidad de que una variable aleatoria se desvíe de su valor central. Es un hecho bien conocido que la media de variables aleatorias independientes e igualmente distribuidas tiende a concentrarse en torno a su valor medio. Desde un punto de vista cualitativo, éste es el resultado de la ley de los Grandes Números. Esta memoria se centra más en resultados cuantitativos. En el caso de la media muestral, la desigualdad de *Chebyshev* sería un ejemplo de desigualdad de concentración. Esta desigualdad admite mejoras en muchas direcciones. Por un lado, bajo ciertas condiciones, es posible probar desigualdades exponenciales, más fuertes que la desigualdad cuadrática de *Chebyshev*. Por otro lado, la concentración de variables aleatorias se observa en situaciones mucho más generales que la de sumas de variables independientes.

A lo largo de este trabajo se han estudiado refinamientos en estas dos direcciones. Partiendo del método de *Chernoff*, se han recorrido dos grupos de técnicas principales para la obtención de desigualdades de concentración: métodos basados en martingalas y métodos basados en la entropía. En el primer grupo se incluye la desigualdad de *Efron-Stein*. En origen, ésta es una desigualdad de varianzas, aunque en la memoria se muestra cómo, en ocasiones, se puede aprovechar la idea para obtener cotas exponenciales. En este trabajo se muestran varias aplicaciones de la desigualdad de *Efron-Stein*, incluyendo desigualdades de concentración para el máximo autovalor de ciertas matrices aleatorias y a la demostración de la desigualdad gaussiana de *Poincaré*.

El segundo grupo de técnicas se basa en el uso de la entropía. Dedicamos un capítulo a exponer algunas propiedades de la entropía y de la entropía relativa, incluyendo su relación con las funciones de *Chernoff*, a través del resultado conocido como "Lema de transporte". Estas propiedades se explotan en el último capítulo para obtener desigualdades logarítmicas de *Sobolev*.

Éstas son acotaciones de la entropía de funciones suficientemente suaves por momentos del gradiente, en un sentido que se precisará. El interés de las desigualdades de concentración es que se puede obtener las segundas a partir de las primeras mediante argumentos tipo *Herbst* descritos en esta memoria. Como ilustración de la utilidad y la potencia de este método, la memoria concluye con resultados sobre la concentración de vectores y procesos aleatorios gaussianos, incluyendo la desigualdad de Tsirelson-Ibragimov-Sudakov.

Capítulo 1

Desigualdades básicas

La ley de los grandes números nos dice que el promedio de variables aleatorias independientes e igualmente distribuidas tiende a concentrarse en torno a su valor central. En este sentido podemos interpretar esta ley como un resultado de concentración. Sin embargo, en esta memoria nos centraremos en resultados cualitativos más que en resultados asintóticos como el anterior.

El objetivo será ver bajo qué condiciones se tiene una cota superior para la probabilidad que una variable aleatoria real Z difiera de EZ de una cierta cantidad, es decir,

$$P(Z - EZ \geq t), \quad P(Z - EZ \leq -t) \quad \text{donde } t > 0.$$

A esta cota se la conoce como desigualdad de concentración. Se asume que el valor esperado EZ existe. Además como la probabilidad $P(|Z - EZ| \geq t)$ coincide por

$$P(Z - EZ \geq t) + P(EZ - Z \geq t)$$

se considerará $\tilde{Z} = Z - EZ$ ó $\tilde{Z} = EZ - Z$ para centrarse en cotas exponenciales para $P(Z \geq t)$ con Z variable aleatoria centrada. A $P(Z - EZ \geq t)$ se la llama probabilidad de la cola por la derecha y a $P(Z - EZ \leq -t)$ probabilidad de la cola por la izquierda y así será usual encontrarlo a lo largo del trabajo.

Una desigualdad de concentración básica es la desigualdad de Chevyshev. Si Z es una variable aleatoria con varianza finita entonces

$$P(|Z - EZ| \geq t) \leq \frac{\text{Var(Z)}}{t^2}, \quad t > 0.$$

En la sección (1.1) se discuten las virtudes y las limitaciones de esta cota, así como posibles mejoras. Esta desigualdad nos dice que podemos obtener cotas de concentración para Z si disponemos de cotas para $\mathrm{Var}(Z)$. Por otro lado, de la representación

$$\mathrm{Var(Z)} = \int_0^\infty P(|Z - EZ| > \sqrt{n})dt,$$

deducimos que los problemas de encontrar cotas para la varianza y para las colas están estrechamente relacionados, un hecho al que se vuelve en varias partes de esta memoria.

Aunque discutimos en este trabajo desigualdades de concentración respecto de la media, ésta no es la única elección posible como valor central de una distribución. En la sección (1.1) se prueba resulta equivalente obtener resultados de concentración respecto de la media o de la mediana.

Se describirá el método de *Cramér-Chernoff*, técnica básica para deducir cotas superiores exponenciales para colas de probabilidad. Nos detendremos en dos casos con especial importancia en los que las sumas muestran dos comportamientos diferentes: sub-gaussiano y sub-exponencial.

Aunque no es el objetivo principal, en la sección (1.2) se mostrará una aplicación del método de *Chernoff* a la obtención de desigualdades maximales.

Por último se enunciarán desigualdades de sumas de variables aleatorias, $Z = X_1 + ... + X_n$ con $X_1, ..., X_n$ independientes que en los casos más favorables pueden derivar a colas de probabilidad exponenciales. De forma general se verá en capítulos posteriores que Z se tome como una función de n variables $X_1, ..., X_n$.

1.1. Preliminares

Un resultado elemental pero potente para limitar las probabilidades de cola es la desigualdad de *Chebyshev* que puede ser deducida fácilmente de la desigualdad de *Markov*. La desigualdad de *Chebyshev* es uno de los resultados clásicos más importantes de la teoría de probabilidad.

1.1 Teorema (Desigualdad de Markov). *Sea Y una variable aleatoria no negativa e integrable , para todo $t > 0$.*

$$P(Y \geq t) \leq \frac{EY}{t}.$$

Aplicando la desigualdad de *Markov* a variables $Y = |Z - EZ|$ se pueden obtener desigualdades de concentración y si además se exigen condiciones de integrabilidad más fuertes sobre Z se produce una mejora en la cota. La idea es aplicar la desigualdad de *Markov* a unas funciones convenientes. Si Φ es una función no decreciente y no negativa definida en un intervalo $I \subset \mathbb{R}$ y si Y denota una variable aleatoria que toma valores en I, entonces la desigualdad de *Markov* implica que para todo t $\in$ I con $\Phi(t) > 0$,

$$P(Y \geq t) \leq P(\Phi(Y) \geq \Phi(t)) \leq \frac{E\Phi(Y)}{\Phi(t)} \tag{1.1}$$

Tomando $\Phi(t) = t^2$ en $I = (0, \infty)$ e $Y = |Z - EZ|$ en (1.1) se obtiene la desigualdad de *Chebyshev*.

$$P(|Z - EZ| \geq t) \leq \frac{\text{Var(Z)}}{t^2}$$

con Z una variable aleatoria no negativa e integrable , para todo $t > 0$.

De forma general, se puede tomar $\Phi(t) = t^q$ para algún $q > 0$. Entonces para todo $t > 0$ se tiene que

$$P(|Z - EZ| \geq t) \leq \frac{E[|Z - EZ|^q]}{t^q}$$

Uno de los casos más sencillos se da cuando Z es suma de variables aleatorias independientes, es decir, $Z = X_1 + ... + X_n$, en este caso se tiene que $\text{Var(Z)} = \sum_{i=1}^{n} Var(X_i)$, y por tanto, la desigualdad de *Chebyshev* se transforma en

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - EX_i)\right| \geq t\right) \leq \frac{\sigma^2}{nt^2} \quad \text{donde} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var(X_i).$$

Aunque la desigualdad de *Chebyshev* permite manejar con facilidad la concentración de sumas de variables independientes, también presenta algunas limitaciones. Por ejemplo, para t próximo a 0 la cota se hace grande

y la cota no resulta suficientemente útil. Una mejora de esta cota la da la desigualdad de *Chebyshev-Cantelli* en la que el término acotante es siempre menor o igual a uno.

1.2 Teorema (**Desigualdad de Chebyshev-Cantelli**). *Para cualquier variable aleatoria con valores reales Y y sea $t > 0$,*

$$P(Y - EY \geq t) \leq \frac{Var(Y)}{Var(Y) + t^2}. \tag{1.2}$$

Demostración. Sea $Z = Y - EY$, entonces $EZ = 0$. Como $t = E(t - Z)$ se tiene

$$t = E(t - Z) \leq E((t - Z)I(Z < t))$$

y elevando ambos miembros de la desigualdad al cuadrado se tiene

$$\begin{aligned} t^2 &\leq [E((t - Z)I(Z < t))]^2 \\ &\leq E(t - Z)^2 P(Z < t) \\ &= (EZ^2 - 2tEZ + t^2)(P(Z < t)) \\ &= (\mathrm{Var(Z)} + t^2)P(Z < t). \end{aligned}$$

Luego,

$$P(Z < t) \geq \frac{t^2}{\mathrm{Var(Z)} + t^2}.$$

Es decir,

$$1 - P(Z < t) = P(Z \geq t) \leq \frac{\mathrm{Var(Z)}}{\mathrm{Var(Z)} + t^2} = \frac{Var(Y)}{Var(Y) + t^2}$$

□

A pesar de la mejora, la cota de *Chebyshev- Cantelli* y la de *Chebyshev* son del mismo orden para t grande. Posteriormente veremos cómo usar de forma distinta la desigualdad de *Markov* para obtener cotas exponenciales para las cotas.

Aunque hasta ahora, y en casi toda la memoria, tratamos la concentración respecto de la media, probar la concentración en torno a la media o en torno a la mediana resulta equivalente como se puede ver a continuación,

1.3 Teorema. *Sea MZ la mediana de una variable aleatoria Z de cuadrado integrable (es decir, $P(Z \geq MZ) \geq 1/2$ y $P(Z \leq MZ) \geq 1/2$). Se tiene que $|MZ - EZ| \leq \sqrt{\mathrm{Var(Z)}}$.*

Demostración. Equivaldría a ver:

1. $MZ - EZ \leq \sqrt{\text{Var(Z)}}$
2. $EZ - MZ \leq \sqrt{\text{Var(Z)}}$

Como

$$MZ - EZ \leq \sqrt{\text{Var(Z)}} \Leftrightarrow P(Z \geq EZ + \sqrt{\text{Var(Z)}}) \leq \frac{1}{2}$$
$$\Leftrightarrow P(Z - EZ \geq \sqrt{\text{Var(Z)}}) \leq \frac{1}{2}.$$

Y por (1.2)

$$P(Z - EZ \geq \sqrt{\text{Var(Z)}}) \leq \frac{\text{Var(Z)}}{\text{Var(Z)} + \text{Var(Z)}} = \frac{1}{2}$$

se tiene 1.
Y del mismo modo se tiene para 2.

□

Terminamos esta sección con una desigualdad de carácter distinto, pero relacionada, porque muestra que es poco probable que una variable aleatoria tome valores muy por debajo de su valor medio.

1.4 Teorema (Desigualdad de Paley-Zygmund). *Para cualquier variable aleatoria no negativa Y con $a \in (0, 1)$,*

$$P(Y < aEY) \leq 1 - (1 - a^2)\frac{(EY)^2}{E[Y^2]}.$$

Demostración. Como

$$P(Y - EY \geq (a - 1)EY) = 1 - P(Y - EY < -(1 - a)EY)$$

y por (1.2) se tiene

$$P(Y - EY < -(1 - a)EY) \geq 1 - \frac{Var(Y)}{Var(Y) + (1 - a)^2(EY)^2} = \frac{(1 - a)^2(EY)^2}{Var(Y) + (1 - a)^2(EY)^2}$$
$$\geq (1 - a)^2\frac{(EY)^2}{EY^2}).$$

siendo la última desigualdad cierta pues $EY^2 - (1 - (1 - a)^2(EY)^2) \geq EY^2$

□

1.2. Método de Cramér-Chernoff

Este método determina la mejor cota posible para una cola de probabilidad que uno puede obtener usando la desigualdad de *Markov* con una función exponencial $\Phi(t) = e^{\lambda t}$ en (1.1). Sea Z una variable aleatoria real, para algún $\lambda > 0$ la desigualdad de *Markov* implica

$$P(Z \geq t) \leq e^{-\lambda t} E e^{\lambda Z}.$$

Como esta desigualdad es cierta para todo valor de $\lambda \geq 0$, se puede elegir λ tal que minimice la cota superior. Definiendo el logaritmo de la función generadora de momentos $\psi_Z(\lambda) = \log E e^{\lambda Z}$ para todo $\lambda \geq 0$, e introduciendo,

$$\psi *_Z (t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda)) \quad (\text{transformación de } \textit{Cramér} \text{ de } Z)$$

se obtiene la desigualdad de *Chernoff*

$$P(Z \geq t) \leq e^{-\psi *_Z (t)}.$$

Como $\psi_Z(0) = 0$, $\psi *_Z$ es una función no negativa. Si EZ existe entonces la convexidad de la función exponencial y la desigualdad de *Jensen* implica que $\psi_Z(\lambda) \geq EZ$ y además para valores negativos de λ, $\lambda t - \psi_Z(\lambda) \leq 0$ siempre que $t \geq EZ$, por lo que se puede extender el supremo a todo valor $\lambda \in \mathbb{R}$ en la definición de la transformacion de *Cramér*:

$$\psi *_Z (t) = \sup_{\lambda \in \mathbb{R}}(\lambda t - \psi_Z(\lambda)) \tag{1.3}$$

A la expresión a la derecha de la igualdad se la conoce como función dual de *Fenchel-Legendre* de ψ_Z. Así, para todo $t \geq EZ$, la transformación de *Cramér* $\psi *_Z (t)$ coincide con la función dual de *Fenchel-Legendre*.

Es claro que la desigualdad de *Chernoff* es trivial siempre que $\psi *_Z (t) = 0$. Esto ocurre si $\psi_Z(\lambda) = \infty$ para todo λ positivo o si $t \leq EZ$. Las desigualdades son no triviales si existe un $\lambda > 0$ tal que $E e^{\lambda Z} < \infty$. El resultado siguiente muestra una fórmula útil para la inversa de la función dual convexa de *Fenchel-Legendre*. La demostración de este resultado y de otros similares puede encontrarse en [1].

1.5 Lema. *Sea ψ una función convexa y continuamente diferenciable definida en el intervalo [0,b) donde $0 < b \leq \infty$. Se asume que $\psi(0) = \psi'(0) = 0$ y se considera*

$$\psi *_Z (t) = \sup_{\lambda \in (0,b)} (\lambda t - \psi(\lambda)) \quad \text{para todo} \quad t \geq 0.$$

Entonces $\psi*$ es una función convexa no negativa y no decreciente en $[0,\infty)$. Además, para todo $y \geq 0$ el conjunto $\{t \geq 0 : \psi*(t) > y\}$ es no vacío y la función inversa de $\psi*$ definida por

$$\psi *^{-1}(y) = \text{ínf}\{t \geq 0 : \psi *(t) > y\},$$

puede también escribirse como

$$\psi *^{-1}(y) = \inf_{\lambda \in (0,b)} \left[\frac{y + \psi(\lambda)}{\lambda}\right]$$

Como ilustración de la potencia del método de *Chernoff*, probamos a continuación una cota para probabilidades binomiales.

1.6 Teorema. *Sea D y n números enteros positivos con $1 \leq D \leq n/2$. Se cumple que*

$$\sum_{j=0}^{D} \binom{n}{j} \leq \left(\frac{en}{D}\right)^{D}$$

Demostración. Sea Z una variable aleatoria con distribución binomial de parámetros n y 1/2. Recuérdese que

$$P(Z = k) = \binom{n}{k}\frac{1}{2^n} \quad \text{con k=0,..,n}$$

y que por tanto

$$P(Z \leq D) = \frac{1}{2^n}\sum_{j=0}^{D}\binom{n}{k} = P(Z \geq n - D),$$

siendo la última igualdad cierta pues $Z \stackrel{d}{=} N - Z$. Z expresarse como suma de n variables de *Bernoulli* independientes $Z \stackrel{d}{=} Z_1 + ... + Z_n$ con $Z_i \sim B(1/2)$. Además por la desigualdad de *Chernoff* se tiene

$$P(Z \leq D) = P(Z \geq n - D) \leq e^{-\psi_Z^*(n-D)} \quad \text{con} \quad \psi_Z^*(\lambda) = \sup_{s>0}(\lambda s - \psi_Z(s))$$

y $\psi_Z(s) = \log Ee^{sZ}$. Por otro lado

$$Ee^{sZ} = E[e^{sZ_1}...e^{sZ_n}] = [Ee^{sZ_1}]^n = [\frac{1}{2} + \frac{1}{2}e^s] = [\frac{1}{2}(1 + e^s)]^n.$$

Un cálculo simple muestra que

$$\psi_Z^*(n-D) = (n-D)\log\frac{n-D}{D} - \log\frac{n}{D} + n\log 2.$$

Por lo tanto,

$$\begin{aligned} P(Z \leq D) &= \sum_{j=0}^{D} \binom{n}{k} \\ &\leq \frac{1}{2^n}\Big(\frac{n}{D}\Big)^n \Big(\frac{D}{n-D}\Big)^{n-D} \\ &\leq \Big(\frac{n}{D}\Big)^D \frac{1}{2^n}\left(1+\frac{D}{n-D}\right)^{n-D} \qquad \leq \frac{1}{2^n}\Big(\frac{n}{D}\Big)^D e^D, \end{aligned}$$

donde hemos usado que $1 + x \leq e^x, x \in \mathbb{R}$. Esto completa la demostración.

□

Muchas e importantes clases de variables aleatorias tienen colas de probabilidad que decrecen al menos tan rápidamente como variables aleatorias con distribución normal.

1.7 Definición. Una variable aleatoria centrada X se dice que es sub-Gaussiana con factor varianza ν si

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2} \quad \text{para todo} \quad \lambda \in \mathbb{R}.$$

Se denota por $\mathcal{G}(\nu)$ a la colección de variables aleatorias sub-Gaussianas.

Una variable aleatoria centrada $X \in \mathcal{G}(\nu)$ si la función generadora de momentos de X, está dominada por la función generadora de una variable aleatoria normal centrada de varianza ν. Además la desigualdad de *Chernoff* implica que si
$X \in \mathcal{G}(\nu)$,

$$\text{máx}\{P(X > t), P(-X > t)\} \leq e^{\frac{-t^2}{2\nu}}$$

El siguiente resultado caracteriza de forma más precisa cuándo una variable aleatoria es sub-gaussiana.

1.8 Teorema. *Sea X una variable aleatoria centrada. Si para algún $\nu > 0$*

$$\text{máx}\{P(X > t), P(-X > t)\} \leq e^{\frac{-t^2}{2\nu}} \quad \textit{para todo} \quad x > 0 \tag{1.4}$$

entonces para todo entero $q \geq 1$,

$$E[X^{2q}] \leq 2q!(2\nu)^q \leq q!(4\nu)^q. \tag{1.5}$$

Recíprocamente si para alguna constante positiva C, $E[X^{2q}] \leq q!C^q$, entonces $X \in \mathcal{G}(4C)$ (y por consiguiente (1.8) se cumple para $\nu = 4C$)

La condición para los momentos de X (1.5) es equivalente a otra condición usada a menudo como definición alternativa de variables aleatorias sub-gaussianas: Para algún $\alpha > 0$ se tiene que $Ee^{\alpha x^2} \leq 2$.

Las variables aleatorias acotadas son una clase importante de variables aleatorias sub-Gaussianas. La propiedad sub-Gaussiana para variables aleatorias acotadas se establece por el siguiente lema:

1.9 Lema (Lema de Hoeffding). *Sea Y una variable aleatoria con $EY = 0$ que toma valores en un intervalo acotado [a,b] y sea $\psi_Y(\lambda) = \log Ee^{\lambda Y}$. Entonces*

$$Var(Y) = \psi_Y''(\lambda) \leq \frac{(b-a)^2}{4} \quad \text{e} \quad Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right).$$

Demostración. En primer lugar nótese que

$$\left|Y - \frac{(b+a)}{2}\right| \leq \frac{(b-a)}{2}$$

y además,

$$\text{Var}(Y) = \text{Var}(Y - (a+b)/2) \leq \frac{(b-a)^2}{4}.$$

Sea $\mathbf{P}$ la distribución de Y y sea $\boldsymbol{P_\lambda}$ la distribución de probabilidad con densidad

$$x \to e^{-\psi_Y(\lambda)}e^{\lambda x}$$

con respecto a $\mathbf{P}$. Como $\boldsymbol{P_\lambda}$ está concentrado en [a,b], la varianza de la variable aleatoria Z con distribución $\boldsymbol{P_\lambda}$ acotada por $(b-a)^2/4$. Por consiguiente, mediante cálculos elementales se tiene,

$$\begin{aligned}\psi_Y''(\lambda) &= e^{-\psi_Y(\lambda)}E[Y^2e^{\lambda Y}] - e^{-2\psi_Y(\lambda)}(E[Ye^{\lambda Y}])^2 \\ &= \text{Var}(Y) \leq \frac{(b-a)^2}{4}.\end{aligned}$$

Como $\psi_Y(0) = \psi'_Y(0) = 0$, por el teorema de Taylor se tiene que para un $\zeta \in [0, \lambda]$,

$$\psi_Y(\lambda) = \psi_Y(0) + \lambda\psi'_Y(0) + \frac{\lambda^2}{2}\psi''_Y(\zeta)$$

lo que probaría la propiedad sub-gaussiana.

□

1.10 Definición. Una variable aleatoria centrada X se dice que es sub-exponencial en la cola derecha con factor varianza ν y parámetro escala c si

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2(1-c\lambda)} \quad \text{para todo} \quad \lambda \quad \text{tal que} \quad 0 < \lambda < \frac{1}{c}.$$

Se denota a la colección de tales variables aleatorias por $\Gamma_+(\nu, c)$.

De forma similar, una variable aleatoria centrada X se dice que es sub-exponencial en la cola izquierda con factor varianza ν y parámetro escala c si -X es sub-exponencial en la cola derecha con factor varianza ν y parámetro escala c. Se denota a la colección de tales variables aleatorias por $\Gamma_-(\nu, c)$.

1.11 Definición. X se dice que es simplemente sub-exponencial con factor varianza ν y parámetro escala c si X es sub-exponencial a ambos lados con el mismo factor varianza ν y parámetro escala c. La colección de tales variables aleatorias por $\Gamma(\nu, c)$. Obsérvese que $\Gamma(\nu, 0) = \mathcal{G}(\nu)$. Si se considera una variable aleatoria Y con distribución gamma y parámetros $a, b \geq 0$ entonces su versión centrada $X = Y - EY$ es un típico ejemplo de variable sub-exponencial.

Por aplicación directa del método de *Chernoff* se obtiene que si $X \in \Gamma(\nu, c)$, entonces para todo $t > 0$,

$$\text{máx}\{P(X > \sqrt{2\nu t} + ct), P(-X > \sqrt{2\nu t} + ct)\} \leq e^{-t}$$

El siguiente resultado proporciona un recíproco a este hecho.

1.12 Teorema. *Sea X una variable aleatoria centrada. Si para algún $\nu > 0$*

$$\text{máx}\{P(X > \sqrt{2\nu t} + ct), P(-X > \sqrt{2\nu t} + ct)\} \leq e^{-t} \quad \textit{para todo} \quad t > 0 \tag{1.6}$$

entonces para todo entero $q \geq 1$,

$$E[X^{2q}] \leq q!(8\nu)^q + (2q)!(4C)^{2q}. \tag{1.7}$$

Recíprocamente si para algunas constantes positiva A y B, $E[X^{2q}] \leq q!A^q + (2q)!A^{2q}$, entonces $X \in \Gamma(4(A + B^2), 2B)$ (y además (1.6) se cumple para $\nu = 4(A + B^2$ y $C = 2B)$.

1.3. Desigualdades maximales

Aquí se verá como la tranformación de *Cramér* de variables aleatorias de una colección finita puede ser usada para acotar la máxima esperanza de estas variables aleatorias.

1.13 Teorema. *Sean $Z_1, ..., Z_N$ variables aleatorias con valores reales tales que para todo $\lambda \in (0, b)$ e $i = 1, ..., N$, $\psi_{Z_i}(\lambda) = \log Ee^{\lambda Z_i} \leq \psi(\lambda)$ sonde ψ es una función convexa y continuamente diferenciable en $[0, b)$ con $0 < b \leq \infty$ tal que $\psi(0) = \psi'(0) = 0$. Entonces*

$$E \max_{i=1,...,N} Z_i \leq \psi *^{-1} (\log N).$$

En particular, si Z_i son variables aleatorias sub-Gaussianas con factor varianza ν, esto es $\psi(\lambda) = \lambda^2\nu/2$ para todo $\lambda \in (0, \infty)$, entonces

$$E \max_{i=1,...,N} Z_i \leq \sqrt{2\nu \log N}.$$

Demostración. Por la desigualdad de *Jensen*,

$$\exp\left(\lambda E \max_{i=1,...,N} Z_i\right) \leq E \exp\left(\lambda \max_{i=1,..,N} Z_i\right) = E \max_{i=1,..,N} \exp(\lambda Z_i)$$

para cualquier $\lambda \in (0, b)$. Así, definiendo $\psi_{Z_i}(\lambda) = \log E \exp(\lambda Z_i)$,

$$\exp(\lambda E \max_{i=1,...,N}) Z_i \leq \sum_{i=1}^{N} E \exp(\lambda Z_i) \leq N \exp(\psi(\lambda)).$$

Además, para cualquier $\lambda \in (0, b)$,

$$\lambda E \max_{i=1,..,N} Z_i - \psi(\lambda) \leq \log N,$$

lo que implica usando el lema (1.5), que

$$E \max_{i=1,..,N} Z_i \leq \inf_{\lambda \in (0,b)} \Big(\frac{\log N + \psi(\lambda)}{\lambda} \Big) = \psi^{*^{-1}}(\log N).$$

Ahora bien, como el inferior se alcanza cuando $\lambda = \sqrt{2v \log N/\nu}$ se tiene que

$$E \max_{i=1,..,N} Z_i \leq \sqrt{2\nu \log N}.$$

□

1.14 Corolario. *Sean $Z_1, ..., Z_N$ variables aleatorias con valores reales pertenecientes a $\Gamma_+(\nu, c)$. Entonces*

$$E \max_{i=1,...,N} Z_i \leq \sqrt{2\nu \log N} + c \ \log N.$$

Demostración. Por el Teorema anterior se tiene que

$$E \max_{i=1,..,N} Z_i \leq \inf_{\lambda \in (0,b)} \Big(\frac{\log N + \psi(\lambda)}{\lambda} \Big) = \psi^{*^{-1}}(\log N).$$

por lo que se trata de calcular la función $\psi^{*^{-1}}(\lambda)$. Como $\psi(\lambda) = \nu\lambda^2/(2-2c\lambda)$ se sigue mediante cálculos elementales que para todo $t > 0$

$$\psi * (t) = \sup_{\lambda \in (0,1/c)} \Big(t\lambda - \frac{\lambda^2 \nu}{2(1 - c\lambda)} \Big) = \frac{\nu}{c^2} h_1(\frac{ct}{\nu})$$

siendo $h_1(u) = 1 + u - \sqrt{1 + 2u}$ con $u > 0$. Como h_1 es una función creciente de $(0, \infty)$ a $(0, \infty)$ con función inversa $h_1^{-1}(u) = u + \sqrt{2u}$, finalmente se tiene que

$$\psi^{*^{-1}}(u) = \sqrt{2\nu u} + cu.$$

De modo que evaluando la función $\psi^{*^{-1}}(u)$ en $\log N$ se llega a

$$E \max_{i=1,...,N} Z_i \leq \sqrt{2\nu \log N} + c \ \log N.$$

1.4. Desigualdades clásicas para sumas de variables aleatorias independientes

En esta sección enunciamos algunas desigualdades de concentración clásicas para sumas de variables independientes. Estas desigualdades se obtienen mediante un uso más o menos directo del método de *Chernoff* y nos sirven como referencia para garantizaciones posteriores. Tratamos en primer lugar la desigualdad de *Hoeffding*, válida para sumandos acotados. Posteriormente enunciamos variantes (desigualdades de *Bennet* y de *Bernstein*) en las que se relaja la hipótesis de acotación.

Si $X_1, ..., X_N$ son variables aleatorias independientes con media finita tales que para algún intervalo I no vacío, $Ee^{\lambda X_i}$ es finito para todo $i \leq n$ y todo $\lambda \in I$ entonces definiendo

$$S = \sum_{i=1}^{n} (X_i - EX_i),$$

por independencia para todo $\lambda \in I$,

$$\psi_S(\lambda) = \sum_{i=1}^{n} \log Ee^{\lambda(X_i - E_i)}.$$

Si X_i toma valores en $[a_i, b_i]$ para todo $i \leq n$, entonces como

$$\left| S - \frac{(b_i - a_i)}{2} \right| \leq \sum_{i=1}^{n} \left| X_i - \frac{(b_i + a_i)}{2} \right| \leq \sum_{i=1}^{n} \frac{(b_i - a_i)}{2}$$

$$\text{y,} \quad Var(S) = Var\left(S - \frac{(b_i - a_i)}{2} \right) \leq \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{4}$$

entonces por el Lema de *Hoeffding*

$$\psi_S(\lambda) = \frac{\lambda^2}{2} \psi"_S(\theta) \leq \frac{\lambda^2}{2} \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{4} \quad \text{con} \quad \theta \in [0, \lambda].$$

Ahora un uso directo de la cota de *Chernoff* prueba el siguiente resultado

1.15 Teorema (**Desigualdad de Hoeffding**). *Sean $X_1, ..., X_N$ variables aleatorias tales que X_i toma valores en $[a_i, b_i]$ casi seguro para todo $i \leq n$. Sea $S = \sum_{i=1}^{n}(X_i - EX_i)$. Entonces para todo $t > 0$*

$$P(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Para relajar la hipótesis de acotación podemos observar que como la función generadora de momentos logarítmicos de las sumas de variables aleatorias independientes es igual a la suma de funciones generadoras de momentos logarítmicos de sumandos centrados, esto es,

$$\psi_S(\lambda) = \sum_{i=1}^{n}\left(\log Ee^{\lambda X_i} - \lambda EX_i)\right).$$

usando que $\log\ u \leq u - 1$ para todo $u > 0$, se obtiene

$$\psi_S(\lambda) \leq \sum_{i=1}^{n} E[e^{\lambda X_i} - \lambda X_i - 1)].$$

Las desigualdades de *Bennet* y de *Bernstein*, enunciadas a continuación pueden ser derivadas de esta cota bajo condiciones diferentes de integrabilidad para los X_i.

1.16 Teorema (**Desigualdad de Bennet**). *Sean $X_1, ..., X_N$ variables aleatorias independientes con varianza finita tal que $X_i \leq b$ para algún $b > 0$ casi seguro para todo $i \leq n$. Sea $S = \sum_{i=1}^{n}(X_i - EX_i)$ y $\nu = \sum_{i=1}^{n} E[X_i^2]$. Si escribimos $\phi(u) = e^u - u - 1$ para $u \in \mathbb{R}$, entonces, para todo $\lambda > 0$,*

$$\log Ee^{\lambda s} \leq n \log\left(1 + \frac{\nu}{nb^2}\phi(b\lambda)\right) \leq \frac{\nu}{b^2}\phi(b\lambda),$$

y para cualquier $t > 0$,

$$P(S \geq t) \leq \exp\left(-\frac{\nu}{b^2}h(\frac{bt}{\nu})\right)$$

donde $h(u) = (1 + u)\log(1 + u) - u$ para $u > 0$.

La desigualdad $h(u) \geq \frac{u^2}{2(1+u/3)}$ implica que bajo las condiciones del anterior Teorema

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(\nu + bt/3)})\right)$$

desigualdad conocida como desigualdad de *Bernstein*.

1.17 Teorema (**Desigualdad de Bernstein**). *Sean $X_1, ..., X_N$ variables aleatorias independientes con valores reales. Se asume que existen números positivos ν y c tales que $\sum_{i=1}^{n} E[X_i^2] \leq \nu$ y*

$$\sum_{i=1}^{n} E[(X_i)_+^q] \leq \frac{q!}{2}\nu c^{q-2} \quad \text{para todo entero} \quad q \geq 3,$$

donde $x_+ = \text{máx}(x, 0)$.
Si $S = \sum_{i=1}^{n}(X_i - EX_i)$ entonces para todo $\lambda \in (0, 1/c)$ y $t > 0$,

$$\psi_S(\lambda) \leq \frac{\nu\lambda^2}{2(1 - c\lambda)}$$

y

$$\psi_S * (t) \geq \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)$$

donde $h_1(u) = (1 + u) - \sqrt{(}1 + 2u)$ para $u > 0$. En particular, para todo $t > 0$,

$$P(S \geq \sqrt{2\nu t} + ct) \leq e^{-t}.$$

1.18 Corolario. *Sean $X_1, ..., X_N$ variables aleatorias independientes con valores reales que satisfacen las condiciones del anterior teorema y sea $S = \sum_{i=1}^{n}(X_i - EX_i)$. Entonces para todo $t > 0$,*

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(\nu + ct)}\right)$$

Se puede ver que cuando t se hace lo suficientemente grande la desigualdad dada en este último corolario tiene el mismo comportamiento que e^{-t} en vez de e^{-t^2}, que es el comportamiento que garantiza la desigualdad de *Hoeffding*. También se puede ver que para grandes valores de t la desigualdad de *Bernstein* pierde el factor logarítimico en el exponente, en cambio cuando ν/b tiene un valor moderado las desigualdades de *Bennet* y de *Bernstein* tienen un comportamiento similar.

Capítulo 2

Métodos de Martingala

El resultado principal en este capítulo es la desigualdad de *Efron-Stein*, una desigualdad para las varianzas de funciones de variables aleatorias independientes. Esto se describe en la sección (2.1). En algunos casos es posible obtener también desigualdades exponenciales a partir de este método,como se puede ver en la sección (2.2). La sección (2.3) particulariza la desigualdad de *Efron-Stein* a algunas funciones de tipo especial.

Finalmente, en la sección (2.4) se muestran dos ejemplos de aplicación: una desigualdad de concentración para autovalores de matrices aleatorias y una prueba de la desigualdad de *Efron-Stein*.

Formalmente, sea $f : \mathcal{X}^n \longrightarrow \mathbb{R}$ una función con valores reales de n variables donde $\mathcal{X}$ es algún espacio medible. Si $X_1, ..., X_n$ son variables aleatorias independientes que toman valores en $\mathcal{X}$, entonces podemos definir la variable aleatoria con valores reales $Z = f(X_1, ..., X_n)$. Cabe señalar que X_i puede tener diferentes distribuciones, la única suposición esencial es la independencia. Se asume que Z tiene varianza finita y el propósito será encontrar cotas superiores generales.

2.1. La desigualdad de Efron-Stein

La desigualdad de *Efron-Stein* proporciona una cota en términos de variaciones locales de la función f. La demostración de la desigualdad de *Efron-Stein* está basada en un argumento de martingalas, que dará lugar a a generalizaciones presentadas en secciones posteriores.

Antes que dar una cota para la varianza de funciones generales $Z = f(X_1, ..., X_n)$ se considerará el caso especial en el que las variables $X_1, ..., X_n$ tienen valores reales y $Z = X_1 + ... + X_n$. En este caso se tiene Var(Z) = $\sum_{i=1}^{n} Var(X_i)$.

En primer lugar cabe hacer un comentario acerca de la notación que se va a utilizar a lo largo de esta sección.

Notación. Para cualquier función integrable $Z = f(X_1, ..., X_n)$, definimos

$$E_i = \int_{\mathbb{X}^{n-i}} f(X_1, ..., X_i, x_{i+1}, ..., x_n) d\mu_{i+1}(x_{i+1}), ..., d\mu_n(x_n)$$

donde, para todo $i = 1, ..., n$, μ_i denota la distribución de probabilidad de X_i. Por el Teorema de *Fubini* E_i está bien definido y es finito con probabilidad 1. Además denotamos por $E^{(i)}$ la esperanza condicionada a $X^{(i)} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$, es decir,

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

Nuestra presentación de la desigualdad de *Efron-Stein* se basa en el uso del método de martingala. Detallamos a continuación algunos aspectos básicos de este concepto. Sea un espacio de probabilidad definido por $(\Omega, \mathbb{F}, P)$, donde Ω es el espacio de muestra, $\mathcal{F}$ es la σ-álgebra asociada a Ω y P es la medida de probabilidad. Se llama filtración a una familia $\{\mathcal{F}_t\}_{t>0}$ de sub-σ-álgebras tal que $\mathcal{F}_t \subset \mathcal{F}_r$ si $t < r$. En las misma condiciones nombradas anteriormente sea $\mathbb{F}$ una filtración de σ-álgebras: $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_T \subset \mathcal{F}$. Sea $\{X_n\} = X_1, .., X_n$ una sucesión de variables aleatorias que forman un proceso estocástico. Entonces, el proceso estocástico $\{X_n, n > 0\}$ adaptado a la filtración $\mathbb{F}$ recibe el nombre de martingala si $E(X_m|\mathcal{F}_n) = X_n$ con $m \geq n$ y donde $\mathcal{F}_s$ es cualquier sub-σ- álgebra de la filtración $\mathbb{F}$.

Para acotar la varianza de una función general expresaremos $Z - EZ$ como una suma de diferencias de martingalas para la filtración de *Doob* y usaremos la ortogonalidad de estas diferencias (en $\mathbb{L}_2$). La filtración de *Doob* es la "natural", es decir, si $X_1, ..., X_n$ son variables aleatorias independientes y F_i es la sigma-álgebra generada por $X_1, ..., X_i$ la filtración de *Doob* es la colección $F_1, ..., F_n$. Si denotamos como E_i la esperanza condicionada a $X_1, ..., X_i$ y se asume que $E_0 = E$, se puede definir

$$\Delta_i = E_i Z - E_{i-1} Z$$

para todo $i = 1, ..., n$. Así,

$$Z - EZ = \sum_{i=1}^{n} \Delta_i$$

y se tiene

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

Ahora bien si $j > i$

$$\begin{aligned} E_i\Delta_j &= E\left[\Delta_j | X^{(i)}\right] \\ &= E\left[E_j Z - E_{j-1} Z | X^{(i)}\right] \\ &= E\left[E(Z|X_1, ..., X_j) - E(Z|X_1, ..., X_{j-1})|X_1, ..., X_i\right] \\ &= E(Z|X_1, ..., X_i) - E(Z|X_1, ..., X_i) = 0. \end{aligned}$$

lo que implica que

$$E_i[\Delta_j \Delta_i] = \Delta_i E_i \Delta_j$$

y por eso $E[\Delta_j \Delta_i] = 0$. Por consiguiente, se obtiene la siguiente fórmula de aditividad de la varianza:

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

Recordando la notación anteriormente definida y haciendo uso del Teorema de *Fubini* se tiene

$$E_i[E^{(i)Z}] = E_{i-1} Z. \tag{2.1}$$

2.1 Teorema (Desigualdad de Efron-Stein). *Sean $X_1, ..., X_n$ variables aleatorias independientes y sea $Z = f(X_1, ..., X_n)$ una función de cuadrado integrable. Entonces*

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

Además si $X_1', ..., X_n'$ son copias independientes de $X_1, ..., X_n$ y se define para todo $i = 1, ..., n$,

$$Z_i' = f(X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n)$$

entonces

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

donde $x_+ = \text{máx}(x, 0)$ *y* $x_- = \text{máx}(-x, 0)$ *denotan la parte positiva y negativa del número real* x*.Asimismo,*

$$\nu = \text{ínf} \sum_{i=1}^{n} E\left[(Z - Z_i)^2\right],$$

donde el ínfimo se toma entre la clase de todas las variables Z_i *que son* $X^{(i)}$ *medibles y de cuadrado integrables* $i = 1, ..., n$*.*

Demostración. Se empezará demostrando el primer enunciado. Nótese que haciendo uso de (2.1) se puede escribir

$$\Delta_i = E_i[Z - E^{(i)}Z].$$

Ahora bien, por la desigualdad de *Jensen*

$$\Delta_i^2 \leq E_i[(Z - E^{(i)}Z)^2]$$

y como $\text{Var(Z)} = \sum_{i=1}^{n} E[\Delta_i^2]$ se obtiene que

$$\text{Var(Z)} \leq \sum_{i=1}^{n} [(Z - E^{(i)}Z)^2].$$

Para probar la igualdad para ν, sea

$$\begin{aligned}
\text{Var}^{(i)}Z &= (\text{VarZ}|X^{(i)}) \\
&\overset{def}{=} E\left[(Z - E[Z|X^{(i)}])^2|X^{(i)}\right] \\
&= E\left[(Z - E^{(i)}Z)^2|X^{(i)}\right] \\
&= E^{(i)}\left[(Z - E^{(i)}Z)^2\right].
\end{aligned}$$

Entonces,

$$\nu = \sum_{i=1}^{n} E\left[(Z - E^{(i)}Z)^2\right].$$

Para probar el segundo enunciado nótese que si X e Y son variables aleatorias independientes idénticamente distribuidas con valores reales entonces

$$E[(X - Y)^2] = E[X^2 + Y^2 - 2XY] = 2E[X^2] - 2[EX]^2 = 2\text{Var(X)},$$

es decir

$$\mathrm{Var(X)} = \frac{1}{2}E[(X-Y)^2]$$

volviendo al enunciado como Z_i' es una copia independiente de Z y usando la condicionalidad a $X^{(i)}$ se tiene

$$\mathrm{Var}^{(i)}Z = \frac{1}{2}E^{(i)}[(Z-Z_i')^2]$$

y además esto equivale a

$$E^{(i)}\left[(Z-Z_i')_+^2\right] = E^{(i)}\left[(Z-Z_i')_-^2\right].$$

La última igualdad se obtiene del hecho de que para toda variable aleatoria real X, $\mathrm{Var(X)} = \text{ínf}_{a\in\mathbb{R}}\, E[(X-a)^2]$. Y usando la condicionalidad a $X^{(i)}$ sobre este argumento se tiene para todo $i = 1, ..., n$

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

Nótese que el ínfimo se alcanza siempre que $Z_i = E^{(i)}Z$.

□

Obsérvese que si $\sum_{i=1}^n X_i$ con $X_1, ..., X_n$ variables aleatorias independientes (con varianza finita) la desigualdad de *Efron-Stein* se convierte en igualdad.

2.2. Desigualdad exponencial

El propósito de esta sección es ver dos formas diferentes en las que la desigualdad de *Efron-Stein* puede ser usada de una forma simple para obtener cotas exponenciales en colas de probabilidad de funciones con diferencias acotadas, las cuales se verán con más detenimiento en (2.3). En los argumentos, de hecho, basta con una condición más suave que la de diferencias acotadas, la propiedad que existe una constante positiva ν tal que

$$\sum_{i=1}^n (Z-Z_i)_+^2 \leq \nu \tag{2.2}$$

ocurre con probabilidad 1.

Un camino para obtener cotas exponenciales es el siguiente: se aplica la desigualdad de *Efron-Stein* a $e^{\lambda Z/2}$ con $\lambda > 0$ y posteriormente el teorema del valor medio

$$\begin{aligned} Ee^{\lambda Z} - (E[e^{\lambda Z/2}]^2) &\leq E\Big[\sum_{i=1}^{n}(e^{\lambda Z/2} - e^{\lambda Z_i'/2})_+^2\Big] \\ &\leq \frac{\lambda^2}{4} E\sum_{i=1}^{n}\Big[e^{\lambda Z}(Z - Z_i')_+^2\Big] \end{aligned}$$

Ahora se puede hacer uso de la condición (2.2) y deducir que

$$Ee^{\lambda Z} - (E[e^{\lambda Z/2}]^2) \leq \frac{\nu\lambda^2}{4} Ee^{\lambda Z}$$

o equivalentemente que

$$\Big(1 - \frac{\nu\lambda^2}{4}\Big)F(\lambda) \leq (F(\lambda/2))^2,$$

donde $F(\lambda) = Ee^{\lambda(Z-EZ)}$ o equivalentemente

$$\Big(1 - \frac{\nu\lambda^2}{4}\Big)F(\lambda) \leq (F(\lambda/2))^2,$$

con $F(\lambda) = Ee^{\lambda(Z-EZ)}$. Esta última desigualdad controla la función generadora de momentos y su solución se prueba haciendo uso del siguiente lema cuya demostración se puede ver en [1].

2.2 Lema. *Sea* $g : (0,1) \to (0,\infty)$ *una función tal que el límite* $\lim_{x\to 0}(g(x)-1)/x = 0$. *Si para todo* $x \in (0,1)$

$$(1 - x^2)g(x) \leq g(x/2)^2,$$

entonces

$$g(x) \leq (1 - x^2)^{-2}.$$

Como $F(0) = 1$ y $F'(0) = 0$, $\lim_{\lambda\to 0}(F(\lambda) - 1)/\lambda = 0$ y se puede aplicar el lema (2.2) a la función que envía x en $F(2x\nu^{-1/2})$ y se llega a

$$F(\lambda) \leq \Big(1 - \frac{\lambda^2\nu}{4}\Big)^{-2} \tag{2.3}$$

para todo $\lambda \in (0, 2\nu^{-1/2})$. Por tanto, la desigualdad de *Efron-Stein* puede ser usada para probar la integrabilidad exponencial de Z. Además, como por (2.3) $F(\nu^{-1/2}) \leq 2$, por la desigualdad de *Markov*, para todo $t > 0$,

$$P(Z - EZ \geq t) \leq 2e^{-t/\sqrt{\nu}}. \tag{2.4}$$

Otra forma de explotar (2.3) es acotar $-\log(1-u)$ por $u(1-u)^{-1}$ y concluir que para todo $\lambda \in (0, 2\nu^{-1/2})$

$$\log F(\lambda) \leq \frac{\lambda^2\nu}{2(1-(\lambda^2\nu/4))}) \leq \frac{\lambda^2\nu}{2(1-(\lambda\sqrt{\nu}/2))}.$$

Esta cota para la función generadora de momentos señala que $Z-EZ$ es una variable aleatoria sub-exponencial con factor varianza ν y parámetro escala $c = \sqrt{\nu}/2$. Como se vió en la sección (1.2) del primer capítulo, para todo $t > 0$,

$$P(Z - EZ \geq \sqrt{2\nu t} + ct) \leq e^{-t}. \tag{2.5}$$

Como $c = \sqrt{\nu}/2$,se puede ver que si t no es muy pequeño ($t \geq 1$), el término lineal en la expresión $\sqrt{2\nu t} + ct$ domina a la otra. Por esta razón no se puede considerar a (2.5) como una desigualdad sub-gaussiana.

En las siguiente subsecciones se hará uso de la desigualdad de *Efron-Stein* para ejemplos típicos.

2.3. Algunos casos especiales

2.3 Definición. Se dice que una función $f : \mathcal{X}^n \longrightarrow \mathbb{R}$ tiene la propiedad de diferencias acotadas si para alguna constante no negativa $c_1, ..., c_n$,

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

En otras palabras si se cambia la i-ésima variable de f mientras se mantengan el resto fijas, el valor de la función no cambiará más que c_i.

La desigualdad de *Efron-Stein* implica lo siguiente

2.4 Corolario. *Si f tiene la propiedad de diferencias acotadas con constantes $c_1, ..., c_n$, entonces*

$$\text{Var(Z)} \leq \frac{1}{4}\sum_{i=1}^{n} c_i^2$$

Demostración. Por la desigualdad de *Efron-Stein*,

$$\text{Var(Z)} \leq \inf_{Z_i} \sum_{i=1}^{n} E[(Z - Z_i)^2]$$

donde el ínfimo se toma entre la clase de todas las variables Z_i que son $X^{(i)}$ medibles y de cuadrado integrables $i = 1, ..., n$.. Entre ellas elegimos

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

Y como

$$(Z - Z_i)^2 \leq \frac{c_i^2}{4} \qquad \text{y} \qquad \text{Var(Z)} \leq \sum_{i=1}^{n} E[Z - Z_i]^2$$

se sigue que

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

Además, si se aplica (2.4) se tiene

$$P(Z - EZ \geq t) \leq 2e^{\frac{-2t}{\sqrt{\sum_{i=1}^{n} c_i^2}}}.$$

□

2.5 Definición. Se dice que una función no negativa $f : \mathcal{X}^n \longrightarrow [0, \infty)$ tiene la propiedad de ser autoacotante si existen funciones $f_i : \mathcal{X}^{n-1} \longrightarrow \mathbb{R}$ tales que para todo $x_1, ..., x_n \in \mathcal{X}$ y todo $i = 1, ..., n$

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

y también

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

Para funciones autoacotantes claramente se tiene,

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

y además la última expresión de ν en el Teorema de *Efron-Stein* implica lo siguiente:

2.6 Corolario. *Si f tiene la propiedad de ser autoacotante, entonces*

$$\text{Var(Z)} \leq EZ$$

Demostración. Por el Teorema de *Efron-Stein* se tiene,

$$\begin{aligned}\text{Var(Z)} &\leq E\left[\sum_{i=1}^{n}(f(x_1,...,x_n)-f_i(x_1,...,x_{i-1},x_{i+1},...,x_n))^2\right]\\ &\leq E\left[f(x_1,...,x_n)\right]=EZ.\end{aligned}$$

□

Una clase importante de funciones que satisfacen la propiedad de ser autoacotante son las llamadas funciones configuración.

2.7 Definición. Suponemos que tenemos una propiedad Π definida sobre la unión de productos finitos de un conjunto $\mathcal{X}$, esto es, una secuencia de conjuntos

$$\Pi_1 \subset \mathcal{X},\ \Pi_2 \subset \mathcal{X}^2, ..., \Pi_n \subset \mathcal{X}^n.$$

Se dice que $(x_1,...,x_m) \in \mathcal{X}^m$ satisface la propiedad Π si $(x_1,...,x_m) \in \Pi_m$. Se asume que Π es hereditaria en el sentido que si $(x_1,...,x_m)$ satisface Π entonces cualquier sub-secuencia $(x_{i_1},...,x_{i_k})$ también lo hace.

2.8 Definición. La función $f : \mathcal{X}^n \longrightarrow \mathbb{N}$ de Π que asigna cualquier vector $x = (x_1,...,x_n)$ la longitud de la sub-secuencia más grande en Π es una función configuración asociada a la propiedad Π, es decir,

$$f(x_1,...,x_n) = \text{máx}\{k \in \mathbb{N} : \exists i_1 \leq ... \leq i_k \in \{1,..,n\}/(x_{i_1},...,x_{i_k}) \in \Pi\}\}$$

2.9 Corolario. *Sea f una función configuración y sea $Z = f(X_1,...,X_n)$ donde $X_1,...,X_n$ son variables aleatorias independientes entonces*

$$\text{Var(Z)} \leq EZ$$

Demostración. Haciendo uso del Corolario (2.6) es suficiente ver que cualquier función de configuración tiene la propiedad de ser autoacotante. Sea $Z_i = f(X^{(i)}) = f(X_1,...,X_{i-1},X_{i+1},...,X_n)$. La condición $0 \leq Z - Z_i \leq 1$ se satisface trivialmente. Por otra parte, se asume que $Z = k$ y sea $\{X_{i_1},...,X_{i_k}\} \subset \{X_1,...,X_n\}$ una subsecuencia (existe pues f es función de configuración) de cardinal k tal que $f_k(X_{i_1},...,X_{i_k}) = k$. Claramente, si el índice i es tal que $i \notin \{i_1,...,i_k\}$ entonces $Z = Z_i$ y además

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

también se satisface, lo que concluye la prueba pues f tendría la propiedad de ser autoacotante.

□

2.4. Ejemplos y aplicaciones

Para ilustrar el hecho de que las funciones configuración aparecen de forma natural en numerosas aplicaciones a continuación se verán algunos ejemplos en diferentes campos.

2.4.1. El mayor autovalor de una matriz simétrica aleatoria

Sea A una matriz simétrica real cuyas entradas $X_{i,j} \quad 1 \leq i \leq j \leq n$ son variables aleatorias independientes con valor absoluto acotado por 1. Sea $Z = \lambda_1$ el mayor autovalor de A. Esta propiedad se necesita para poder acotar la varianza de Z, esto es, si $v = (v_1, ..., v_n) \in \mathbb{R}$ es el autovector asociado a λ_1 con $||v|| = 1$, entonces

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

Se considera ahora la matriz simétrica $A'_{i,j}$ obtenida de reemplazar $X_{i,j}$ en A por una copia independiente $X'_{i,j}$ mientras se mantiene el resto de las variables fijas. Sea $Z'_{i,j}$ el mayor autovalor de la matriz obtenida. Entonces se tiene

$$\begin{aligned}(Z - Z'_{i,j})_+ &\leq (v^T A v - v^T A'_{i,j} v)\mathbb{1}_{\{Z > Z'_{i,j}\}} \\ &= (v^T (A - A'_{i,j}) v)\mathbb{1}_{\{Z > Z'_{i,j}\}} \\ &\leq 2(v_i v_j (X_{i,j} - X'_{i,j}))_+ \\ &\leq 4|v_i v_j|.\end{aligned}$$

Y por consiguiente,

$$\sum_{1 \leq i \leq j \leq n} (Z - Z'_{i,j})_+^2 \leq \sum_{1 \leq i \leq j \leq n} 16|v_i v_j|^2 \leq 16 \left(\sum_{1 \leq i \leq j \leq n} v_i^2 \right)^2 = 16$$

Tomando esperanza a ambos lados y usando la desigualdad de *Efron-Stein* se tiene

$$\text{Var(Z)} \leq \sum_{i=1}^{n} E[(Z - Z'_{i,j})^2] \leq 16.$$

En consecuencia, la varianza está acotada por una constante independiente del tamaño de la matriz y de la distribución de las entradas. La única condición que se necesita es la independencia y la acotación de las entradas; no tienen que tener la misma distribución. Además, si se aplica (2.4) se tiene

$$P(Z - EZ \geq t) \leq 2e^{-t/4}.$$

La desigualdad de *Efron-Stein* se puede aplicar satisfactoriamente para probar una cota fuerte para la varianza de una función suave de un vector aleatorio con distribución normal estándar.

2.4.2. Desigualdad gaussiana de Poincaré.

Sea $X = (X_1, ..., X_n)$ un vector de variables aleatorias independientes igualmente distribuidas con distribución normal estándar (es decir, $X \sim \gamma_n$, $\gamma_n \sim N(0, I_n)$).
Sea $f : \mathbb{R}^n \to \mathbb{R}$ cualquier función de clase $\mathbb{C}^1$. Entonces

$$\mathrm{Var}(f(X)) \leq \mathrm{E}[||\nabla f(X)||^2] = \int_{\mathbb{R}^n} ||\nabla f(X)||^2 d\gamma_n(x).$$

Demostración. Se asume que $\mathrm{E}[||\nabla f(X)||^2] < \infty$, pues en otro caso la desigualdad es trivial. Sean $\epsilon_1, ..., \epsilon_n$ variables aleatorias independientes *Rademacher*, es decir, $P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$ con $i = 1, ..., n$, y se define $S_n = 1/(\sqrt{n}) \sum_{j=1}^{n} \epsilon_j$. En primer lugar veamos qué ocurre cuando f es una función de soporte compacto y en segundo lugar se hará uso de que toda función se puede extender a una de soporte compacto.

- Primer caso: Suponemos $f \in \mathscr{C}_c^2(\mathbb{R})$. Por la desigualdad de Efron Stein se tiene

  $$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathrm{E}[\mathrm{Var}^{(i)}](Z)$$

  y como se puede escribir

  $$\mathrm{Var}^{(i)}(f(S_n)) = \frac{1}{4}\Big(f\Big(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\Big) - f\Big(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\Big)\Big)^2$$

  uniendo estos dos resultados se llega a

  $$Var(f(S_n)) \leq \frac{1}{4}\sum_{i=1}^{n} \mathrm{E}\Big(f\Big(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\Big) - f\Big(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\Big)\Big)^2 \quad (2.6)$$

pues $Z = f(X) \sim f(S_n)$. El teorema central del límite implica que S_n converge en distribución a X, donde X tiene una distribución normal estándar. Por consiguiente $\mathrm{Var}(f(S_n))$ converge a $\mathrm{Var}(f(X))$. Se considera el desarrollo de Taylor de f en torno a S_n,

$$f(Y) = f(S_n) + f'(S_n)(Y - S_n) + R^2(f)$$

con $R^2(f)$ el resto de Taylor de orden dos, es decir,

$$R^2(f) = f^{II)}(\zeta)\frac{(Y - S_n)^2}{2!}.$$

Sea K el supremo del valor absoluto de la segunda derivada de f y haciendo uso ahora del corolario 2.4.1 de [3] se tiene que para todo i,

$$\left|f\Big(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\Big) - f\Big(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\Big)\right| \leq \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{2K}{n}$$

y además,

$$\frac{n}{4}\Big(f\Big(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\Big) - f\Big(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\Big)\Big)^2 \leq f'(S_n)^2 + \frac{2K}{\sqrt{n}}|f'(S_n)| + \frac{K^2}{n}.$$

Esto y el teorema central del límite no llevan a que

$$\limsup_{n\to\infty} \sum_{i=1}^{n} E\Big[\Big(f\Big(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\Big) - f\Big(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\Big)\Big)^2\Big] = E[f'(X)^2],$$

y haciendo uso de (2.6) se llega a que

$$\mathrm{Var}(f(X)) \leq E[f'(X)].$$

- Segundo caso: $f \in \mathscr{C}^1(\mathbb{R})$. Suponemos sin pérdida de generalidad que $\int_{-\infty}^{\infty} f(x)\varphi(x)dx = 0$ entonces

$$\mathrm{Var}(f(x)) = \int_{-\infty}^{\infty} f(x)^2\varphi(x)dx = \lim_{k\to\infty}\int_{-k}^{k} f(x)^2\varphi(x)dx.$$

Además,

$$\int_{-k}^{k} f(x)^2\varphi(x)dx = \lim_{\epsilon\to 0}\int_{\mathbb{R}} f_\epsilon(x)^2\varphi(x)dx,$$

donde

$$f_\epsilon(x) = \int_{-\epsilon}^{\epsilon} \frac{1}{\epsilon}\eta\Big(\frac{y}{\epsilon}\Big) f\,|_{[-k,k]}\,(x-y) \quad \text{y}$$

$$\eta(y) = \begin{cases} C \exp\left(\frac{1}{x^2-1}\right) & \text{si } |x| < 1 \\ 0 & \text{resto} \end{cases}$$

y C elegida de forma que $\int_{\mathbb{R}} \eta(y)dy = 1$. Por el Teorema 6 página 630 de [4] se tiene,

$$\lim_{\epsilon \to 0} \int_{-\infty}^{\infty} f_\epsilon(x)\varphi(x)dx = \int_{-k}^{k} f(x)^2\varphi(x)dx.$$

y

$$\lim_{\epsilon \to 0} \int_{-\infty}^{\infty} f_\epsilon(x)^2\varphi(x)dx = \int_{-k}^{k} f(x)^2\varphi(x)dx.$$

Por lo tanto, $\lim_{\epsilon \to 0} \mathrm{Var}(f_\epsilon(x)) = \mathrm{Var}(f\mid_{[-k,k]}(x))$. Por otro lado, por el primer caso (porque f_ϵ es $\mathscr{C}_c^\infty(\mathbb{R})$)

$$\mathrm{Var}(f_\epsilon(x)) \leq \int_{\mathbb{R}} (f_\epsilon')^2(x)\varphi(x)dx.$$

Pero también se tiene (de nuevo por el Teorema 6 página 630 de [4])

$$\lim_{\epsilon \to 0} \int_{\mathbb{R}} (f_\epsilon'(x))^2\varphi(x)dx = \int_{-k}^{k} f'(x)^2\varphi(x)dx.$$

De aquí concluimos que

$$\int_{-k}^{k} f(x)^2\varphi(x)dx \leq \int_{-k}^{k} f'(x)^2\varphi(x)dx.$$

Ahora bien, por el teorema de la convergencia monótona

$$\lim_{k \to \infty} \int_{-k}^{k} f(x)^2\varphi(x)dx = \int_{-\infty}^{\infty} f(x)^2\varphi(x)dx$$

$$\lim_{k \to \infty} \int_{-k}^{k} f'(x)^2\varphi(x)dx = \int_{-\infty}^{\infty} f'(x)^2\varphi(x)dx$$

y concluimos que

$$\int_{-\infty}^{\infty} f(x)^2\varphi(x)dx \leq \int_{-\infty}^{\infty} f'(x)^2\varphi(x)dx$$

como queríamos probar.

□

Capítulo 3

Entropía de variables aleatorias

En este capítulo se introducirá la noción de entropía relativa que es una herramienta clave para probar desigualdades de concentración. Muchas de las más importantes desigualdades de concentración se pueden obtener por el método de la entropía.

En la sección (3.1) se presentan varios resultados de representación de la entropía relativa mediante fórmulas variacionales o de dualidad. Como consecuencia de estas representaciones, en la sección (3.2) se demuestra el "Lema de Transporte", un resultado clave para relacionar la entropía relativa con la concentración de variables aleatorias. Una aplicación de este lema se presenta en la sección (3.3) , donde se demuestra la desigualdad de *Pinsker*. Finalmente, la sección (3.4) prueba una importante propiedad de sub-aditividad de la entropía de variables independientes. En este caso presentamos el resultado en la forma general Φ- entropías, que incluyen el caso de la entropía clásica, pero también la desigualdad de *Efron-Stein* como casos particulares.

Si Φ denota la función $\Phi(x) = x\log(x)$ definida en $[0,\infty)$ (donde $0\log 0$ se define como 0) y $(\Omega, \mathcal{A}, \mathbf{P})$ es un espacio de probabilidad, entonces Y una variable aleatoria no negativa definida en tal espacio y tal que Y sea integrable, esto es, si $EY = \int_\Omega Y(\omega)d\mathbf{P}(\omega) < \infty$, se define la entropía de Y por

$$\mathrm{Ent}(Y) = E\Phi(Y) - \Phi(EY). \tag{3.1}$$

$\mathrm{Ent}(Y)$ está bien definida para toda variable aleatoria no negativa.

Como Φ es una función convexa, por la desigualdad de *Jensen*, $\mathrm{Ent}(Y)$ es una cantidad no negativa (quizás infinita). Además $\mathrm{Ent}(Y) < \infty$ si y sólo si $\Phi(Y)$ es integrable.

Si Y es una variable aleatoria no negativa con $EY = 1$, se puede definir otra medida de probabilidad en $(\Omega, \mathcal{A}, \mathbf{P})$ por $\mathbf{Q}(A) = \int_A Y(\omega) d\mathbf{P}(\omega) = E[Y\mathbb{1}_A]$ para todo $A \in \mathcal{A}$. Escribimos $\mathbf{Q} = Y\mathbf{P}$ para tal medida de probabilidad. Entonces la divergencia de *Kullback-Leibler*(o entropía relativa) de $\mathbf{Q}$ respecto a $\mathbf{P}$ se define por $D(\mathbf{Q}||\mathbf{P}) = \mathrm{Ent(Y)}$

3.1. Dualidad y fórmulas variacionales

El siguiente resultado da una caracterización alternativa a la entropía relativa.

3.1 Teorema (**Fórmula de dualidad de Entropía**). *Sean Y una variable aleatoria no negativa definida en el espacio de probabilidad $(\Omega, \mathcal{A}, \boldsymbol{P})$ tal que $E\Phi(Y) < \infty$. Entonces se tiene la fórmula de la dualidad*

$$\mathrm{Ent(Y)} = \sup_{U \in \mathcal{U}} E[UY]$$

donde el supremo se toma sobre el conjunto $\mathcal{U}$ de todas las variables aleatorias $U : \Omega \longrightarrow \bar{\mathbb{R}}$ con $Ee^U = 1$. Además si U es tal que $E[UY] \leq \mathrm{Ent(Y)}$ para toda variable aleatoria no negativa Y tal que $\Phi(Y)$ es integrable y $EY = 1$, entonces $Ee^U \leq 1$.

3.2 Comentarios.

- Bajo el estudio de la función

  $$g(x) = xu - x\log x \quad \text{con} \quad u \in \mathbb{R}, x > 0$$

  se tiene que e^{u-1} es un punto crítico de $g(x)$, es decir,

  $$\sup_{x>0}(xu - \Phi(x)) = e^{u-1}.$$

  Por tanto, si $\Phi(Y)$ es integrable y $Ee^u = 1$, se tiene,

  $$UY \leq \Phi(Y) + \frac{1}{e}e^u.$$

  Además, U_+Y es integrable y se puede definir $E[UY]$ como $E[U_+Y] - E[U_-Y]$ (donde U_+ y U_- denotan las partes positivas y negativas de U.). Por ello el lado derecho de la igualdad del teorema está siempre bien definido.

- (**Formulación alternativa de la Fórmula de la dualidad**). La fórmula de la dualidad se puede reescribir como

$$\mathrm{Ent(Y)} = \sup_{T} E[Y(\log T - \log(ET))] \tag{3.2}$$

donde el supremo se toma sobre todas las variables aleatorias integrables y no negativas.

Demostración. Para probar la fórmula de dualidad de entropía obsérvese que para toda variable aleatoria U con $Ee^u = 1$, se tiene,

$$\mathrm{Ent}[Y] - \mathrm{E}[UY] = \mathrm{Ent}_{e^u P}[Ye^{-u}]$$

pues es inmediato al aplicar la definición (3.1). Nótese que $\mathrm{Ent}_{e^u P}$ se define como la entropía en la que las esperanzas se toman con respecto a la medida de probabilidad $e^U P$. Esto nos muestra que

$$\mathrm{Ent}[Y] - \mathrm{E}[UY] \leq 0$$

y se de la igualdad cuando $e^U = Y/\mathrm{E}[Y]$. Y así queda probada la fórmula de la dualidad. Para ver la última parte del teorema considérese U tal que $\mathrm{E}[UY] \leq \mathrm{Ent}[Y]$ para toda variable aleatoria no negativa Y en $(\Omega, \mathcal{A})$ tal que $\Phi(Y)$ es integrable. Si $Ee^U = 0$ no hay nada que probar. De otra forma, dado un entero positivo n suficientemente grande, se puede elegir $Y = e^{U \wedge n}/x_n$ con $x_n = \mathrm{E}[e^{U \wedge n}]$ lo que nos lleva a que

$$\mathrm{E}[UY_n] \leq \mathrm{Ent}[Y_n] = \mathrm{E}\Phi(Y_n) - \Phi(\mathrm{E}Y_n)$$

y además,

$$\frac{1}{x_n} E[Ue^{\mathrm{mín}(U,n)}] \leq \frac{1}{x_n}[E[(\mathrm{mín}(U,n))e^{\mathrm{mín}(U,n)}] - \log x_n].$$

Por tanto, $\log x_n = \log Ee^{\mathrm{mín}(U,n)} \leq 0$ y tomando el límite cuando $n \to \infty$, se tiene por el Teorema de la convergencia monótona que $Ee^U \leq 1$, lo que concluye la prueba del teorema.

□

El teorema previo hace posible establecer una dualidad entre entropía y las funciones generadoras de momentos.

3.3 Corolario. *Sea Z una variable aleatoria integrable con valores reales. Entonces para todo $\lambda \in \mathbb{R}$,*

$$\log Ee^{\lambda(E-EZ)} = \sup_{\boldsymbol{Q}<<\boldsymbol{P}} [\lambda(E_{\boldsymbol{Q}}Z - EZ) - D(\boldsymbol{Q}||\boldsymbol{P})]$$

donde el supremo se toma de todas las medidas de probabilidad $\boldsymbol{Q}$ absolutamente continuas con respecto a $\boldsymbol{P}$, y $E_{\boldsymbol{Q}}$ denota la integración con respecto a la medida $\boldsymbol{Q}$.(E es la integración con respecto a $\boldsymbol{P}$).

Demostración. Sea $\mathbf{Q}$ una medida de probabilidad absolutamente continua con respecto a $\mathbf{P}$. Tomando $Y = d\mathbf{Q}/d\mathbf{P}$ y eligiendo $U = \lambda(Z - EZ) - \psi_{Z-EZ}(\lambda)$, y como $Ee^U = 1$ se sigue de la fórmula del Teorema (3.1) que

$$D(\mathbf{Q}||\mathbf{P}) = \mathrm{Ent(Y)} \geq E[UY] = \lambda(E_{\mathbf{Q}}Z - EZ) - \psi_{Z-EZ}(\lambda),$$

o equivalentemente que

$$\psi_{Z-EZ}(\lambda) \geq \lambda(E_{\mathbf{Q}}Z - EZ) - D(\mathbf{Q}||\mathbf{P})$$

y además

$$\psi_{Z-EZ}(\lambda) = \log Ee^{\lambda(Z-EZ)} \geq \sup_{\mathbf{Q}'<<\mathbf{P}} [\lambda(E_{\mathbf{Q}'}Z - EZ) - D(\mathbf{Q}'||\mathbf{P}].$$

Recíprocamente, tomando

$$U = \lambda(Z - EZ) - \sup_{\mathbf{Q}'<<\mathbf{P}} [\lambda(E_{\mathbf{Q}'}Z - EZ) - D(\mathbf{Q}'||\mathbf{P}]$$

para toda variabales aleatoria no negativa Y tal que $EY = 1$,

$$E[UY] \leq \mathrm{Ent(Y)}.$$

Por consiguiente, por el Teorema (3.1), $Ee^U \leq 1$ lo que quiere decir que

$$\psi_{Z-EZ}(\lambda) = \log Ee^{\lambda(Z-EZ)} \leq \sup_{\mathbf{Q}'<<\mathbf{P}} [\lambda(E_{\mathbf{Q}'}Z - EZ) - D(\mathbf{Q}'||\mathbf{P}].$$

□

La fórmula de la dualidad implica la propiedad siguiente de la divergencia de *Kullback-Leibler*.

3.4 Corolario. *Sean* $\boldsymbol{P}$ *y* $\boldsymbol{Q}$ *dos distribuciones de probabilidad en el mismo espacio. Entonces*

$$D(\boldsymbol{Q}||\boldsymbol{P}) = \sup_Z [E_{\boldsymbol{Q}} Z - \log E e^Z]$$

donde el supremo se toma entre todas las variables aleatorias tales que $Ee^Z < \infty$.

La fórmula de la dualidad para la entropía y sus corolarios tienen muchas útiles consecuencias. Entre ellas se encuentran las tres siguientes.

3.5 Teorema (**Convexidad de la divergencia de Kullback-Leibler**). *Para cualquier medida* $\boldsymbol{P}$ *en* $\mathcal{X}$*, la función* $\boldsymbol{Q} \longrightarrow D(\boldsymbol{P}||\boldsymbol{Q})$ *es convexa en el conjunto de distribuciones de probabilidad sobre* $\mathcal{X}$.

Demostración. Sea Z una variable aleatoria simple, $Z = \sum_{i \in I} a_i I_{A_i}$ se tiene $E_{\mathbf{Q}}(Z) = \sum_{i \in I} a_i \mathbf{Q}(A_i)$. Ahora bien, si $\mathbf{Q_1}$ y $\mathbf{Q_2}$ son dos medidas de probabilidad

$$\begin{aligned} E_{\lambda \mathbf{Q_1} + (1-\lambda)\mathbf{Q_2}}(Z) &= \sum_{i \in I} a_i (\lambda \mathbf{Q_1}(A_i) + (1-\lambda)\mathbf{Q_2}(A_i)) \\ &= \lambda \sum_{i \in I} a_i \mathbf{Q_1}(A_i) + (1-\lambda) \sum_{i \in I} a_i \mathbf{Q_2}(A_i)) \\ &= \lambda E_{\mathbf{Q_1}}(Z) + (1-\lambda) E_{\mathbf{Q_2}}(Z). \end{aligned}$$

Recuérdese que E es la integración con respecto a $\mathbf{P}$.

$$\begin{aligned} D(\lambda \mathbf{Q_1} + (1-\lambda)\mathbf{Q_2}||\mathbf{P}) &= \sup_Z [E_{\lambda \mathbf{Q_1} + (1-\lambda)\mathbf{Q_2}}(Z) - \log E e^Z] \\ &= \sup_Z [\lambda (E_{\mathbf{Q_1}}(Z) - \log E e^Z) + (1-\lambda)(E_{\mathbf{Q_2}}(Z) - \log E e^Z)] \\ &\leq \sup_Z [\lambda (E_{\mathbf{Q_1}}(Z) - \log E e^Z)] + \sup_Z [(1-\lambda)(E_{\mathbf{Q_2}}(Z) - \log E e^Z)] \\ &= \lambda D(\mathbf{Q_1}||\mathbf{P}) + (1-\lambda) D(\mathbf{Q_2}||\mathbf{P}) \end{aligned}$$

□

3.6 Teorema (**Divergencia de Kullback-Leibler y Transformación de Legendre de la función generadora de momentos logarítmicos**). *Sea* Z *una variable aleatoria con valores reales. Recuérdese que* $\psi_Z(\lambda) = \log E e^{\lambda Z}$ *para todo* $\lambda \in \mathbb{R}$. *Sea* $\psi*(t) = \sup_{\lambda \in \mathbb{R}} [\lambda t - \psi_{Z-EZ}(\lambda)]$. *Entonces se tiene para todo* $t > 0$ *que*

$$\psi*(t) = \text{ínf}\{D(\boldsymbol{Q}||\boldsymbol{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\}$$

Demostración. En primer lugar cabe señalar que como se vió en (1.3) $\psi *_Z (t) = \sup_{\lambda \geq 0}[\lambda t - \psi_Z(\lambda)]$ se puede extender a $\sup_{\lambda \in \mathbb{R}}[\lambda t - \psi_{Z-EZ}(\lambda)]$ por lo que la desigualdad formulada equivale a ver que

$$\sup_{\lambda \geq 0}[\lambda t - \psi_Z(\lambda)] = \text{ínf}\{D(\mathbf{Q}||\mathbf{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\}.$$

Sea $\mathbf{Q}$ una medida de probabilidad absolutamente continua con respecto a $\mathbf{P}$. Tomando $Y = d\mathbf{Q}/d\mathbf{P}$ y eligiendo $U = \lambda(Z - EZ) - \psi_{Z-EZ}(\lambda)$, se tiene $Ee^U = 1$ y se sigue de la fórmula del Teorema (3.1)

$$D(\mathbf{Q}||\mathbf{P}) = \text{Ent(Y)} \geq E[UY] = \lambda(E_{\mathbf{Q}}Z - EZ) - \psi_{Z-EZ}(\lambda),$$

Si $D(\mathbf{Q}||\mathbf{P})$ es tal que cumple $E_{\mathbf{Q}}(Z) - EZ \geq t$

$$\{D(\mathbf{Q}||\mathbf{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\} \geq \lambda t - \psi_{Z-EZ}(\lambda) \quad \forall t, \forall \lambda \geq 0.$$

En particular,

$$\text{ínf}\{D(\mathbf{Q}||\mathbf{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\} \geq \sup_{\lambda \geq 0}[\lambda t - \psi_Z(\lambda)]$$

o lo que es lo mismo,

$$\text{ínf}\{D(\mathbf{Q}||\mathbf{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\} \geq \sup_{\lambda \in \mathbb{R}}[\lambda t - \psi_Z(\lambda)].$$

Recíprocamente,
Si $E_{\mathbf{Q}}(Z) - EZ \geq t$, por el corolario (3.3) se tiene
$\sup_{\lambda \geq 0}[\lambda t - D(\mathbf{Q}||\mathbf{P})] \leq \psi_{Z-EZ}(\lambda) \Leftrightarrow \sup_{\lambda \geq 0}[\lambda t - \psi_{Z-EZ}(\lambda)] \leq \text{ínf}\{D(\mathbf{Q}||\mathbf{P}) : E_{\mathbf{Q}}(Z) - EZ \geq t\}$

□

3.7 Teorema (La divergencia de Kullback-Leibler respecto a la distribución producto.). *Sea $\boldsymbol{P}$ la probabilidad definida en $\mathcal{X} \times \mathcal{Y}$. Sea $\boldsymbol{P_x}$ y $\boldsymbol{P_y}$ sus distribuciones marginales y sean $\boldsymbol{Q_x}$ y $\boldsymbol{Q_y}$ distribuciones de probabilidad definidas en $\mathcal{X}$ e $\mathcal{Y}$. Se tiene que*

$$D(\boldsymbol{P}||\boldsymbol{Q_x} \otimes \boldsymbol{Q_y}) = D(\boldsymbol{P}||\boldsymbol{P_x} \otimes \boldsymbol{P_y}) + D(\boldsymbol{P_x}||\boldsymbol{Q_x}) + D(\boldsymbol{P_y}||\boldsymbol{Q_y})$$

Demostración. Si $\mathbf{Q} << \mathbf{P}$,

$$\mathrm{Ent}\Big(\frac{d\mathbf{Q}}{d\mathbf{P}}\Big) = D(\mathbf{Q}||\mathbf{P}) \quad \text{con} \quad Y = \frac{d\mathbf{Q}}{d\mathbf{P}}.$$

Ahora bien,

$$D(\mathbf{Q}||\mathbf{P}) = \mathrm{Ent}(Y) = \int_\Omega Y \log Y d\mathbf{P} = \int \frac{d\mathbf{Q}}{d\mathbf{P}} \log\Big(\frac{d\mathbf{Q}}{d\mathbf{P}}\Big) d\mathbf{P}.$$

Aplicado a este problema,

$$D(\mathbf{P}||\boldsymbol{Q_x} \otimes \boldsymbol{Q_y}) = \int \frac{d\mathbf{P}}{d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y})} \log\Big(\frac{d\mathbf{P}}{d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y})}\Big) d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y}).$$

Haciendo uso de la descomposición

$$\frac{d\mathbf{P}}{d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y})} = \frac{d\mathbf{P}}{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})} \cdot \frac{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})}{d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y})}$$

se tiene,

$$D(\mathbf{P}||\boldsymbol{Q_x} \otimes \boldsymbol{Q_y}) = \int \log\Big(\frac{d\mathbf{P}}{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})}\Big) d\mathbf{P} + \int \log\Big(\frac{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})}{d(\boldsymbol{Q_x} \otimes \boldsymbol{Q_y})}\Big) d\mathbf{P}.$$

Recuérdese ahora que si $\tilde{\mathbf{P}} = \boldsymbol{P_x} \otimes \boldsymbol{P_y}$ y $\tilde{\mathbf{Q}} = \boldsymbol{Q_x} \otimes \boldsymbol{Q_y}$

$$\frac{d\tilde{\mathbf{P}}}{d\tilde{\mathbf{Q}}}(x,y) = \frac{d\boldsymbol{P_x}}{d\boldsymbol{Q_x}}(x) \cdot \frac{d\boldsymbol{P_y}}{d\boldsymbol{Q_y}}(y).$$

De modo que,

$$\begin{aligned} D(\mathbf{P}||\boldsymbol{Q_x} \otimes \boldsymbol{Q_y}) &= \int \log\Big(\frac{d\mathbf{P}}{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})}\Big) d\mathbf{P} + \int \Big[\log \frac{d\boldsymbol{P_x}}{d\boldsymbol{Q_x}}(x) \log \frac{d\boldsymbol{P_y}}{d\boldsymbol{Q_y}}(y)\Big] d\mathbf{P} \\ &= \int \log\Big(\frac{d\mathbf{P}}{d(\boldsymbol{P_x} \otimes \boldsymbol{P_y})}\Big) d\mathbf{P} + \int \log \frac{d\boldsymbol{P_x}}{d\boldsymbol{Q_x}}(x) d\mathbf{P} + \int \log \frac{d\boldsymbol{P_y}}{d\boldsymbol{Q_y}}(y) d\mathbf{P} \\ &= D(\mathbf{P}||\boldsymbol{P_x} \otimes \boldsymbol{P_y}) + D(\boldsymbol{P_x}||\boldsymbol{Q_x}) + D(\boldsymbol{P_y}||\boldsymbol{Q_y}). \end{aligned}$$

□

La divergencia de *Kullback-Leibler* se puede derivar de la divergencia de *Bregman* como se puede ver a continuación.

3.8 Teorema (La esperanza minimiza la divergencia de Bregman). *Sean $I \subset \mathbb{R}$ un intervalo abierto y sea $f : I \longrightarrow \mathbb{R}$ una función convexa y diferenciable. Para todo $x, y \in I$, la divergencia de Bregman de f para x a y es $f(y) - f(x) - f'(x)(y - x)$. Sea X una variable aleatoria con valores en I. Entonces*

$$E[f(X) - f(EX)] = \inf_{u>0} E[f(X) - f(a) - f'(a)(X - a)].$$

Tomando $f(x) = x \log x$, se obtiene la siguiente fórmula variacional para la entropía.

3.9 Corolario. *Sea Y una variable aleatoria no negativa tal que $E\Phi(Y) < \infty$. Entonces*

$$\mathrm{Ent(Y)} = \inf_{u>0} E[Y(\log Y - \log u) - (Y - u)].$$

3.2. Lema de transporte

La fórmula de la dualidad del Corolario (3.3) permite relacionar la propiedad de concentración de una variable aleatoria Z alrededor de su esperanza con lo que se llama coste de transporte, esto es, el "precio" que uno tiene que pagar cuando se calcula la expectativa de Z bajo una medida de probabilidad $\mathbf{Q}$ en lugar de la medida de probabilidad original $\mathbf{P}$.

3.10 Lema. *Sea Z una variable aleatoria integrable con valores reales. Sea Φ una función convexa y continuamente diferenciable en un intervalo $[0, b)$ (puede que no acotado) y se asume que $\Phi(0) = \Phi'(0) = 0$. Se define para todo $x \geq 0$,*

$$\Phi^*(x) = \sup_{\lambda \in (0,b)} (\lambda x - \Phi(\lambda)),$$

y sea para todo $t \geq 0$

$$\Phi^{*-1}(t) = \inf\{x \geq 0 : \Phi^*(x) > t\}.$$

Entonces los dos enunciados siguientes son equivalentes:

(i) $\forall \lambda \in (0, b), \quad \log Ee^{\lambda(Z - EZ)} \leq \Phi(\lambda)$

(ii) Para cualquier medida $\boldsymbol{Q}$ *absolutamente continua con respecto a* $\boldsymbol{P}$ *tal que* $D(\boldsymbol{Q}||\boldsymbol{P}) < \infty$,

$$E_{\boldsymbol{Q}}Z - EZ \leq \Phi^{*-1}[D(\boldsymbol{Q}||\boldsymbol{P})]. \tag{3.3}$$

En particular, dado $\nu > 0$

$$\log Ee^{\lambda(Z-EZ)} \leq \frac{\nu\lambda^2}{2} \quad \forall\lambda > 0$$

si, y sólo si, para cualquier medida de probabilidad $\boldsymbol{Q}$ *absolutamente continua respecto a* $\boldsymbol{P}$ *y tal que* $D(\boldsymbol{Q}||\boldsymbol{P}) < \infty$,

$$E_Q Z - EZ \leq \sqrt{2\nu D(\boldsymbol{Q}||\boldsymbol{P})}.$$

Demostración. Haciendo uso del Corolario (3.3) se tiene que

$$\sup_{\mathbf{Q}<<\mathbf{P}} [\lambda(E_{\mathbf{Q}} - EZ) - D(\mathbf{Q}||\mathbf{P})] \leq \Phi(\lambda),$$

lo que equivale a

$$E_{\mathbf{Q}} - EZ \leq \inf_{\lambda\in(0,b)} \left(\frac{\Phi(\lambda) + D(\mathbf{Q}||\mathbf{P})}{\lambda} \right)$$

para toda distribución $\mathbf{Q}$ que es absolutamente continua con respecto a $\mathbf{P}$. Sin embargo, se sigue del Lema (1.5) que

$$\Phi^{*-1}(D(\mathbf{Q}||\mathbf{P})) = \inf_{\lambda\in(0,b)} \left(\frac{\Phi(\lambda) + D(\mathbf{Q}||\mathbf{P})}{\lambda} \right)$$

lo que lleva a que (i) es equivalente a (ii).

□

3.3. Desigualdad de Pinsker

En esta sección se verá un resultado fundamental conocido como desigualdad de *Pinsker* que es la base del éxitoso método para probar desigualdades de concentración llamado "método de transporte". La desigualdad de *Pinsker* relaciona la entropía de dos distribuciones de probabilidad en un espacio medible $(\Omega,\mathcal{A})$. En primer lugar véase la definición de variación total.

3.11 Definición. La variación total o la distancia variacional entre $\mathbf{P}$ Y $\mathbf{Q}$ está definida por

$$V(\mathbf{P}, \mathbf{Q}) = \sup_{A \in \mathcal{A}} |\mathbf{P}(A) - \mathbf{Q}(A)|$$

pudiéndose representar también por $d_{TV}(\mathbf{P}, \mathbf{Q})$.

Además la variación total se puede definir como la mitad de la distancia $\mathbb{L}_1(||f||_1 = \left(\int_X |f|^1 dm\right)$), esto es, si λ es la medida dominante común de $\mathbf{P}$ y $\mathbf{Q}$ con $p(x) = d\mathbf{P}/d\lambda$ y $q(x) = d\mathbf{Q}/d\lambda$ sus densidades respectivas, entonces

$$V(\mathbf{P}, \mathbf{Q}) = \mathbf{P}(A^*) - \mathbf{Q}(A^*) = \frac{1}{2} \int |p(x) - q(x)| d\lambda(x)$$

donde $A^* = \{x : p(x) \geq q(x)\}$. Además se puede probar que $V(\mathbf{P}, \mathbf{Q}) = \text{mín}\, \mathbf{P}\{X \neq Y\}$, donde el mínimo se toma sobre todos los pares de distribuciones comunes para las variables aleatorias (X, Y) cuyas distribuciones marginales son $X \sim \mathbf{P}$ y $Y \sim \mathbf{Q}$.

De forma similar a como ocurre en el caso continuo se tiene,

3.12 Teorema. *Sean $\boldsymbol{P}$ y $\boldsymbol{Q}$ distribuciones de probabilidad en el mismo conjunto discreto $\mathcal{X}$. Entonces,*

$$V(\boldsymbol{P}, \boldsymbol{Q}) = \boldsymbol{P}(A^*) - \boldsymbol{Q}(A^*) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\boldsymbol{P}(x) - \boldsymbol{Q}(x)|$$

donde $A^* = \{x : \boldsymbol{P}(x) \geq \boldsymbol{Q}(x)\}$. Además, $V(\boldsymbol{P}, \boldsymbol{Q}) = \text{mín}\, \boldsymbol{P}\{X \neq Y\}$, donde el mínimo se toma sobre todos los pares de distribuciones comunes para las variables aleatorias (X, Y) cuyas distribuciones marginales son $X \sim \boldsymbol{P}$ y $Y \sim \boldsymbol{Q}$.

Demostración. Sean $A^* = \{x : \mathbf{P}(x) \geq \mathbf{Q}(x)\}$ y $B^* = \{x : \mathbf{P}(x) \leq \mathbf{Q}(x)\}$. Puesto que

$$\sum_{x \in A^*} (\mathbf{P}(x) - \mathbf{Q}(x)) = -\sum_{x \in B^*} (\mathbf{P}(x) - \mathbf{Q}(x)) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{P}(x) - \mathbf{Q}(x)|.$$

En particular,

$$\frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{P}(x) - \mathbf{Q}(x)| \leq V(\mathbf{P}, \mathbf{Q}).$$

Por otro lado, si $C^* \subset \mathcal{X}$

$$\mathbf{P}(C^*) - \mathbf{Q}(C^*) = (\mathbf{P}(C^* \cap A^*) - \mathbf{Q}(C^* \cap A^*)) + (\mathbf{P}(C^* \cap B^*) - \mathbf{Q}(C^* \cap B^*))$$

y de aquí se deduce

$$-\mathbf{P}(A^*) - \mathbf{Q}(A^*) \leq \mathbf{P}(C^*) - \mathbf{Q}(C^*) \leq \mathbf{P}(A^*) - \mathbf{Q}(A^*).$$

□

Como aplicación de esta caracterización de la distancia de variación total, probamos a continuación la siguiente versión de la ley de los sucesos raros.

La Ley de los Eventos Raros fue demostrada por *Poisson* en su libro *Recherches sur la Probabilités des Jugements en Matière Criminelle et Matière Civile*. Surge del estudio de una variable aleatoria X que mide el "número de éxitos" en n ensayos *Bernoulli* con parámetro p. El término raro se refiere a que la probabilidad de éxito $p > 0$ es pequeña.

3.13 Teorema (**Ley de eventos raros**). *Sea $\boldsymbol{P}$ la distribución de probabilidad de una suma de n variables aleatorias independientes $X_1, ..., X_n$ con distribución Bernoulli con parámetros $p_1, ..., p_n$. Sea $P_o(\mu)$ la distribución Poisson con esperanza $\mu = \sum_{i=1}^{n} p_i$.*
Entonces $V(\boldsymbol{P}, P_o(\mu)) \leq \sum_{i=1}^{n} p_i^2$.

Demostración. Se utilizará la letra $\mathbb{P}$ para denotar la medida de probabilidad y $\mathbf{P}$ y $\mathbf{Q}$ serán distribuciones de probabilidad. Recuérdese que si $\mathbf{P}$ y $\mathbf{Q}$ son distribuciones de probabilidad en el mismo conjunto discreto $\mathcal{X}$ entonces

$$V(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{P}(x) - \mathbf{Q}(x)|.$$

Teniéndose además que

$$V(\mathbf{P}, \mathbf{Q}) = \text{mín}\, \mathbb{P}\{X \neq Y\},$$

donde el mínimo se toma sobre todos los pares de distribuciones comunes para las variables aleatorias (X, Y) cuyas distribuciones marginales son $X \sim \mathbf{P}$ y $Y \sim \mathbf{Q}$. Ahora bien, para $\mathbf{P} = B(p)$ y $\mathbf{Q} = P_o(p)$ se tiene,

$$\begin{aligned}
V(B(p), P_o(p)) &= \frac{1}{2} \sum_{j=0}^{\infty} |\mathbf{P}(j) - \mathbf{Q}(j)| \\
&= \frac{1}{2} \left[\sum_{j=2}^{\infty} \frac{e^{-p} p^j}{j!} + |(1-p) - e^{-p}| + |p - e^{-p} p| \right] \\
&= \frac{1}{2} \left[e^{-p}(e^p - 1 - p) + e^{-p} - (1-p) + p(1 - e^{-p}) \right] \\
&= \frac{1}{2} \left(2p - 2pe^{-p} \right) \\
&= p(1 - e^{-p}) \leq p^2.
\end{aligned}$$

pues recuérdese que $1-p \leq e^{-p}$. Ahora bien como

$$V(B(p), P_o(p) = p(1-e^{-p}) \leq p^2$$

en particular existen (X_i, Y_i) con $X_i \sim B(p_i)$, $Y_i \sim P_o(p_i)$ y $\mathbb{P}\{X_i \neq Y_i\} \leq p_i^2$. Sean $(X_1, Y_1), ..., (X_n, Y_n)$ variables aleatorias independientes tales que $X_1 + ... + X_n \sim \mathbf{P}$ y $Y_1 + ... + Y_n \sim P_o(\mu)$ se tiene

$$\mathbb{P}\{X_1 + ... + X_n \neq Y_1 + ... + Y_n\} \leq \sum_{i=1}^{n} \mathbb{P}\{X_i \neq Y_i\} \leq \sum_{i=1}^{n} p_i^2.$$

De modo que $V(\mathbf{P}, P_o(\mu)) \leq \sum_{i=1}^{n} p_i^2$

□

Completamos la sección con el resultado anunciado que relaciona la distancia en variación total con la entropía relativa.

3.14 Teorema (**Desigualdad de Pinsker**). *Sean* $\boldsymbol{P}$ *y* $\boldsymbol{Q}$ *distribuciones de probabilidad en* $(\Omega, \mathcal{A})$ *tal que* $\boldsymbol{Q} << \boldsymbol{P}$. *Entonces,*

$$V(\boldsymbol{P}, \boldsymbol{Q})^2 \leq \frac{1}{2} D(\boldsymbol{Q}||\boldsymbol{P})$$

Demostración. Sea la variable aleatoria Y tal que $\mathbf{Q} = Y\mathbf{P}$ y sea $A^* = \{Y \geq 1\}$ el conjunto que alcanza el máximo en la definición de variación total entre $\mathbf{P}$ y $\mathbf{Q}$. Entonces siendo $Z = \mathbb{1}_{\{A^*\}}$,

$$V(\mathbf{P}, \mathbf{Q}) = \mathbf{Q}(A^*) - \mathbf{P}(A^*) = E_{\mathbf{Q}} Z - EZ$$

se sigue del lema de *Hoeffding* (1.9) que para todo $\lambda > 0$

$$\Psi_{Z-EZ}(\lambda) = \log E e^{\lambda(Z-EZ)} \leq \frac{\lambda^2}{8}$$

que por (3.3), lleva a,

$$E_{\mathbf{Q}} Z - EZ \leq \sqrt{2 \, \frac{1}{4} \, D(\mathbf{Q}||\mathbf{P})}$$

□

3.4. Sub-Aditividad de la Entropía

A continuación se verá una desigualdad que servirá como la base del llamado "método de la entropía" para probar desigualdades de concentración. Además se verá una versión mucho más general con consecuencias más importantes.

Sean $X_1, ..., X_n$ variables aleatorias independientes y $Z = f(X_1, ..., X_n)$. La base del método de la entropía es una extensión útil de la desigualdad de *Efron-Stein*. Recuérdese que la desigualdad de *Efron-Stein* establece que

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

Es claro que esta ecuación equivale a tomar $\Phi(x) = x^2$ en

$$E\Phi(Z) - \Phi(EZ) \leq \sum_{i=1}^{n} E\left[E^{(i)}\Phi(Z) - \Phi(E^{(i)}Z)^2\right].$$

De hecho, esta desigualdad es cierta para una clase más amplia de funciones convexas Φ. Estudiamos ahora el caso $\Phi(x) = x\log x$.

3.15 Teorema (Sub-aditividad de la entropía). *Sean $X_1, ..., X_n$ variables aleatorias independientes y sea $Y = f(X_1, ..., X_n)$ una función medible no negativa de estas variables tal que $\Phi(Y) = Y\log Y$ es integrable. Para todo $1 \leq i \leq n$, se denota por $E^{(i)}$ a la esperanza condicionada a $X^{(i)} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$ y $\mathrm{Ent}^{(i)}(Y)$ la entropía condicional de Y, $\mathrm{Ent}^{(i)}(Y) = \mathrm{Ent}^{(i)}\Phi(Y) - \Phi(\mathrm{Ent}^{(i)}Y)$. Entonces,*

$$\mathrm{Ent(Y)} \leq E\sum_{i=1}^{n} \mathrm{Ent}^{(i)}(Y).$$

Demostración. Sea $E_i[\cdot] = E[\cdot|X_1, ..., X_i]$ la varianza condicional para $i = 1, ..., n$ y se asume que $E_0 = E$. Se tiene la descomposición:

$$Y(\log Y - \log(EY)) = \sum_{i=1}^{n} Y(\log(E_iY) - \log(E_{i-1}Y)).$$

Haciendo ahora uso de (3.2) y de que $E_i[E^{(i)Z}] = E_{i-1}Z$ pues $X_1, ..., X_n$ son independientes se tiene

$$E^{(i)}[Y(\log(E_iY) - \log(E^{(i)}[E_iY]))] \leq \mathrm{Ent}^{(i)}(Y).$$

Uniendo ambos resultados y tomando esperanzas a ambos lados

$$\begin{aligned} E[Y(\log Y - \log(EY))] &= \sum_{i=1}^{n} E\left[E^{(i)}[Y(\log(E_i Y) - \log(E^{(i)}[E_i Y]))]\right] \\ &= \sum_{i=1}^{n} E\left[Y(\log(E_i Y) - \log(E_{i-1} Y]\right] \\ &\leq \sum_{i=1}^{n} E\left[\mathrm{Ent}^{(i)}(Y)\right]. \end{aligned}$$

□

Esta propiedad de sub-aditividad es válida para otros tipos de Φ- entropía, concepto que introducimos a continuación

3.16 Definición. Sea $\Phi : [0, \infty) \longrightarrow \mathbb{R}$ una función convexa y se asigna a toda variable aleatoria no negativa Z e integrable el número

$$H_\Phi(Z) = E\Phi(Z) - \Phi(EZ).$$

Por la desigualdad de *Jensen*, $H_\Phi(Z)$ es siempre no negativa. A $H_\Phi(Z)$ se le llama Φ-entropía de Z.

Recuérdese que $Z = f(X_1, ..., X_n)$ será la función a estudiar siendo ésta medible no negativa donde $X_1, ..., X_n$ son variables aleatorias independientes que toman valores en un conjunto $\mathcal{X}$. La propiedad clave que se necesita es la siguiente desigualdad se sub-aditividad de Φ-entropía.

$$H_\Phi(Z) \leq E \sum_{i=1}^{n} H_\Phi^{(i)}(Z)$$

donde $H_\Phi^{(i)}(Z) = E^{(i)}\Phi(Z) - \Phi(E^{(i)}Z)$ es la entropía condicional y como ya s eha mencionado anteriormente $E^{(i)}$ denota la esperanza condicionada a $X^{(i)}$. Obsérvese que para $\Phi(x) = x^2$, la propiedad de sub-aditividad es justo la desigualdad de *Efron-Stein*, mientras que con $\Phi(x) = x \log x$ se reduce a la desigualdad de sub-aditividad de la entropía ordinaria.

Aquí se verá que las Φ-entropías son sub-aditivas para una amplia clase de funciones convexas Φ. De hecho, se caracteriza la clase de funciones Φ que dan lugar a los funcionales de entropía con la propiedad de sub-aditividad.

En primer lugar cabe señalar que la sub-aditividad es equivalente a una desigualdad de tipo *Jensen*. En realidad para n=2 y tomando $Z = f(X_1, X_2)$, la propiedad de sub-aditividad se reduce a

$$H_\Phi\left(\int f(x, X_2)d\mu_1(x)\right) \leq H_\Phi(f(x, X_2)) \tag{3.4}$$

donde μ_1 denota la distribución de X_1. Por otra parte, (3.4) implica la propiedad de aditividad. En efecto, sea Y_1 igualmente distribuida a X_1 y sea Y_2 igualmente distribuida a la n-1-tupla $X_2, ..., X_n$. Sean μ_1 y μ_2 sus distribuciones correspondientes. Entonces, $Z = f(Y_1, Y_2)$ es una función medible. Por el teorema de *Tonelli-Fubini*,

$$\begin{aligned}
H_\Phi(Z) = \int\int \Big(\Phi\Big(f(y_1, y_2)\Big) &- \Phi\Big(\int f(y_1', y_2)d\mu_1(y_1')\Big) \\
&+ \Phi\Big(\int f(y_1', y_2)d\mu_1(y_1')\Big) \\
&- \Phi\Big(\int\int f(y_1', y_2')d\mu_1(y_1')d\mu_2(y_2')\Big)\Big)d\mu_1(y_1)d\mu_2(y_2) \\
\\
= \int \Big(\int \Big[\Phi\Big(f(y_1, y_2)\Big) &- \Phi\Big(\int f(y_1', y_2)d\mu_1(y_1')\Big]\, d\mu_1(y_1)\Big)d\mu_2(y_2) \\
&+ \int \Big(\Phi\Big(\int f(y_1', y_2)d\mu_1(y_1')\Big) \\
&- \Phi\Big(\int\int f(y_1', y_2')d\mu_1(y_1')d\mu_2(y_2')\Big)\Big)d\mu_2(y_2) \\
\\
= \int H_\Phi\Big(f(Y_1, y_2)\Big)d\mu_2(y_2) &+ H_\Phi\Big(\int f(y_1', Y_2)d\mu_1(y_1')\Big) \\
\leq \int H_\Phi\Big(f(Y_1, y_2)\Big)d\mu_2(y_2) &+ \int H_\Phi(f(y_1', Y_2))d\mu_1(y_1')
\end{aligned}$$

donde el último paso se sigue de (3.4). En otras palabras se tiene

$$H_\Phi(Z) \leq EH_\Phi^{(1)} + \int H_\Phi(f(x_1, X_2, ..., X_n))d\mu_1(x_1).$$

Por inducción, (3.4) lleva a la propiedad de sub-aditividad para todo n. En consecuencia, la propiedad de sub-aditividad de H_Φ es equivalente a lo que podríamos llamar propiedad de *Jensen*, esto es, (3.4). Esto implica que para probar la sub-aditividad de una Φ-entropía es suficiente ver que ésta cumple la propiedad de *Jensen*.

3.17 Teorema (**Sub-aditividad de la Φ-entropía**). *Sea $\mathcal{C}$ la clase de funciones $f : [0, \infty) \longrightarrow \mathbb{R}$ que son continuas y convexas en $[0, \infty)$, doblemente diferenciables en $(0, \infty)$ y tal que o Φ es una función afín o Φ'' es estrictamente positiva y $1/\Phi''$ es cóncavo. Para todo $\Phi \in \mathcal{C}$, el funcional H_Φ es sub-aditivo.*

La demostración de este teorema se basa principalmente en la fórmula de la dualidad para la Φ-entropía de la forma

$$H_\Phi(Z) = \sup_{T \in \mathcal{T}} E[\psi_1(T)Z + \psi_2(T)]$$

con funciones convenientes ψ_1 y ψ_2 y una clase adecuada de variables no negativas $\mathcal{T}$. Tal fórmula implica obviamente que el funcional H_Φ es convexa. Por otra parte, esto también implica la propiedad de *Jensen* y además la propiedad de sub-aditividad para H_Φ siguiendo el simple argumento: Sea $Z = f(Y_1, Y_2)$ una función donde $Y_1 = X_1$ y $Y_2 = (X_2, ..., X_n)$. Entonces

$$\begin{aligned}
H_\Phi\Big(\int f(y_1, Y_2) d\mu_1(y_1) \Big) &= \sup_{T \in \mathcal{T}} \int \left[\Psi_1(T(y_2)) \int f(y_1, y_2) d\mu_1(y_1) + \Psi_2(T(y_2)) \right] d\mu_2(y_2) \\
&\qquad \text{(Por el Teorema de Fubini)} \\
&= \sup_{T \in \mathcal{T}} \int \left(\int [\Psi_1(T(y_2)) f(y_1, y_2) + \Psi_2(T(y_2))] d\mu_2(y_2) \right) d\mu_1(y_1) \\
&\leq \int \left(\sup_{T \in \mathcal{T}} \int [\Psi_1(T(y_2)) f(y_1, y_2) + \Psi_2(T(y_2))] d\mu_2(y_2) \right) d\mu_1(y_1) \\
&= \int H_\Phi(f(y_1, Y_2)) d\mu_1(y_1).
\end{aligned}$$

Por consiguiente, para completar la prueba del teorema, el siguiente lema es suficiente.

3.18 Lema (**Fórmula de Dualidad para Φ-entropías.**). *Sea $\Phi \in \mathcal{C}$ y $Z \in \mathbb{L}_1^+$.Si $\Phi(Z)$ es integrable entonces*

$$H_\Phi(Z) = \sup_{T \in \mathbb{L}_1^+, T \neq 0} \{E[(\Phi'(T) - \Phi'(ET))(Z - T) + \Phi(T)] - \Phi(ET)\}$$

Demostración. El caso en que Φ es afín es trivial. En este caso H_Φ es cero y lo mismo ocurre con la expresión definida por la fórmula de la dualidad.

Nótese que la expresión entre llaves para $T = Z$ es igual a $H_\Phi(Z)$ por lo que basta comprobar la desigualdad

$$H_\Phi(Z) \geq E[(\Phi'(T) - \Phi'(ET))(Z - T) + \Phi(T)] - \Phi(ET)$$

bajo la suposición que $\Phi(Z)$ es integrable y $T \in \mathbb{L}_1^+$. Se asume que Z y T están acotadas y acotadas no próximas a cero. Para todo $\lambda \in [0, 1]$, se tiene $T_\lambda = (1 - \lambda)Z + \lambda T$ y

$$g(\lambda) = E[(\Phi'(T_\lambda) + \Phi'(ET_\lambda))(Z - T_\lambda)] + H_\Phi(T_\lambda).$$

El objetivo es ver que la función g es no creciente en $[0, 1]$. Nótese que $Z - T_\lambda = \lambda(Z - T)$ y usando los supuestos de acotación para derivar en g se tiene,

$$\begin{aligned} g'(\lambda) &= E[(\Phi'(T_\lambda) - \Phi'(ET_\lambda))(Z - T_\lambda)] \\ &\quad - \lambda E[((-Z + T)\Phi''(T_\lambda) - (-EZ + ET)\Phi''(ET_\lambda))(Z - T_\lambda)] \\ &\quad + E[-Z + T]E\Phi'(T_\lambda) - (-EZ + ET)\Phi'(ET_\lambda) \\ &= E[(\Phi'(T_\lambda) - \Phi'(ET_\lambda))(Z - T_\lambda)] \\ &\quad - \lambda\Big(E[(Z - T)^2\Phi''(T_\lambda)] - (E(Z - T))^2\Phi''(ET_\lambda)\Big) \\ &\quad + E[\Phi'(T_\lambda)(T - Z)] - \Phi'(ET_\lambda)E[T - Z]. \end{aligned}$$

Esto es,

$$g'(\lambda) = -\lambda\Big(E[(Z - T)^2\Phi''(T_\lambda)] - (E(Z - T))^2\Phi''(ET_\lambda)\Big).$$

Ahora, por la desigualdad de *Cauchy-Schwarz* se tiene,

$$\begin{aligned} (E[Z - T])^2 &= \left(E\left[(Z - T)\sqrt{\Phi''(T_\lambda)}\frac{1}{\sqrt{\Phi''(T_\lambda)}}\right]\right)^2 \\ &\leq E\Big[\frac{1}{\Phi''(T_\lambda)}\Big]E[(Z - T)^2\Phi''(T_\lambda)]. \end{aligned}$$

Usando la concavidad de $1/\Phi''$, la desigualdad de *Jensen* implica

$$E\left[\frac{1}{\Phi''(T_\lambda)}\right] \leq \frac{1}{\Phi''(ET_\lambda)},$$

lo que lleva a,

$$(E[Z - T])^2 \leq \frac{1}{\Phi''(ET_\lambda)}E[(Z - T)^2\Phi''(T_\lambda)]$$

que es equivalente a $g'(\lambda) \leq 0$. Y además $g(1) \leq g(0) = H_\Phi(Z)$. Esto quiere decir que para todo T,

$$E[(\Phi'(T) - \Phi'(ET))(Z - T)] + H_\Phi(T) \leq H_\Phi(Z).$$

En el caso general se considera las secuencias

$$Z_n = \text{mín}\left\{ \text{máx}\left\{Z, \frac{1}{n}\right\}, n\right\} \qquad \text{y} \qquad T_k = \text{mín}\left\{ \text{máx}\left\{T, \frac{1}{k}\right\}, k\right\}$$

y el objetivo es tomar el límite en k cuando $n \to \infty$ en la desigualdad

$$H_\Phi(Z_n) = E\Phi(Z_n) - \Phi(EZ_n) \geq E[(\Phi'(T_k) - \Phi'(ET_k))(Z_n - T_k) + \Phi(T_k)] - \Phi(ET_k)$$

que podemos escribirlo también como

$$E[\Psi(Z_n, T_k)] \geq -\Phi'(ET_k)E[Z_n - T_k] - \Phi(ET_k) + \Phi(EZ_n), \tag{3.5}$$

donde $\Psi(z,t) = \Phi(z) - \Phi(t) - (z-t)\Phi'(t)$. Como tenemos que ver que

$$E[\Psi(Z,T)] \geq -\Phi'(ET)E[Z - T] - \Phi(ET) + \Phi(EZ), \tag{3.6}$$

con $\Psi \leq 0$ se puede asumir que $\Psi(Z,T)$ es integrable pues si no lo es (3.6) se verifica trivialmente. A continuación se tomará el límite en (3.5) cuando $n \to \infty$. Primero se estudiará el término a la izquierda. Nótese que $\Psi(z,t)$ como función de t, decrece en $(0,z)$ y crece en (z,∞). De froma similar si $\Psi(z,t)$ es una función de z, $\Psi(z,t)$ decrece en $(0,t)$ y crece en $(t,+\infty)$. Por lo tanto, para todo t, $\Psi(Z_n,t) \leq \Psi(1,t) + \Psi(z,t)$ mientras que para todo z, $\Psi(z,T_k) \leq \Psi(z,1) + \Psi(z,T)$. Por consiguiente, dado k,

$$\Psi(Z_n, T_k) \leq \Psi(1, T_k) + \Psi(Z, T_k),$$

como $\Psi((Z \vee \frac{1}{n}) \wedge n), T_k) \to \Psi(z, T_k)$ para todo z, se puede aplicar el teorema de la convergencia dominada para concluir que $E\Psi(Z_n, T_k)$ converge a $E\Psi(Z, T_k)$ cuando $n \to \infty$. Por tanto, se tiene

$$E\Psi(Z, T_k) \geq -\Phi(ET_k)E[Z - T_k] - \Phi(ET_k) + \Phi(EZ).$$

Ahora también se tiene que $\Psi(Z,T_k) \leq \Psi(Z,1) + \Psi(Z,T)$ y se puede aplicar el teorema de la convergencia dominada para garantizar que $E\Psi(Z,T_k)$ converge a $E\Psi(Z,T)$ cuando $k \to \infty$. Tomando el límite cuando $k \to \infty$ implica que (3.6) se cumple para todos las $T, Z \in \mathbb{L}_1^+$ tales que $\Phi(Z)$ sea integrable y $ET > 0$. si $Z \neq 0$ (3.6) se cumple para $T = Z$,salvo en un conjunto de medida nula, mientras que si $Z = 0$ se cumple para $T = 1$,salvo en un conjunto de medida nula y la prueba del lema está completa para todos los casos generales.

□

3.19 Comentarios.

- Nótese que ya que el supremo en la fórmula (3.5) se alcanza para $T = Z$ (ó $T = 1$ si $Z = 0$). La fórmula de la dualidad continua siendo cierta si el supremo se restringe a la clase $\mathcal{T}_\Phi$ de variables T tales que $\Phi(T)$ es integrable. Por lo que se puede escribir una fórmula alternativa

$$H_\Phi(Z) = \sup_{T \in \mathcal{T}_\Phi} \{E[(\Phi'(T) - \Phi'(ET))(Z - T)] + H_\Phi(T)\}.$$

- El lema (3.5) generaliza la fórmula de la dualidad del teorema (3.6) para la entropía usual. Si se toma $\Phi(x) = x \log x$ se tiene

$$\mathrm{Ent}(Z) = \sup_T \{E[(\log(T) - \log(ET))Z]\}$$

donde el supremo se extiende al conjunto de las variables aleatorias integrables y no negativas T con $ET > 0$. Otro caos de interés se da cuando se toma $\Phi(x) = x^p \quad \text{con} \quad p \in (1, 2]$. En este caso se tiene

$$H_\Phi(Z) = \sup_T \{pE[(Z(T^{p-1} - (ET)^{p-1})] - (p-1)H_\Phi(T)\}$$

donde el supremo se extiende al conjunto de las variables aleatorias no negativas en $\mathbb{L}_p$.

- Por simplicidad nos hemos centrado en variables aleatorias no negativas y funciones convexas Φ en $[0, \infty)$. Esta restricción se puede suprimir y considerar Φ como una función convexa en $\mathbb{R}$ y definir la $\Phi-$entropía de una variable aleatoria Z integrable y con valores reales por la misma fórmula que en el caso no negativo. Asumiendo ahora que Φ es diferenciable en $\mathbb{R}$ y doblemente diferenciable en $\mathbb{R} \setminus \{0\}$, la prueba de la fórmula de dualidad anterior se puede fácilmente adaptar para cubrir el caso en el que $1/\Phi''$ se puede extender a una función cóncava en $\mathbb{R}$. En particular si $\Phi(x) = |x|^p$, donde $p \in (1, 2]$,uno tiene,

$$H_\Phi(Z) = \sup_T \left\{ pE\left[Z\left(\frac{|T|^p}{T} - \frac{|ET|^p}{ET} \right) \right] - (p-1)H_\Phi(T) \right\}$$

donde el supremo se extiende a $\mathbb{L}_p$. Nótese que para p=2, la fórmula se reduce a la fórmula clásica para la varianza

$$\mathrm{Var}(Z) = \sup_T \{2\mathrm{Cov}(Z, T) - \mathrm{Var}(T)\}$$

donde el supremo se extiende al conjunto de las variables de cuadrado integrable. Esto quiere decir que la desigualdad de la sub-aditividad para la $\Phi-$entropía (3.17) también es cierta para funciones Φ convexas en $\mathbb{R}$ con la condición de que $1/\Phi''$ es la restricción a $\mathbb{R} \setminus \{0\}$ de una función cóncava en $\mathbb{R}$.

Capítulo 4

Métodos basados en la entropía

En este capítulo se probarán una desigualdad conocida como desigualdad logarítmica de *Sobolev*. El resultado más simple se verá en la primera sección. Esta desigualdad es sorprendentemente útil, se podrá ver como puede ser usada para probar una desigualdad de concentración exponencial mediante un argumento llamado de *Herbst*. Además se verá como los argumentos del caso de *Bernoulli* se pueden extender a variables aleatorias gaussianas obteniendo así una útil desigualdad de concentración.

4.1. Distribuciones simétricas de Bernoulli

El propósito de esta sección es probar una versión más simple de una familia de desigualdades conocidas como "desigualdades logarítmicas de *Sobolev*".

Se considerarán funciones con valores reales definidas en el hipercubo binario $\{-1,1\}^n$. Sea $X = (X_1, ..., X_n)$ un vector binario distribuido uniformemente en el hipercubo $\{-1,1\}^n$, es decir, $P(X_i = -1) = P(X_i = 1) = 1/2$, con $X_1, ..., X_n$ independientes. Se considera la variable aleatoria $Z = f(X)$ con $f : \{-1,1\}^n \to \mathbb{R}$, la desigualdad logarítmica de *Sobolev* presentada aquí relaciona el funcional de entropía, es decir,

$$\mathrm{Ent}(f) = E[f(X)\log(f(X))] - Ef(X)\log Ef(X),$$

para funciones $f \geq 0$ (se usará indistintamente $\mathrm{Ent}(f)$ y $\mathrm{Ent}(Z)$ para referirse a la entropía de $Z = f(X)$) con una cantidad relacionada con la desigualdad

de *Efron-Stein*,

$$\mathcal{E}(f) := \frac{1}{4}E\Big[\sum_{i=1}^{n}(f(X) - f(\tilde{X}^{(i)}))^2\Big]$$

donde $\tilde{X}^{(i)} = (X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n)$ es la variable que se obtiene al sustituir la i-ésima componente de X por una copia independiente X_i'. Recuérdese que por la desigualdad de *Efron-Stein*, $\mathrm{Var}(f(X)) \leq \mathcal{E}(f)$. Como X está uniformemente distribuida $\mathcal{E}(f)$ puede escribirse de una forma más conveniente,

$$\begin{aligned}\mathcal{E}(f) &= \frac{1}{4}E\Big[\sum_{i=1}^{n}(f(X) - f(\bar{X}^{(i)}))^2\Big] \\ &= \frac{1}{4}E\Big[\sum_{i=1}^{n}(f(X) - f(\bar{X}^{(i)}))_+^2\Big]\end{aligned}$$

donde $\bar{X}^{(i)} = (X_1, ..., X_{i-1}, -X_i, X_{i+1}, ..., X_n)$ es la variable que se obtiene al sustituir la i-ésima componente de X por una copia independiente. Se puede definir $\nabla f(x) = (\nabla_1 f(x), ..., \nabla_n f(x))$ como el vector gradiente discreto donde sus componentes vienen definidas como $\nabla_i f(x) = (f(x) - f(\bar{x^{(i)}}))/2$. Con esta notación, la cota de la varianza por el método de *Efron-Stein* es justo la esperanza de la norma al cuadrado del vector gradiente discreto: $\mathcal{E}(f) = E||\nabla f(x)||^2$.

En este contexto, se conoce al siguiente resultado como desigualdad logarítmica de *Sobolev*.

4.1 Teorema (Desigualdad logarítmica de *Sobolev* para una distribución simétrica *Bernoulli*). *Sea $f : \{-1, 1\}^n \to \mathbb{R}$ una función real definida en el hipercubo binario n-dimensional y se asume que X está distribuida uniformemente sobre $\{-1, 1\}$. Entonces,*

$$\mathrm{Ent}(f^2) \leq 2\mathcal{E}(f).$$

Demostración. Primero se probará para $n = 1$, esto es, cuando $f : \mathbb{R} \to \mathbb{R}$ es una función continuamente diferenciable. La clave de la prueba es la subaditividad de la entropía (Teorema (3.15)), esta propiedad implica que si $Z = f(X)$,

$$\mathrm{Ent}(Z^2) \leq E\sum_{i=1}^{n}\mathrm{Ent}^{(i)}(Z^2) \tag{4.1}$$

donde

$$\mathrm{Ent}^{(i)}(Z^2) = \mathrm{E}^{(i)}[Z^2 \log Z^2] - \mathrm{E}^{(i)}[Z^2]\log(\mathrm{E}^{(i)}[Z^2]).$$

Además es suficiente ver que para todo $i = 1, ..., n$

$$\mathrm{Ent}^{(i)}(Z^2) \leq \frac{1}{2}\mathrm{E}^{(i)}\Big[\Big(f(X) - f(\bar{X}^{(i)})\Big)^2\Big].$$

Dada cualquier realización de $X^{(i)}$ Z puede tomar dos valores diferentes con la misma probabilidad, si a y b son estos valores, $f(X^{(i)})$ vale b si Z vale a y vale a si Z vale b. De modo que la desigualdad deseada (4.1) toma la forma

$$\frac{a^2}{2}\log(a^2)+\frac{b^2}{2}\log(b^2)-\frac{a^2+b^2}{2}\log\frac{a^2+b^2}{2} \leq \frac{1}{2}\Big(\frac{1}{2}(a-b)^2+\frac{1}{2}(a-b)^2\Big) = \frac{1}{2}(a-b)^2$$

y queda demostrar que esta desigualdad es válida para todo $a, b \in \mathbb{R}$. Como $(|a|-|b|)^2 \leq (a-b)^2$, se asume, sin perdida de generalidad, que tanto a como b son no negativas así como se puede asumir que $a \geq b$. Para cualquier valor fijo $b \geq 0$ se define la función

$$h(a) = \frac{a^2}{2}\log(a^2) + \frac{b^2}{2}\log(b^2) - \frac{a^2+b^2}{2}\log\frac{a^2+b^2}{2} - \frac{1}{2}(a-b)^2 \quad \forall a \in [b, \infty).$$

Como $h(b) = 0$, es suficiente ver que $h'(b) = 0$ y que h es cóncava en $[b, \infty)$ de modo que así $h(a) \leq 0$ tal y como queremos demostrar. Si se calcula la derivada de la función se tiene

$$h'(a) = a\log\frac{2a^2}{a^2+b^2} - (a-b)$$

y en concreto $h'(b) = 0$, además

$$h''(a) = 1 + \log\frac{2a^2}{a^2+b^2} - \frac{2a^2}{a^2+b^2} \leq 1 - 1 = 0$$

pues se tiene que $\log(x) - x \leq -1$ y como $h''(a)$ es negativa ó cero, $h(a)$ será cóncava y así $h(a) \leq 0$.

□

La desigualdad logarítmica de *Sobolev* es un resultado más fuerte que una desigualdad de *Poincaré*, como prueban los dos teoremas siguientes,

4.2 Teorema. *Para cualquier variable aleatoria no negativa Z, se tiene* $\mathrm{Var}(Z) \leq \mathrm{Ent}(Z^2)$.

Demostración. Se define el funcional

$$\Psi_p(Z) = \mathrm{E}[z^2] - (\mathrm{E}[z^p])^{2/p} \quad \text{con} \quad p \in [1, 2).$$

Nótese que si $p = 1$, $\Psi_1(Z) = \mathrm{Var}(Z)$. Ahora bien, si se considera el funcional $U(p) = \Psi_p(Z)/((1/p) - (1/2))$, $U(1) = 2\mathrm{Var}(Z)$. Además

$$2\mathrm{Var}(Z) = U(1) \leq U(p).$$

Se verá que el límite de $U(p)$ cuando p tiende a 2 es $2\mathrm{Ent}(Z^2)$ y se concluirá que $2\mathrm{Var}(Z) \leq 2\mathrm{Ent}(Z^2)$. En primer lugar veamos que la función $f(p) = \mathrm{E}[z^p]$ es derivable y se calculará el valor de su derivada. Por el teorema 25.12 de [5] como

$$[z^p]' = \lim_{h\to 0} \frac{z^{p+h} - z^p}{h} = \lim_{h\to 0} z^p \frac{z^h - 1}{h} = z^p \log z$$

basta ver que

$$\Big|\frac{z^h - 1}{h}\Big|$$

está acotada por una función integrable de modo que

$$z^p\Big(\frac{z^h - 1}{h}\Big)$$

será uniformemente integrable y se cumplirá que

$$\mathrm{E}'[z^p] = \lim_{h\to 0} \frac{\mathrm{E}(z^{p+h}) - \mathrm{E}(z^p)}{h} = \lim_{h\to 0} \mathrm{E}\Big(\frac{z^{p+h} - z^p}{h}\Big) = \mathrm{E}[z^p \log z]$$

es decir, que $\mathrm{E}'[z^p] = \mathrm{E}[z^p \log z]$. Para obtener la acotación se redefine la función como

$$z^h - 1 = \exp(h \log z) - \exp(0 \log z) = \int_0^h \exp(s \log z) \log z ds = \log z \int_0^h \exp(s \log z) ds.$$

y se analizan los siguientes casos:

- si $0 \leq z \leq 1$ y $h > 0$ entonces $|z^h - 1| \leq h|\log z|$.
- si $0 \leq z \leq 1$ y $h < 0$ entonces $|z^h - 1| \leq |h||\log z||z|^h$.
- si $z > 1$ y $h > 0$ entonces $|z^h - 1| \leq hz^h \log z$.
- si$z > 1$ y $h < 0$ entonces $|z^h - 1| \leq |h| \log z$.

de modo que para todo z y todo h se tiene,

$$z^p\Big|\frac{z^h-1}{h}\Big| \leq (\log z)z^p + (\log z)z^{p+h}.$$

De este modo se puede concluir que $f(p) = \mathrm{E}[z^p]$ es derivable y que su derivada es $\mathrm{E}'[z^p] = \mathrm{E}[z^p \log z]$. Si llamamos $G(p) = \mathrm{E}[z^p]^{2/p}$ y $g(p) = \log G(p) = 2/p \log \mathrm{E}[z^p]$ entonces

$$g'(p) = \frac{2}{p}\frac{\mathrm{E}[z^p \log z]}{\mathrm{E}[z^p]} - \frac{2}{p^2}\log \mathrm{E}[z^p]$$

y se puede escribir $G(p) = e^{g(p)}$ de modo que

$$\begin{aligned} G'(p) &= g'(p)e^{G(p)} \\ &= \Big(\frac{2}{p}\frac{\mathrm{E}[z^p \log z]}{\mathrm{E}[z^p]} - \frac{2}{p^2}\log \mathrm{E}[z^p]\Big)(\mathrm{E}[z^p])^2/p \\ &= \frac{2}{p^2}\Big(\mathrm{E}[z^p \log z^p] - \mathrm{E}[z^p]\log \mathrm{E}[z^p]\Big)(\mathrm{E}[z^p])^{2/p-1} \end{aligned}$$

Luego,$\Psi_p'(Z) = G'(p)$. De este modo se tiene,

$$\begin{aligned} \lim_{p\to 2^-} U(p) &= \lim_{p\to 2^-} \frac{\Psi_p(Z)}{\frac{1}{p}-\frac{1}{2}} = \lim_{p\to 2^-} \frac{-G'(p)}{-\frac{1}{p^2}} = \frac{\frac{2}{4}(\mathrm{E}[z^2\log z^2] - \mathrm{E}[z^2]\log \mathrm{E}[z^2])}{\frac{1}{4}} \\ &= 2\mathrm{Ent}(Z^2). \end{aligned}$$

El siguiente paso será ver que

$$U(p) = \frac{\Psi_p(Z)}{\frac{1}{p}-\frac{1}{2}}$$

es no decreciente para ello se verá primero que la función $\alpha(t) = t\log(\mathrm{E}[z^{1/t}])$ es una función convexa para $t \in (1/2, 1]$. Sea $a \in [0,1]$ y $t_1, t_2 \in (1/2,1]$ y recuérdese que α es convexa si

$$\begin{aligned} \alpha(at_1 + (1-a)t_2) &\leq a\alpha(t_1) + (1-a)\alpha(t_2) \\ &= at_1\log(\mathrm{E}[z^{1/t_1}]) + (1-a)t_2\log(\mathrm{E}[z^{1/t_2}]) \\ &= \log(\mathrm{E}[z^{1/t_1}])^{(at_1} + \log(\mathrm{E}[z^{1/t_2}])^{(1-a)t_2}. \end{aligned}$$

Ahora bien, por la desigualdad de *Hölder* para

$$1/p = (at_1)/(at_1 + (1-a)t_2) \quad , \quad 1/q = ((1-a)t_2)/(at_1 + (1-a)t_2)$$

y

$$X = Z^{1/t_1 p} \quad , \quad Y = Z^{1/t_2 p}$$

se tiene

$$\begin{aligned} \mathrm{E}[XY] = \mathrm{E}[Z^{1/(at_1+(1-a)t_2)}] &\leq (\mathrm{E}[X^p])^{1/p} (\mathrm{E}[Y^q])^{1/q} \\ &= (\mathrm{E}[Z^{1/t_1}) ^{\frac{at_1}{at_1+(1-a)t_2}} (\mathrm{E}[Z^{1/t_2}) ^{\frac{(1-a)t_2}{at_1+(1-a)t_2}} \end{aligned}$$

y elevando a ambos miembros de la desigualdad a $at_1 + (1-a)t_2$ y aplicando logaritmos se obtiene el resultado, por lo que α es convexa. Por tanto, como la exponencial es creciente y convexa, entonces la función $\beta(t) = e^{2\alpha(t)} = (\mathrm{E}f^{1/t})^{2t}$ es convexa. Además vamos a comprobar que

$$\frac{\beta(t) - \beta(1/2)}{t - 1/2} \quad \text{es no decreciente en} \quad (1/2, 1]$$

para ello veamos que si $t < t'$ se tiene

$$\frac{\beta(t) - \beta(1/2)}{t - 1/2} \leq \frac{\beta(t') - \beta(1/2)}{t' - 1/2}.$$

En primer lugar nótese que

$$t = \frac{t' - t}{t' - 1/2} \frac{1}{2} + \frac{t - 1/2}{t' - 1/2} t'$$

y que $\frac{t'-t}{t'-1/2} \in (0,1)$, $\frac{t-1/2}{t'-1/2} \in (0,1)$ y como β era convexa

$$\beta(t) \leq \frac{t' - t}{t' - 1/2} \beta(1/2) + \frac{t - 1/2}{t' - 1/2} \beta(t')$$

lo que implica

$$\beta(t) - \beta(1/2) \leq \frac{t - 1/2}{t' - 1/2} \beta(t') - \frac{t - 1/2}{t' - 1/2} \beta(1/2)$$

y pasando el factor común $t - 1(/2)$ al otro miembro de la desigualdad se tiene

$$\frac{\beta(t) - \beta(1/2)}{t - 1/2} \leq \frac{\beta(t') - \beta(1/2)}{t' - 1/2}.$$

De aquí se concluye que la función

$$U(p) = \frac{\Psi_p(Z)}{\frac{1}{p} - \frac{1}{2}}$$

es no decreciente en $p \in [1, 2)$

□

4.3 Teorema. *La desigualdad logarítmica de Sobolev para una distribución simétrica Bernoulli* (4.1) *implica que para toda función*

$$f : \{-1, 1\}^n \to \mathbb{R}, \quad \mathrm{Var}(f(X)) \leq \mathcal{E}(f).$$

Demostración. En primer lugar se verá que

$$\mathrm{Ent}((1+\epsilon f)^2) = 2\epsilon^2 \mathrm{Var}(f(X)) + \epsilon^3 L(\epsilon)$$

con $L(\epsilon) = L_1(\epsilon) - L_2(\epsilon)$ acotada en un entorno de 0 y ϵ suficientemente pequeño.
Sea $0 < ||f||_\infty < \infty$. Si se aplica la definición de entropía a la función $((1+\epsilon f)^2)$ se tiene

$$\mathrm{Ent}((1+\epsilon f)^2) = \mathrm{E}((1+\epsilon f)^2 \log(1+\epsilon f)^2) - \mathrm{E}(1+\epsilon f)^2 \log[\mathrm{E}((1+\epsilon f)^2)].$$

Llámense A y B a cada uno de los sumandos de modo que

$$\mathrm{Ent}((1+\epsilon f)^2) = A - B.$$

Ahora bien, recuérdese que el desarrollo limitado del $\log(1+x)$ es el siguiente

$$\log(1+x) = x - \frac{x^2}{2} + x^3 R(x) \quad \text{si} \quad |x| < 1$$

con

$$R(x) = \sum_{n=3}^{\infty} \frac{(-1)^{n-1}}{n} x^{n-3}.$$

$$R(x) \in [\frac{1}{6}, \frac{3}{6}] \quad \text{si} \quad |x| \leq \delta_0 \quad \text{con} \quad \delta_0 > 0.$$

Tomo $\epsilon > 0$ tal que $2\epsilon||f||_\infty + \epsilon^2||f||_\infty^2 < 1/6$ entonces

$$\log(1+\epsilon f) = \epsilon f - \frac{(\epsilon f)^2}{2} + \epsilon^3 f^3 R(\epsilon f).$$

De modo que

$$\begin{aligned}
A &= 2\mathrm{E}\Big[(1+\epsilon^2 f^2 + 2\epsilon f)(\epsilon f - \frac{\epsilon^2 f^2}{2} + \epsilon^3 f^3 R(\epsilon f)\Big] \\
&= 2\mathrm{E}\Big[\epsilon f - \frac{\epsilon^2 f^2}{2} + \epsilon^3 f^3 R(\epsilon f) + 2\epsilon^2 f^2 + 2\epsilon^4 f^4 R(\epsilon f) - \frac{1}{2}\epsilon^4 f^4 + \epsilon^5 f^5 R(\epsilon f)\Big] \\
&= 2\epsilon \mathrm{E} f + 3\epsilon^2 \mathrm{E} f^2 + 2\epsilon^3 \mathrm{E}(f^3 R(\epsilon f)) + 2\epsilon^4 \mathrm{E}(2F^4 R(\epsilon f) - \frac{1}{2} f^4) + 2\epsilon^5 \mathrm{E}(f^5 R(\epsilon f)) \\
&= 2\epsilon \mathrm{E} f + 3\epsilon^2 \mathrm{E} f^2 + \epsilon^3 L_1(\epsilon) \quad \text{con} \quad L_1(\epsilon) \quad \text{acotado en un entorno de 0.}
\end{aligned}$$

y

$$\begin{aligned}
B &= \mathrm{E}(1+\epsilon^2 f^2+2\epsilon f)\log(1+\epsilon^2\mathrm{E}f^2+2\epsilon\mathrm{E}f)\\
&= \mathrm{E}(1+\epsilon^2 f^2+2\epsilon f)\Big[\epsilon^2\mathrm{E}f^2+2\epsilon\mathrm{E}f-\frac{(2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2)^2}{2}\Big]\\
&\quad+\mathrm{E}(1+\epsilon^2 f^2+2\epsilon f)\Big[(2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2)^3 R((2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2))\Big]\\
&= 2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2-\frac{(2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2)^2}{2}\\
&\quad+(2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2)^3 R((2\epsilon\mathrm{E}f+\epsilon^2\mathrm{E}f^2))+4\epsilon^2\mathrm{E}^2 f+\epsilon^3 L_2(\epsilon)\\
&\quad\text{con}\quad L_2(\epsilon)\quad\text{acotado en un entorno de 0.}
\end{aligned}$$

Llegando así a que

$$\mathrm{Ent}((1+\epsilon f)^2) = A - B = 2\epsilon^2\mathrm{Var}(f(x)) + \epsilon^3 L(\epsilon)$$

con $L(\epsilon) = L_1(\epsilon) - L_2(\epsilon)$ acotada en un entorno de 0. Por otro lado aplicando la desigualdad gaussiana de *Poincaré* a la función $((1+\epsilon f)^2)$ se tiene

$$2\epsilon^2\mathrm{Var}(f(x)) + \epsilon^3 L(\epsilon) \leq 2\mathcal{E}((1+\epsilon f)^2) = 2\epsilon^2\mathcal{E}(f).$$

Dividiendo ambos miembros de la desigualdad por ϵ^2 y haciendo tender ϵ a 0 se tiene

$$\mathrm{Var}(f(X)) \leq \mathcal{E}(f).$$

□

El nombre de desigualdad de *Poincaré* aplicado a la cota del Teorema (4.3) se justifica por analogía con la desigualdad gaussiana de *Poincaré* del capítulo 3, sustituyendo $E||\nabla f(X)||$ por su versión discreta, $\mathcal{E}(f)$.

4.2. Argumento de Herbst

El siguiente argumento, atribuido a *Herbst*(ver [1]) proporciona una desigualdad de concentración para $Z = f(X)$ a partir de una desigualdad logarítmica de *Sobolev*. Recuérdese que $f : \{-1,1\}^n \to \mathbb{R}$ y $X = (X_1, ..., X_n)$ es un vector binario distribuido uniformemente en el hipercubo $\{-1,1\}$.

Se usará la desigualdad logarítmica de *Sobolev* para distribución *Bernoulli* simétrica para la función no negativa $g(x) = e^{\lambda f(x)/2}$ donde $\lambda \in \mathbb{R}$ es un

parámetro cuyo valor se optimizará luego. Ahora bien, si se calcula la entropía de g^2 se tiene,

$$\mathrm{Ent}(g^2) = \mathrm{Ent}(e^{\lambda f(x)}) = \lambda E[Ze^{\lambda Z}] - Ee^{\lambda Z} \log Ee^{\lambda Z}.$$

Haciendo uso de que $F(\lambda) = Ee^{\lambda Z}$ es la función generadora de momentos de Z y $F'(\lambda) = E[Ze^{\lambda Z}]$ es su derivada, se puede escribir,

$$\mathrm{Ent}(g^2) = \lambda F'(\lambda) - F(\lambda) \log F(\lambda).$$

Por el Teorema (4.1), se tiene

$$\begin{aligned}
\mathrm{Ent}(g^2) &\leq \frac{1}{2} \sum_{i=1}^{n} \mathrm{E}\Big[\Big(e^{\lambda f(x)/2} - e^{\lambda f(\bar{x}^{(i)})/2}\Big)^2\Big] \\
&= \sum_{i=1}^{n} \mathrm{E}\Big[\Big(e^{\lambda f(x)/2} - e^{\lambda f(\bar{x}^{(i)})/2}\Big)_+^2\Big]
\end{aligned}$$

donde se usa el hecho de que X y $\bar{X}^{(i)}$ tienen la misma distribución. Por otro lado como la función exponencial es convexa se tiene que para todo número real $z > y$,

$$e^{z/2} - e^{y/2} \leq \frac{(z-y)}{2} e^{z/2}.$$

Por lo que ahora se tiene,

$$\begin{aligned}
\mathrm{Ent}(g^2) &\leq \frac{\lambda^2}{4} \sum_{i=1}^{n} \mathrm{E}\Big[\Big(f(x) - f(\bar{x}^{(i)})\Big)_+^2 e^{\lambda f(x)}\Big] \\
&= \frac{\lambda^2}{4} \mathrm{E}\Big[e^{\lambda f(x)} \sum_{i=1}^{n} \Big(f(x) - f(\bar{x}^{(i)})\Big)_+^2\Big].
\end{aligned}$$

Sea

$$\nu = \max_{x \in \{-1,1\}^n} \sum_{i=1}^{n} \Big(f(x) - f(\bar{x}^{(i)})\Big)_+^2$$

entonces

$$\mathrm{Ent}(g^2) = \mathrm{Ent}(e^{\lambda f(x)}) \leq \frac{\nu\lambda^2}{4} \mathrm{E}e^{\lambda f(x)}.$$

Expresando esta desigualdad en términos de la función generadora de momentos F, se tiene,

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \frac{\nu\lambda^2}{4} F(\lambda).$$

Ahora el objetivo es resolver ésta desigualdad diferencial. Para ello se dividen ambas partes por el término positivo $\lambda^2 F(\lambda)$,teniendo así

$$\frac{1}{\lambda}\frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) \leq \frac{\nu}{4}.$$

Llamando $G(\lambda) = \log F(\lambda)$ se observa que $G'(\lambda) = \frac{F'(\lambda)}{F(\lambda)}$. Por lo que la desigualdad diferencial se puede reescribir como

$$\Big(\frac{G(\lambda)}{\lambda}\Big)' \leq \frac{\nu}{4}.$$

Por la regla de L'Hôpital se tiene

$$\lim_{\lambda\to 0}\frac{G(\lambda)}{\lambda} = \lim_{\lambda\to 0}\frac{\log F(\lambda)}{\lambda} = \lim_{\lambda\to 0}\frac{F'(\lambda)}{F(\lambda)} = \frac{F'(0)}{F(0)} = \mathrm{E}Z.$$

Ahora cabe distinguir dos casos, si $\lambda > 0$ ó si $\lambda < 0$.

- Si $\lambda > 0$, integrando la desigualdad entre 0 y λ, se tiene

  $$\frac{G(\lambda)}{\lambda} \leq \mathrm{E}Z + \frac{\lambda\nu}{4}$$

  ,es decir,

  $$G(\lambda) \leq \lambda\mathrm{E}Z + \frac{\lambda^2\nu}{4}$$

  tomando exponenciales a ambos lados

  $$F(\lambda) = \mathrm{E}e^{\lambda Z} \leq e^{\lambda\mathrm{E}Z + \frac{\lambda^2\nu}{4}}$$

  y por la desigualdad de *Markov* se tiene

  $$P(Z > EZ + t) \leq \inf_{\lambda>0} F(\lambda)e^{-\lambda(\mathrm{E}Z+t)} \leq \inf_{\lambda>0} e^{\lambda^2\nu/4-\lambda t} = e^{-\frac{t^2}{\nu}}$$

  donde la última igualdad se obtiene de estudiar los extremos de la función $\lambda^2\nu/4 - \lambda t$.

- Si $\lambda < 0$, se estudia de forma similar. Se integra $G'(\lambda) = \frac{F'(\lambda)}{F(\lambda)}$ entre $-\lambda$ y 0 y se obtiene

  $$F(\lambda) = \mathrm{E}e^{\lambda Z} \leq e^{\lambda\mathrm{E}Z + \frac{\lambda^2\nu}{4}}$$

  lo que implica la desigualdad de cola por la izquierda,

  $$P(Z < EZ - t) \leq \inf_{\lambda<0} F(\lambda)e^{-\lambda(\mathrm{E}Z-t)} \leq \inf_{\lambda<0} e^{\lambda^2\nu/4+\lambda t} = e^{-\frac{t^2}{\nu}}.$$

  Como consecuencia de lo anterior

  $$P(Z < EZ - t) \leq e^{-\frac{t^2}{\nu}}, t > 0.$$

4.3. Desigualdad gaussiana logarítmica de Sobolev

En esta sección se usa la desigualdad logarítmica de *Sobolev* en la distribución simétrica de *Bernoulli* para derivar a un resultado análogo bajo la distribución gaussiana en $\mathbb{R}^n$. Además del interés propio de las desigualdades logarítmicas de *Sobolev* para el hipercubo binario éstas se pueden usar como resultados intermedios útiles para probar la desigualdad gaussiana logarítmica de *Sobolev* y una serie de resultados relacionados.

4.4 Teorema (Desigualdad gaussiana logarítmica de Sobolev). *Sea $X = (X_1, ..., X_n)$ un vector de n variables aleatorias independientes con distribución normal estándar y sea $f : \mathbb{R}^n \to \mathbb{R}$ una función continuamente diferenciable. Entonces,*

$$\mathrm{Ent}(f^2) \leq 2E[||\nabla f(x)||^2].$$

Demostración. Primero se probará para $n = 1$, esto es, cuando $f : \mathbb{R} \to \mathbb{R}$ es una función continuamente diferenciable en la recta real y X es una variable aleatoria con distribución normal estándar. Véase que si $E[f'(x)^2] = \infty$, no hay nada que probar por lo que se asume $E[f'(x)^2] < \infty$. Por un argumento de como el usado en el Teorema visto en (2.4.2) es suficiente probar el teorema para funciones doblemente diferenciables con soporte acotado.

Sean $\epsilon_1, \epsilon_2, ..., \epsilon_n$ variables aleatorias *Rademacher* e independientes. Recuérdese de la prueba de la desigualdad gaussiana de *Poincaré* que

$$\lim_{n\to\infty} \mathrm{E}\Big[\sum_{j=1}^{n} \Big| f\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big) - f\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i - 2\frac{\epsilon_j}{\sqrt{n}}\Big)\Big|^2\Big] = 4\mathrm{E}[f'(X)^2].$$

Por otra parte, para cualquier función f continua uniformemente acotada, por el teorema central del límite se tiene

$$\lim_{n\to\infty} \mathrm{Ent}\Big[f^2\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big)\Big] = \mathrm{Ent}[f(X)^2].$$

La prueba se completa haciendo uso de la desigualdad logarítmica de *Sobolev* para una distribución simétrica *Bernoulli* (4.1) que afirma que,

$$\mathrm{Ent}\Big[f^2\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big)\Big] \leq \frac{1}{2}\mathrm{E}\Big[\sum_{j=1}^{n} \Big| f\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big) - f\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i - 2\frac{\epsilon_j}{\sqrt{n}}\Big)\Big|^2\Big]. \tag{4.2}$$

La extensión del resultado para $n \geq 1$ se sigue del teorema de la subaditividad de la entropía (3.15) por el que se tiene

$$\mathrm{Ent}(f^2) \leq \sum_{i=1}^{n} \mathrm{E}\Big[\mathrm{E}^{(i)}[f(X)^2 \log f(X)^2] - \mathrm{E}^{(i)}[f(X)^2] \log \mathrm{E}^{(i)}[f(X)^2]\Big].$$

Y por el resultado probado para $n = 1$ (4.2) se llega a

$$\mathrm{E}^{(i)}[f(X)^2 \log f(X)^2] - \mathrm{E}^{(i)}[f(X)^2] \log \mathrm{E}^{(i)}[f(X)^2] \leq 2\mathrm{E}^{(i)}[(\partial_i f(X))^2]$$

y como $||\nabla f(x)||^2 = \sum_{i=1}^{n} (\partial_i f(X))^2$ se completa la prueba.

□

4.4. Concentración gaussiana: Desigualdad de Tsirelson-Ibragimov-Sudakov

Igual que la desigualdad logarítmica de *Sobolev* para distribuciones *Bernoulli* simétricas conduce mediante el argumento de *Herbst* a desigualdades exponenciales de concentración, la desigualdad logarítmica gaussiana de *Sobolev* lleva a desigualdades de colas exponenciales para funciones suaves de variables aleatorias independientes normales. A este resultado se le conoce como desigualdad de concentración gaussiana y es el que aparece a continuación.

4.5 Teorema. *Sea $X = (X_1, ..., X_n)$ un vetcor de n variables aleatorias independientes con distribución normal estándar y sea $f : \mathbb{R}^n \to \mathbb{R}$ una función L-Lipschitziana, esto es, que existe una constante $L > 0$ tal que para todo $x, y \in \mathbb{R}^n$*

$$|f(x) - f(y)| \leq L||x - y||.$$

Entonces, para todo $\lambda \in \mathbb{R}$,

$$\log \mathrm{E} e^{\lambda(f(x) - \mathrm{E}f(x))} \leq \frac{\lambda^2}{2} L^2.$$

Demostración. De nuevo por un argumento de densidad como el del Teorema visto en (2.4.2) se puede asumir que f es diferenciable con gradiente uniformemente acotado por L. También se puede asumir que $\mathrm{E}f(x) = 0$. Usando

la desigualdad logarítmica gaussiana de *Sobolev* para la función $e^{\lambda f/2}$, se obtiene,

$$\mathrm{Ent}(e^{\lambda f}) \leq 2E[||\nabla e^{\lambda f(x)/2}||^2] = \frac{\lambda^2}{2} E[e^{\lambda f(x)}||\nabla f(x)||^2] \leq \frac{\lambda^2}{2} L^2 E e^{\lambda f(x)}.$$

Denotando $F(\lambda) = Ee^{\lambda f(x)}$ y $F'(\lambda) = E[f(x)e^{\lambda f(x)}]$ es su derivada, se puede escribir,

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

para resolver ésta desigualdad diferencial, se dividen ambas partes por el término positivo $\lambda^2 F(\lambda)$, teniendo así

$$\frac{1}{\lambda}\frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) \leq \frac{\nu}{4}.$$

Llamando $G(\lambda) = \log F(\lambda)$ se observa que $G'(\lambda) = \frac{F'(\lambda)}{F(\lambda)}$. Por lo que la desigualdad diferencial se puede reescribir como

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

Por la regla de L'Hôpital se tiene

$$\lim_{\lambda\to 0} \frac{G(\lambda)}{\lambda} = \lim_{\lambda\to 0} \frac{\log F(\lambda)}{\lambda} = \lim_{\lambda\to 0} \frac{F'(\lambda)}{F(\lambda)} = \frac{F'(0)}{F(0)} = \mathrm{E}Z.$$

Ahora cabe distinguir dos casos, si $\lambda > 0$ ó si $\lambda < 0$.

- Si $\lambda > 0$, integrando la desigualdad entre 0 y λ, se tiene

  $$\frac{G(\lambda)}{\lambda} \leq \mathrm{E}Z + \lambda\frac{L^2}{2},$$

  es decir,

  $$G(\lambda) \leq \lambda \mathrm{E}Z + \lambda\frac{L^2}{2}$$

  tomando exponenciales a ambos lados

  $$F(\lambda) = \mathrm{E}e^{\lambda Z} \leq e^{\lambda \mathrm{E}Z + \lambda\frac{L^2}{2}}$$

  y como $EZ = 0$

  $$F(\lambda) \leq e^{\lambda^2\frac{L^2}{2}},$$

  es decir,

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

- Si $\lambda < 0$, se estudia de forma similar. Se integra $G'(\lambda) = \frac{F'(\lambda)}{F(\lambda)}$ entre $-\lambda$ y 0 y se obtiene

$$F(\lambda) = \mathrm{E}e^{\lambda Z} \leq e^{\lambda \mathrm{E}Z + \lambda \frac{L^2}{2}}$$

y como $EZ = 0$

$$F(\lambda) \leq e^{\lambda^2 \frac{L^2}{2}},$$

es decir,

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

□

Al igual como ocurría para la desigualdad logarítmica de *Sobolev* para una distribución simétrica *Bernoulli* con la cota sub-Gaussiana obtenida para la función generadora de momentos se llega a una desigualdad de cola exponencial vía la desigualdad de *Markov*. Más precisamente se obtiene el siguiente resultado

4.6 Teorema (**Desigualdad de concentración gaussiana**). *Sea* $X = (X_1, ..., X_n)$ *un vector de n variables aleatorias independientes con distribución normal estándar y sea* $f : \mathbb{R}^n \to \mathbb{R}$ *una función L-Lipschitziana entonces, para todo* $t > 0$,

$$P(f(x) - Ef(x) \geq t) \leq e^{-t^2/2L^2}.$$

Se puede ver que, al contrario de lo que ocurría en la desigualdad de concentración para variables aleatorias simétricas de *Bernoulli*, la cota obtenida no depende de la dimensión n.

Una aplicación significativa del teorema anteriormente visto es la siguiente

4.7 Ejemplo (**Norma de un vector gaussiano**). Sea $X = X_1, ..., X_n$ un vector cuyas variables siguen la distribución normal con media 0 y matriz de covarianzas Γ. Sea $p \geq 1$ y se considera la variable aleatoria Z con valores reales definida por la p-norma de X, esto es,

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

Como Γ es semidefinida positiva, existe una matriz A $n \times n$ que satisface $A^T A = \Gamma$. Entonces el vector gaussiano X se distribuye como AY donde $Y = (Y_1, ..., Y_n)$ se distribuye conforme a la distribución gaussiana canónica, es decir, las componentes de Y son variables aleatorias independientes con distribución normal estándar. Entonces $f(y) = ||Ay||_p$ es una función lipschitziana de $\mathbb{R}^n$ a $\mathbb{R}$ con constante *Lipschitz* L igual al operador norma de A que hace corresponder l^2 a l^p, esto es,

$$L = ||A||_{l^2 \to l^p} \stackrel{def}{=} \sup_{y \in \mathbb{R}^n : ||y||_2 = 1} ||Ay||_p.$$

Con esta notación,

$$P(||X||_p - E||X||_p \geq t) \leq e^{-t^2/2||A||^2}.$$

4.5. Desigualdad de concentración para el supremo de procesos gaussianos

En esta sección se ilustrará la desigualdad de concentración gaussiana viendo como ésta implica de una forma simple una desigualdad de concentración para el supremo de un proceso gaussiano. Una característica clave de la desigualdad de concentración gaussiana es que la cota superior no depende de la dimensión n. Esto nos permite extenderlo fácilmente a un escenario de dimensión infinita que se describirá a continuación. Sea $\mathcal{T}$ y sea $(X_t)_{t \in \mathcal{T}}$ un proceso gaussiano indexado por $\mathcal{T}$, es decir, que una variable aleatoria X_t es asignada a todo $t \in \mathcal{T}$ y para toda colección finita $\{t_1, ..., t_n\} \subset \mathcal{T}$ el vector $(X_{t_1}, ..., X_{t_n})$ tiene conjuntamente distribución gaussiana con media 0. Además, se asume que $\mathcal{T}$ está totalmente acotada (es decir, para todo $t > 0$ se puede recubrir por un número finito de bolas de radio t) y el proceso gaussiano es continuo casi seguro, es decir, con probabilidad 1 X_t es una función continua en t.

4.8 Teorema. *Sea $(X_t)_{t \in \mathcal{T}}$ un proceso gaussiano centrado y continuo (casi seguro) e indexado por un conjunto $\mathcal{T}$ precompacto para $d(s,t) = \sqrt{E(X_t - X_s)^2}$. Si*

$$\sigma^2 = \sup_{t \in \mathcal{T}} E[X_t^2],$$

entonces $Z = \sup_{t\in\mathcal{T}} X_t$ *satisface que* $\mathrm{Var}(Z) \leq \sigma^2$ *y para todo* $u > 0$,

$$P(Z - EZ \geq u) \leq e^{-u^2/(2\sigma^2)}$$

y

$$P(EZ - Z \geq u) \leq e^{-u^2/(2\sigma^2)}$$

Demostración. Se asume que $\mathcal{T}$ es un conjunto finito. Si $\mathcal{T}$ no es finito se puede extender a un $\mathcal{T}$ precompacto arbitrario pues en particular $\mathcal{T}$ es separable, por lo que existe un subconjunto $\mathcal{D}$ denso y numerable tal que

$$\sup_{t\in\mathcal{T}} X_t = \sup_{t\in\mathcal{D}} X_t.$$

Sea

$$D = \{d_n\}_{n\geq 1}$$

y

$$Z = \sup_{t\in\mathcal{D}} X_t = \lim_{n\to\infty} \sup(X_{d_1}, ..., X_{d_n}) = \lim_{n\to\infty} Z_n$$

con $Z_n = \sup(X_{d_1}, ..., X_{d_n})$. Por el teorema de la convergencia monótona se tiene que $\lim_{n\to\infty} EZ_n = EZ$. Como $(X_t)_{t\in\mathcal{T}}$ es un proceso gaussiano $P(Z_n \geq u + EZ_n) \leq e^{-u^2/(2\sigma_n^2)}$ con $\sigma_n^2 = \sup_{m\leq n} EX_{d_m}^2$. Además se tiene que

$$\lim_{n\to\infty} \sigma_n^2 = \lim_{n\to\infty} \sup_{m\leq n} EX_{d_m}^2 = \sigma^2 = \sup_{n\geq 1} EX_{d_n}^2.$$

Se considera $\sigma^2 < \infty$ y se tiene $\sigma_n^2 \leq \sigma^2$. Como Z_n es positiva,

$$EZ_n \leq \sqrt{EZ_n^2} \leq \sqrt{\mathrm{Var} Z_n} \leq \sigma_n \leq \sigma.$$

De modo que

$$EZ \leq \sigma < \infty.$$

Como Z_n converge a Z en distribución se tiene,

$$\lim_{n\to\infty} P(Z_n \geq t) = P(Z \geq t) \quad \text{para todo t de continuidad de Z.}$$

En concreto,

$$\lim_{n\to\infty} P(Z_n - EZ_n \geq t) = P(Z - EZ \geq t) \quad \text{para todo t de continuidad de Z.}$$

y

$$\lim_{n\to\infty} e^{-t^2/(2\sigma_n^2)} = e^{-t^2/(2\sigma^2)}$$

de modo que,

$$P(Z - EZ \geq t) \leq e^{-t^2/(2\sigma^2)} \quad \forall t > 0.$$

Se asume por simplicidad que $\mathcal{T} = \{1, ..., n\}$. Sea Γ la matriz de covarianza del vector centrado gaussiano $X = (X_1, ..., X_n)$. Se denota por A la raíz cuadrada de la matriz semidefinida positiva Γ, es decir, $A^T A = \Gamma$. Si $Y = (Y_1, ..., Y_n)$ es un vector de variables aleatorias independientes con distribución normal estándar, entonces $f(Y) = \max_{i=1,...,n}(AY)$ tiene la misma distribución que $\max_{i=1,...,n} X_i$. Por lo tanto se puede aplicar la desigualdad de concentración gaussiana acotando la constante Lipschitziana de f. Por la desigualdad de *Cauchy-Scharwz* para todo $u, v \in \mathbb{R}^n$ y $i = 1, ..., n$

$$|(Au)_i - (Av)_i| = |\sum_j A_{i,j}(u_j - v_j)| \leq \left(A_{i,j}^2\right)^{1/2} ||u - v||.$$

Como $\sum_j A_{i,j}^2 = \text{Var}(X_i)$, se tiene

$$|f(u) - f(v)| \leq \max_{i=1,...,n} |(Au)_i - (Av)_i| \leq \sigma ||u - v||.$$

Además f es lipschitziana con constante σ : $||f||_L \leq \sigma$ y se sigue de la desigualdad de concentración gaussiana que las colas están acotadas, además como $\text{Var}(f(X)) \leq L^2$ y en este caso $L = \sigma$ se obtiene así la cota de la varianza.

□

Bibliografía

[1] S. Boucheron, O. Bousquet, G. Lugosi; *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[2] P. Massart; *Concentration Inequalities and Model Selection*, Lecture Notes in Mathematics Vol. 1896. Ecole d'Eté de Probabilités de Saint-Flour XXXIII- 2003, Springer.

[3] J.M Sanz-Serna; *Diez lecciones de cálculo numérico*, Secretariado de Publicaciones e Intercambio Científico, Universidad de Valladolid, 1998.

[4] L.C Evans; *Partial differential equations*, 2nd edition, American Math Society, 2010

[5] P. Billingsley; *Probability and measure*, 1986. Wiley series in probability and mathematical statistics.