**UNIVERSIDAD DE BUENOS AIRES**
**Facultad de Ciencias Exactas y Naturales**
**Departamento de Matemática**

**Tesis de Licenciatura**

# DISTRIBUCIÓN DE FORMAS CUADRÁTICAS, Y GEODÉSICAS EN LA CURVA MODULAR

**Alejo Salvatore**

**Director: Miguel Walsh**

13 de marzo de 2020

# Índice general

# Introducción

Una forma cuadrática binaria es un polinomio de la forma $q(x,y) = ax^2 + bxy + cy^2$ y su discriminante está dado por $d = b^2 - 4ac$. Las formas cuadráticas binarias han sido objeto de estudio en teoría de números por siglos. El objetivo de esta tesis es demostrar un teorema de equidistribución del conjunto de formas cuadráticas sobre $\mathbb{Z}$ de igual discriminante $d$, cuando $d \to +\infty$. Basta considerar las formas cuadráticas *primitivas*, es decir aquellas cuyos coeficientes no tienen un divisor en común. Las formas cuadráticas binarias primitivas de discriminante $d$ se identifican con el conjunto

$$R_{\text{disc}}(d) = \{(a,b,c) \in \mathbb{Z}^3 : \text{mcd}(a,b,c) = 1,\, b^2 - 4ac = d\}.$$

Un *discriminante* es un entero $d$ para el cual $R_{\text{disc}}(d)$ es no vacío. Resulta natural preguntar cómo se distribuye el conjunto $R_{\text{disc}}(d)$. Es claro que $|d|^{-1/2}R_{\text{disc}}(d)$ está contenido en el conjunto

$$V^+_{\text{disc}}(\mathbb{R}) = \{(a,b,c) \in \mathbb{R}^3 : b^2 - 4ac = 1\} \tag{1}$$

de formas cuadráticas reales de discriminante 1. Este espacio posee una medida $\mu_{\text{disc}}$ definida de la siguiente manera: para $A \subseteq V^+_{\text{disc}}(\mathbb{R})$ medible, $\mu_{\text{disc}}(A)$ es la medida de Lebesgue del cono $C(A) = \{(a,b,c) \in \mathbb{R}^3 : 0 < b^2 - 4ac \leq 1,\ \text{y } (b^2-4ac)^{-1/2}(a,b,c) \in A\}$. En los años 50, Linnik comenzó a estudiar la distribución de $|d|^{-1/2}R_{\text{disc}}(d)$ en $V^+_{\text{disc}}(\mathbb{R})$ y problemas similares sobre la distribución de las soluciones de $Q(a,b,c) = d$ a medida que $d \to \pm\infty$, donde $Q$ es una forma cuadrática ternaria (ver [Lin68]). Este trabajo fue continuado por su estudiante Skubenko. El resultado que buscaban demostrar es el siguiente:

**Teorema 0.1** (Duke)**.** *A medida que $d \to +\infty$ entre los disriminantes positivos, el conjunto*

$$d^{-1/2}R_{\text{disc}}(d) \subseteq V^+_{\text{disc}}(\mathbb{R})$$

*se vuelve equidistribuido en $V^+_{\text{disc}}(\mathbb{R})$ con respecto a la medida $\mu_{\text{disc}}$, en el siguiente sentido: para todo par de funciones $\varphi_1, \varphi_2 \in C_c(V^+_{\text{disc}}(\mathbb{R}))$ tales que $\mu_{\text{disc}}(\varphi_2) \neq 0$, se tiene*

$$\frac{\sum_{x \in R_{\text{disc}}(d)} \varphi_1(x/\sqrt{d})}{\sum_{x \in R_{\text{disc}}(d)} \varphi_2(x/\sqrt{d})} \to \frac{\mu_{\text{disc}}(\varphi_1)}{\mu_{\text{disc}}(\varphi_2)}. \tag{2}$$

En 1962, Skubenko [Sku62] demostró el teorema para los discriminantes $d > 0$ tales que $\left(\frac{d}{p}\right) = 1$ para un primo $p$ fijo, donde $\left(\frac{\cdot}{p}\right)$ es el símbolo de Legendre. Aunque la condición $\left(\frac{d}{p}\right) = 1$ es esencial en el método de demostración de Linnik, resulta arbitraria y se esperaba que el resultado valiera en general. Duke [Duk88] logró eliminar la condición $\left(\frac{d}{p}\right) = 1$ recién dos décadas más tarde, usando métodos de análisis armónico y en particular una cota no trivial para sumas de Kloosterman usando ideas y resultados de Iwaniec [Iwa87].

Una de las ideas clave de la demostración de Duke consiste en reinterpretar el problema en términos de la distribución de un conjunto finito de geodésicas en el campo tangente unitario de la curva modular $Y_0(1)$. A cada forma cuadrática $q(x,y) = ax^2 + bxy + cy^2$ en $R_{\text{disc}}(d)$ se le puede asignar el par de puntos $\frac{-b \pm \sqrt{d}}{2a} \in \mathbb{R}$. En el plano hiperbólico $\mathbb{H} = \{z \in \mathbb{C} : \text{Im}(z) > 0\}$ hay una única geodésica cuyos extremos son estos dos puntos y es la semicircunferencia que los une (en el sentido de geometría euclideana). En $Y_0(1) = \text{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$, esta geodésica se vuelve una curva cerrada $\gamma_q$. Sea

$$\mathcal{G}_d = \bigcup_{q \in R_{\text{disc}}(d)} \gamma_q. \tag{3}$$

El teorema de Duke es equivalente al siguiente resultado:

**Teorema 0.2** (Duke, versión 2)**.** *A medida que $d \to +\infty$ entre los discriminantes fundamentales, el conjunto $\mathcal{G}_d$ se vuelve equidistribuido en $T^1(Y_0(1))$ con respecto a la medida de Haar.*

El enunciado preciso de este teorema será dado en el comienzo del capítulo 2, así como la demostración de la equivalencia entre los teoremas 0.1 y 0.2. Recientemente, Venkatesh, Lindenstrauss, Einsiedler y Michel [Ein+14] hallaron una demostración original del teorema de Duke, usando teoría ergódica, y éste es el enfoque que utilizaremos en esta tesis. La demostración tiene dos ingredientes principales: el *lema básico de Linnik* 4.5, que ya estaba presente en el método original de Linnik, y la unicidad de la medida de máxima entropía para el flujo asociado al grupo uniparamétrico $a_t = \left(\begin{smallmatrix} e^t & 0 \\ 0 & e^{-t} \end{smallmatrix}\right)$ en $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$.

Vale la pena mencionar que un método similar fue utilizado por los mismos autores en [Ein+11], donde demuestran un resultado análogo para $X_3 = \mathrm{PGL}_3(\mathbb{Z})\backslash\mathrm{PGL}_3(\mathbb{R})$. En este caso las geodésicas son reemplazadas por subvariedades riemannianas completas totalmente geodésicas de curvatura seccional 0 maximales. Sea $H$ el subgrupo de matrices diagonales en $\mathrm{PGL}_3(\mathbb{R})$. A cada órbita cerrada $xH$ en $X_3$ le asignan un "discriminante" que mide la complejidad aritmética de la órbita. Luego demuestran el siguiente resultado:

**Teorema 0.3.** *Las órbitas periódicas de $H$ en $X_3$ se pueden agrupar en clases de equivalencia de modo que dos órbitas equivalentes tienen igual volumen y discriminante. Para cada órbita periódica $xH$, sea $Y_{xH}$ la unión de todas las órbitas compactas equivalentes a $xH$. Si $(x_iH)_i$ es una sucesión de órbitas compactas tales que* $\mathrm{disc}(x_iH) \to +\infty$*, entonces*

(*a*) $\mathrm{vol}(Y_{x_iH}) = \mathrm{disc}(x_iH)^{1/2+o(1)}$.
(*b*) *Los conjuntos $Y_{x_iH}$ se vuelven uniformemente distribuidos en $X_3$ con respecto a la medida de Haar.*

Este teorema se puede interpretar de varias maneras y tiene, por ejemplo, el siguiente corolario:

**Corolario 0.4.** *Fijemos un primo $p > 3$. Entonces, cuando $d \to +\infty$ entre los enteros que no son cubos perfectos y tales que $p$ no se parte completamente en $\mathbb{Q}(\sqrt[3]{d})$, la sucesión de conjuntos $d^{-1/3}\{M \in M_3(\mathbb{Z}) : M^3 = dI\}$ se vuelve equidistribuido en $\{M \in M_3(\mathbb{Z}) : M^3 = I\}$.*

En el capítulo 1 desarrollamos los resultados de teoría ergódica que necesitaremos para la demostración. Las principales herramientas son el teorema ergódico puntual de Birkhoff, la descomposición ergódica y las cotas a la entropía que demostraremos en la última sección.

En el capítulo 2 estudiamos las propiedades métricas básicas del plano hiperbólico, definimos el flujo hiperbólico en esta variedad y en cocientes de la misma por retículos, y probamos que es ergódico con respecto a la medida de Haar en este último caso.

El capítulo 3 está dedicado al estudio de formas cuadráticas binarias y cuerpos cuadráticos. En él estudiaremos los espacios cuadráticos y los órdenes en cuerpos cuadráticos. En particular demostraremos la correspondencia entre ideales inversibles sobre estos órdenes y clases de formas cuadráticas binarias módulo la acción de $\mathrm{GL}_2(\mathbb{Z})$. Luego demostraremos una cota sobre representaciones de formas cuadráticas que luego usaremos en el último capítulo para demostrar el lema básico de Linnik, que es uno de los ingredientes clave de la demostración.

Finalmente, en el capítulo 4 daremos la demostración del teorema de Duke.

## Notación asintótica

Dadas dos funciones $f, g$ a valores complejos con igual dominio, la notación $f(x) = O(g(x))$ denota que existe una constante $c > 0$ tal que $|f(x)| \leq c|g(x)|$ para todo $x$ en el dominio de $f$ y $g$. Más en general, dada una función $F : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{C}$, usamos la notación $f(x) = F(x, O(g(x)))$ para denotar que existe una función $r(x) = O(g(x))$ tal que $f(x) = F(x, r(x))$ para todo $x$. Por ejemplo, $f(x) = h(x) + O(g(x))$ se traduce como $f(x) - h(x) = O(g(x))$, mientras que la condición $f(x) = x^{O(1)}$ es equivalente a pedir la existencia de una constante $c > 0$ tal que $|f(x)| \leq x^c$ para todo $x$. En el caso que la constante implícita dependa de algún o varios parámetros, todos estos

parámetros deben ser especificados ubicándolos como subíndice de $O$. Por ejemplo, la ecuación $f(x) = O_\varepsilon(x^\varepsilon)$ significa que existe una constante $C(\varepsilon) > 0$ tal que $|f(x)| \leq C(\varepsilon)x^\varepsilon$.

Dadas dos funciones $f, g : \Omega \to \mathbb{C}$ definidas en un conjunto $\Omega \subseteq \mathbb{R}$ y un punto de acumulación $x_0 \in [-\infty, +\infty]$ de $\Omega$, decimos que $f(x) = o(g(x))$ cuando $x \to x_0$ si

$$\lim_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| = 0.$$

Los mismos comentarios hechos sobre la notación $O$ también son aplicables para la notación $o$. En vez de la notación $f(x) = O(g(x))$ utilizaremos a veces la notación equivalente $f(x) \ll g(x)$, cuya ventaja radica en que es posible concatenar varias de estas expresiones, aprovechando el hecho de que $\ll$ es una relación transitiva entre funciones. Análogamente, escribir $f(x) \ll_{y_1,\ldots,y_n} g(x)$ es equivalente a $f(x) = O_{y_1,\ldots,y_n}(g(x))$.

# Agradecimientos

A mi director de tesis Miguel Walsh, por su paciencia, sus consejos y sugerencias.

A los miembros del jurado, Fernando Cuckierman y Román Sasyk, por su tiempo, sus comentarios y sugerencias.

A todos los excelentes profesores que tuve a lo largo de la carrera. A Gustavo Krimker, por ayudarme en mis primeros pasos en olimpíadas.

A mis compañeros de la facultad, por hacer más amenas las horas de cursada y estudio.

A mis padres por su apoyo y a mi hermana por haber infundido en mí una predilección por la ciencia.

# Índice de Notación

| | |
|---|---|
| $(G : H)$ | Índice de $H$ en $G$ |
| $[\alpha_1, \ldots, \alpha_n]$ | El $\mathbb{Z}$-módulo generado por $\alpha_1, \ldots, \alpha_n$, pág. 52 |
| $[\mathscr{L}_2(\mathbb{R})]$ | El espacio de retículos de $\mathbb{R}^2$ módulo homotecia, pág. 60 |
| $\Delta_G$ | Función modular de un grupo $G$, pág. 14 |
| $\Gamma_q$ | Estabilizador de $q$ en $\mathrm{PGL}_2(\mathbb{Z})$, pág. 35 |
| $\mathscr{I}(\mathcal{O})$ | El grupo de ideales fraccionarios de $\mathcal{O}$, pág. 56 |
| $\mathrm{card}_\mu(\mathscr{P})$ | La cantidad de elementos $B \in \mathscr{P}$ con $\mu(B) > 0$ |
| $\mathrm{Pic}(R)$ | Grupo de Picard de un anillo $R$, pág. 56 |
| $\mathrm{Pic}^+(\mathcal{O})$ | Grupo de clases angostas de un orden $\mathcal{O}$ |
| $\mathrm{Rec}(x, B)$ | $= \{n \in \mathbb{Z} : T^n x \in B\}$ |
| $\mathrm{Reg}(\mathcal{O}_d)$ | El regulador de un orden $\mathcal{O}_d$, pág. 37 |
| $\mathscr{B}^T$ | La $\sigma$-subálgebra de conjuntos esencialmente $T$-invariantes en $\mathscr{B}$ |
| $\mathscr{G}_d$ | Soporte de la medida $\mu_d$ |
| $\mathscr{L}_2^{(1)}(\mathbb{R})$ | El espacio de retículos en $\mathbb{R}^2$ de covolumen 1, pág. 60 |
| $\mathscr{M}(X)$ | Espacio de medidas de Radon complejas, pág. 10 |
| $\mathscr{M}^+(X)$ | Conjunto de medidas de Radon en $X$, pág. 11 |
| $\mathscr{M}^1(X)$ | Conjunto de medidas de probabilidad de Radon en $X$, pág. 11 |
| $\mathscr{M}_H^+(X)$ | Conjunto de medidas de Radon $H$-invariantes a derecha, pág. 11 |
| $\mathcal{O}_d$ | $= \mathbb{Z}\left[\frac{d+\sqrt{d}}{2}\right]$ |
| $\mathcal{O}_K$ | Anillo de enteros de $K$, pág. 51 |
| $\mathcal{O}_{d,1}$ | Grupo de unidades de $\mathcal{O}_d$ de norma 1, pág. 59 |
| $\mathscr{P} \vee \mathscr{Q}$ | Supremo de dos particiones |
| $\mathscr{P}^+$ | Grupo de ideales fraccionarios principales en un orden $\mathcal{O}$ |
| $\mathscr{P}^{(n)}$ | $= \bigvee_{k=0}^{n-1} T^{-k}\mathscr{P}$ |
| $\mu_d$ | Ver pág. 35 |
| $\mu_n \Longrightarrow \mu$ | Convergencia débil de una sucesión de medida, pág. 11 |
| $\mu_n \xrightarrow{w*} \mu$ | Convergencia débil$^*$ de una sucesión de medidas, pág. 11 |

| | |
|---|---|
| $\rho_d$ | Ver pág. 35 |
| $\widetilde{\Gamma}_q$ | El estabilizador de $x_q$ en $A$, pág. 35 |
| $A_N f$ | $= \frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n$, ver (1.5) |
| $B_r^G$ | Bola de radio $r$, centrada en la identidad, pág. 45 |
| $C(d)$ | Cociente de $R_{\text{disc}}(d)$ por la acción de $\text{PGL}_2(\mathbb{Z})$, pág. 35 |
| $d_K$ | Discriminante de un cuerpo de números $K$, pág. 52 |
| $F$ | Dominio fundamental de la curva modular, ver (2.14) |
| $f_*\mu$ | Medida push-forward, pág. 9 |
| $g_q$ | Punto en $\text{PGL}_2(\mathbb{R})$ asociado a la forma $q$, pág. 35 |
| $H_\mu(\mathscr{P})$ | Entropía de una partición $\mathscr{P}$, ver (1.14) |
| $h_\mu(T)$ | Entropía de una transformación medible $T$, ver (1.21) |
| $h_\mu(T, \mathscr{P})$ | $= \lim_{n\to\infty} \frac{1}{n} H(\mathscr{P}^{(n)})$, ver (1.20) |
| $I_H f(x)$ | $= \int_H f(xh) d\lambda_H(h)$, ver (1.3) |
| $I_q$ | Ideal asociado a una forma cuadrática binaria, ver (3.4) |
| $M_q$ | Matriz asociada a una forma cuadrática binaria |
| $R_{\text{disc}}(d)$ | El conjunto de formas cuadráticas binarias enteras primitivas de discriminante $d$ |
| $R_d(n)$ | Cantidad de ideales inversibles de $\mathcal{O}_d$ con norma $n$, pág. 59 |
| $S$ | El dominio fundamental para la acción de $\text{PSL}_2(\mathbb{Z})$ en $\text{PSL}_2(\mathbb{R})$, ver (2.16) |
| $S_v$ | $= \{z \in \Lambda : \langle v, z\rangle \equiv 0 \pmod{p}\}$, ver pág. 68 |
| $V_{\text{disc}}^+(\mathbb{R})$ | Espacio de formas cuadráticas binarias reales de discriminante 1, ver (2.1) |
| $x_q$ | $= \Gamma g_q$, pág. 35 |
| $X_{\geq H}, X_{\leq H}$ | Ver pág. 61 |

*A Florencia*

# Capítulo 1

# Teoría ergódica

## 1.1. Teoría de la medida

Esta sección sirve como repaso de los resultados de teoría de la medida que utilizaremos más adelante, así como para fijar notación.

Sea $X$ un conjunto. Un *álgebra* de subconjuntos de $X$ es una familia $\mathscr{A}$ de subconjuntos de $X$ que cumple: $(i)$ $\varnothing \in \mathscr{A}$; $(ii)$ si $A, B \in \mathscr{A}$ entonces $A \cup B \in \mathscr{A}$, y $(iii)$ si $A \in \mathscr{A}$ entonces $X \setminus A \in \mathscr{A}$. Una *$\sigma$-álgebra* es un álgebra $\mathscr{A}$ que es cerrada por uniones numerables de sus elementos, es decir, si $A_n \in \mathscr{A}$ para cada $n \in \mathbb{N}$, entonces $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$. Notemos que cualquier intersección de álgebras en el conjunto de partes $\mathscr{P}(X)$ es un álgebra, así que para toda colección $\mathscr{S} \subseteq \mathscr{P}(X)$ existe una única álgebra que es minimal con la propiedad de contener a $\mathscr{S}$. Llamamos a esta álgebra el *álgebra generada por $\mathscr{S}$*. El mismo argumento es válido para $\sigma$-álgebras, así que existe para $\mathscr{S} \subseteq \mathscr{P}(X)$ una única $\sigma$-álgebra $\sigma(\mathscr{S})$ que contiene a $\mathscr{S}$ y es minimal con esta propiedad, y que llamamos la *$\sigma$-álgebra generada por $\mathscr{S}$*. Si $X$ es un espacio topológico, su *$\sigma$-álgebra de Borel* es la $\sigma$-álgebra $\mathscr{B}_X$ generada por los abiertos (o equivalentemente, los cerrados) de $X$.

Un *espacio medible* consiste de un conjunto $X$ provisto de una $\sigma$-álgebra $\mathscr{B} \subseteq \mathscr{P}(X)$, y llamamos *conjuntos medibles* a los elementos de $\mathscr{B}$. Si $(X, \mathscr{B})$ es un espacio medible, una *medida* en este espacio es una función $\mu : \mathscr{B} \to [0, +\infty]$ tal que $\mu(\varnothing) = 0$ y $\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n)$ para toda sucesión $\{B_n\}_{n\in\mathbb{N}} \subseteq \mathscr{B}$ de conjuntos disjuntos dos a dos. Análogamente, una *medida signada finita* (resp. *medida compleja*) es una función $\mu : \mathscr{B} \to \mathbb{R}$ (resp. $\mu : \mathscr{B} \to \mathbb{C}$) que es $\sigma$-aditiva. Decimos que una medida $\mu$ es una *medida de probabilidad* si además $\mu(X) = 1$. Decimos que $\mu$ es *finita* (resp. *$\sigma$-finita*) si $\mu(X) < +\infty$ (resp. $X$ es la unión de numerables conjuntos medibles de medida finita). Un *espacio de medida* es una terna $(X, \mathscr{B}, \mu)$ donde $(X, \mathscr{B})$ es un espacio medible y $\mu$ es una medida. Si $\mu(X) = 1$ decimos que $(X, \mathscr{B}, \mu)$ es un *espacio de probabilidad*.

Si $(X, \mathscr{B})$, $(Y, \mathscr{C})$ son espacios medibles, llamamos a una función $f : X \to Y$ *medible* si $f^{-1}(C) \in \mathscr{B}$ para todo conjunto medible $C \in \mathscr{C}$. En el caso que $Y = \mathbb{R}$ o $\mathbb{C}$, tomamos por convención la $\sigma$-álgebra de Borel. Entonces una función $f : (X, \mathscr{B}) \to \mathbb{R}$ es medible si y solo si $f^{-1}(c, +\infty) \in \mathscr{B}$ para todo $c \in \mathbb{R}$. Si $f : (X, \mathscr{B}) \to (Y, \mathscr{C})$ es medible y $\mu$ es una medida en $(X, \mathscr{B})$, entonces $f$ define una *medida push-forward* $f_*\mu$ en $(Y, \mathscr{C})$, dada por $f_*\mu(C) = \mu(f^{-1}(C))$ para todo $C \in \mathscr{C}$.

Un método posible para definir medidas en un espacio medible $(X, \mathscr{B})$ consiste en definir la medida $\mu$ en un subconjunto de $\mathscr{B}$, por ejemplo un álgebra $\mathscr{B}_0 \subseteq \mathscr{B}$ y después probar que $\mu$ se puede extender a $\mathscr{B}$. Si $\mathscr{B}_0$ es un álgebra de subconjuntos de $X$, una *pre-medida* en $\mathscr{B}_0$ es una función $\mu_0 : \mathscr{B}_0 \to [0, +\infty]$ tal que: $(i)$ $\mu_0(\varnothing) = 0$, y $(ii)$ si $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathscr{B}_0$ son conjuntos disjuntos dos a dos tales que $\bigcup_{n=1}^{\infty} A_n \in \mathscr{B}_0$, entonces $\mu_0(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_0(A_n)$. Notemos que una pre-medida es finitamente aditiva ya que podemos tomar $A_n = \varnothing$ en la definición para $n$ grande. Es claro que si $\mu_0$ se puede extender a una medida en una $\sigma$-álgebra más grande entonces debe cumplir estas dos propiedades. El siguiente teorema garantiza que $\mu_0$ se puede extender a la $\sigma$-álgebra generada por $\mathscr{B}_0$:

**Teorema 1.1** (Hahn-Kolmogorov)**.** *Sea $\mathscr{B}_0$ un álgebra de subconjuntos de $X$ y sea $\mathscr{B}$ la $\sigma$-álgebra generada por $\mathscr{B}_0$. Entonces toda pre-medida $\mu_0 : \mathscr{B}_0 \to [0, +\infty]$ se puede extender a una medida*

$\mu : \mathscr{B} \to [0, +\infty]$. *Si* $\mu$ *es* $\sigma$*-finita, entonces tal extensión es única y está dada por*

$$\mu(E) = \text{ínf} \left\{ \sum_{k=1}^{\infty} \mu_0(B_k) : B_k \in \mathscr{B}_0, E \subseteq \bigcup_{k=1}^{\infty} B_k \right\}. \tag{1.1}$$

*Demostración.* Ver [Fol99, Teo. 1.14]. □

**Teorema 1.2** (Teorema de aproximación)**.** *Sea* $(X, \mathscr{B}, \mu)$ *un espacio de medida y sea* $\mathscr{B}_0 \subseteq \mathscr{B}$ *un álgebra que genera a* $\mathscr{B}$ *como* $\sigma$*-álgebra. Entonces para todo* $B \in \mathscr{B}$ *tal que* $\mu(B) < \infty$ *y para todo* $\varepsilon > 0$ *existe* $B_0 \in \mathscr{B}_0$ *tal que* $\mu(B \triangle B_0) < \varepsilon$.

*Demostración.* Ver [Hal50, §12.B] □

Supongamos ahora que $(X, \mathscr{B}, \mu)$ es un espacio de medida donde $X$ es un espacio métrico separable y $\mathscr{B}$ es la $\sigma$-álgebra de Borel. Sea $N \subseteq X$ la unión de todos los abiertos $U \subseteq X$ con $\mu(U) = 0$. Como $X$ es separable, $N$ se puede expresar como la unión de a lo sumo numerables de estos abiertos, así que $\mu(N) = 0$ por $\sigma$-aditividad. Se sigue que $N$ es el mayor abierto con medida 0. El complemento de $N$ se denomina el *soporte* de $\mu$, notado $\text{sop}(\mu)$. El soporte es un subespacio cerrado y se puede caracterizar como el conjunto de puntos $x \in X$ tales que $\mu(U) > 0$ para todo entorno abierto $U$ de $x$.

### 1.1.1. Espacios de Lebesgue

Dado $p \in [1, +\infty)$ definimos $\mathscr{L}^p_\mu(X)$ como el conjunto de todas las funciones $f : X \to \mathbb{C}$ medibles tales que $\int |f|^p \, d\mu < \infty$. Por la desigualdad de Minkowski, $\|f\|_p = \left(\int |f|^p \, d\mu\right)^{1/p}$ es una seminorma en este espacio. Su núcleo es el conjunto $\mathscr{N}$ de funciones medibles que son iguales a 0 c.t.p. Luego el cociente $L^p_\mu(X) = \mathscr{L}^p_\mu(X)/\mathscr{N}$ es un espacio normado que es completo (ver [Fol99, Teo. 6.6]). También usamos la notación alternativa $L^p(X, \mathscr{B}, \mu)$ para especificar la $\sigma$-álgebra $\mathscr{B}$. Para $p = \infty$, $\mathscr{L}^\infty_\mu(X)$ es el conjunto de funciones esencialmente acotadas, es decir, aquellas funciones para las cuales

$$\|f\|_\infty = \text{ínf}\{c \in \mathbb{R} : |f| \leq c \text{ para } \mu\text{-casi todo } x \in X\} < \infty.$$

Al igual que antes, $\|f\|_\infty$ es una norma en el cociente $L^\infty_\mu(X) = \mathscr{L}^\infty_\mu(X)/\mathscr{N}$.

### 1.1.2. Convergencia de medidas

Sea $X$ un espacio localmente compacto y Hausdorff. Una *medida de Radon* en $X$ es una medida de Borel que cumple:

(*i*) $\mu(K) < \infty$ para todo compacto $K \subseteq X$;

(*ii*) $\mu(E) = \text{ínf}\{\mu(U) : U \supseteq E \text{ es abierto}\}$ para todo boreliano $E \in \mathscr{B}_X$;

(*iii*) $\mu(U) = \sup\{\mu(K) : K \subseteq U \text{ es compacto}\}$ para todo abierto $U \subseteq X$.

Denotamos por $\mathscr{M}(X)$ al espacio de medidas de Radon complejas. Se puede ver que cualquier medida de Borel finita es una medida de Radon.

**Teorema 1.3** (Riesz-Markov)**.** *Si* $\varphi : C_c(X) \to \mathbb{C}$ *es una transformación lineal positiva, entonces existe una única medida de Radon* $\mu$ *en* $X$ *tal que* $\varphi(f) = \int f \, d\mu$ *para toda* $f \in C_c(X)$.

*Demostración.* Ver [Fol99, Teo. 7.2]. □

**Corolario 1.4.** $C_c^*(X)$ *es isomorfo como* $\mathbb{C}$*-espacio vectorial a* $\mathscr{M}(X)$

Si $X$ es un espacio localmente compacto y Hausdorff, denotamos por $\mathscr{M}^+(X)$ (resp. $\mathscr{M}^1(X)$) al conjunto de medidas de Radon (resp. medidas de probabilidad de Radon). Si $H$ es un grupo que actúa a derecha en $X$, denotamos por $\mathscr{M}_H^+(X)$ al conjunto de medidas de Radon $H$-invariantes a derecha. Se le puede asignar a todos estos conjuntos una *topología débil** que es la topología inicial con respecto a la familia de mapas $\mu \mapsto \mu(f) = \int f\, d\mu$ donde $f$ varía en $C_c(X)$. En otras palabras, es la topología con sub-base formada por los conjuntos

$$U(\mu_0, f, \varepsilon) = \{\mu \in \mathscr{M}(X) : |\mu(f) - \mu_0(f)| < \varepsilon\},$$

donde $\varepsilon > 0$, $\mu_0 \in \mathscr{M}(X)$ y $f \in C_c(X)$. Bajo el isomorfismo $\mathscr{M}(X) \cong C_c(X)^*$, la topología débil* se corresponde con la topología débil* en $C_c(X)^*$. A su vez podemos considerar la topología inicial en $\mathscr{M}(X)$ con respecto a los mapas $\mu \mapsto \mu(f)$ con $f$ variando en $C_b(X)$, el espacio de funciones continuas acotadas. Esta topología es más fina que la débil* y se suele llamar la *topología débil*. Usamos la notación $\mu_n \xrightarrow{w*} \mu$ para denotar convergencia débil* y la notación $\mu_n \Longrightarrow \mu$ para denotar convergencia débil.

**Lema 1.5.** *Sea $(X, d)$ un espacio métrico y sean $\mu_n, \mu$ medidas de probabilidad en $(X, \mathscr{B})$. Son equivalentes:*

*(i)* $\mu_n \Longrightarrow \mu$.

*(ii)* $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ *para todo cerrado* $F \subseteq X$.

*(iii)* $\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ *para todo abierto* $U \subseteq X$.

*(iv)* $\mu_n(B) \to \mu(B)$ *para todo* $B \in \mathscr{B}$ *tal que* $\mu(\partial B) = 0$.

*Demostración.* Ver [Bil99, Teo. 2.1] □

**Teorema 1.6.** *Sea $X$ un espacio métrico separable y localmente compacto. Entonces la topología débil* y la topología débil en $\mathscr{M}^1(X)$ son metrizables.*

*Demostración.* El espacio normado $(C_c(X), \|\cdot\|_{\sup})$ es separable así que la bola unitaria cerrada $\overline{B}$ de $C_c(X)^*$ es metrizable con la topología débil*. Por el teorema de Riesz-Markov el espacio $\mathscr{M}^1(X)$ con la topología débil* es homeomorfo a un subespacio de $\overline{B}$ y en particular es metrizable. Para la topología débil el teorema se deduce de que esta topología está inducida por una distancia conocida como la *métrica de Prohorov*, cf. [Bil99, §6]. □

Si $X$ es un espacio métrico compacto, entonces $\mathscr{M}^1(X)$ es compacto ya que bajo el isomorfismo $\mathscr{M}(X) \cong C(X)^*$, $\mathscr{M}^1(X)$ se corresponde con un subespacio cerrado de la bola unitaria cerrada, que es débil*-compacta por Banach-Alaoglu. Sin embargo esto no vale cuando $X$ no es compacto y puede suceder que una sucesión de medidas de probabilidad converja a una medida $\mu \in \mathscr{M}(X)$ con $\mu(X) < 1$. Por ejemplo, si tomamos las medidas de Dirac $\delta_n$ en $X = \mathbb{R}$, entonces $\delta_n \xrightarrow{w*} 0$ ya que $\delta_n(f) = f(n)$ es eventualmente cero para cualquier $f \in C_c(\mathbb{R})$. Este fenómeno se suele llamar *escape de masa*. Notar además que $\delta_n$ no converge débilmente a 0.

No obstante, hay una condición sencilla que garantiza la compacidad de subespacios de $\mathscr{M}^1(X)$. Sea $X$ un espacio métrico separable y completo. Decimos que un conjunto $\Pi \subseteq \mathscr{M}^1(X)$ es *tight* si para todo $\varepsilon > 0$ existe un compacto $K \subseteq X$ tal que $\mu(K) > 1 - \varepsilon$ para todo $\mu \in \Pi$. Decimos que una sucesión $(\mu_n)_n \subseteq \mathscr{M}^1(X)$ es *tight* si para todo $\varepsilon > 0$ existe $n_0 \in \mathbb{N}$ y un compacto $K$ tal que $\mu_n(K) > 1 - \varepsilon$ para todo $n \geq n_0$. Observemos que esto es equivalente a que el conjunto $\{\mu_n : n \in \mathbb{N}\}$ sea tight. En efecto, es fácil ver que $\{\mu\}$ es tight para cualquier medida $\mu \in \mathscr{M}^1(X)$ (cf. [Bil99, Teo. 1.3]). De esto se deduce que cualquier conjunto $\Pi \subseteq \mathscr{M}^1(X)$ finito es tight. Si $(\mu_n)_n$ es una sucesión tight, dado $\varepsilon > 0$ existe $n_0 \in \mathbb{N}$ y $K_1 \subseteq X$ compacto tal que $\mu_n(K_1) > 1 - \varepsilon$ para todo $n \geq n_0$. Como $\{\mu_1, \ldots, \mu_{n_0 - 1}\}$ es finto, existe un compacto $K_2$ tal que $\mu_n(K_2) > 1 - \varepsilon$ para todo $n < n_0$. Entonces $\mu_n(K_1 \cup K_2) > 1 - \varepsilon$ para todo $n$. La implicación inversa es evidente.

**Teorema 1.7** (Prohorov)**.** *Si $X$ es un espacio métrico separable, entonces toda sucesión tight $(\mu_n)_n \subseteq \mathscr{M}^1(X)$ posee una subsucesión que converge débilmente.*

*Demostración.* Ver [Bil99, §5]. □

**Proposición 1.8.** *Si $X$ es un espacio métrico localmente compacto y separable, entonces la topología débil* y la débil coinciden en $\mathscr{M}^1(X)$.*

*Demostración.* Como ambas topologías son metrizables, basta ver que tienen las mismas sucesiones convergentes. Es claro que la convergencia débil implica la convergencia débil*. Recíprocamente, supongamos que $(\mu_n)_n$ es una sucesión que converge débil* a $\mu \in \mathscr{M}^1(X)$. Veamos primero que $(\mu_n)_n$ es tight. Si esto no pasa, entonces existe $\eta > 0$ tal que $\liminf_{n\to\infty} \mu_n(K) \leq 1-\eta$ para todo compacto $K$. Por otro lado, como $\{\mu\}$ es tight, existe un compacto $K_1$ tal que $\mu(K_1) > 1-\eta/2$. Sea $f : X \to [0,1]$ una función continua de soporte compacto tal que $f|_{K_1} \equiv 1$. Entonces $\int f\,d\mu \geq \mu(K_1) > 1-\eta/2$. Por otro lado el soporte $K_2$ de $f$ es compacto, así que

$$\liminf_{n\to\infty} \int f\,d\mu_n \leq \liminf_{n\to\infty} \mu_n(K_2) \leq 1-\eta.$$

Pero esto se contradice con que $\mu_n(f) \to \mu(f)$. Por lo tanto $(\mu_n)_n$ es tight.

Como la sucesión es tight, para todo $j \in \mathbb{N}$ existe un compacto $C_j$ tal que $\mu_n(C_j) > 1-1/j$ para todo $n$. Como $X$ es separable y localmente compacto, se puede escribir como unión numerable de bolas abiertas precompactas $B_n$, $n \in \mathbb{N}$. Construimos inductivamente una sucesión de compactos $K_j \supseteq C_j$ tales que $K_j \subseteq K_{j+1}^\circ$ para todo $j$ y $X = \bigcup_{j=1}^\infty K_j$. En primer lugar tomamos $K_1 = C_1$. Supongamos que hemos definido $K_j$ para algún $j \geq 1$. Por compacidad existe $n_j > 0$ tal que

$$K_j \subseteq B_1 \cup \cdots \cup B_{n_j} = W_j.$$

Notar que $\overline{W_j} = \overline{B_1} \cup \cdots \cup \overline{B_{n_j}}$ es compacto. Entonces podemos tomar $K_{j+1} = C_{j+1} \cup \overline{W_j}$. Por el lema de Urysohn existe una función continua $\phi_j : X \to [0,1]$ tal que $\phi_j \equiv 1$ en $K_j$ y $\phi_j \equiv 0$ en $X \setminus K_{j+1}^\circ$. Dada una función $f \in C_b(X)$, definimos $f_j = \phi_j f \in C_c(X)$. Notemos que $f = f_j$ en $K_j$, así que

$$\int_X |f - f_j|\,d\mu_n = \int_{X\setminus K_j} |f|(1-\phi_j)\,d\mu_n \leq \mu(X \setminus K_j)\|f\|_{\sup} \leq \frac{\|f\|_{\sup}}{j}$$

para todo $n \in \mathbb{N}$. Ahora veamos que $\mu_n(f) \to \mu(f)$. Tenemos que

$$\begin{aligned} |\mu(f) - \mu_n(f)| &\leq |\mu(f) - \mu(f_j)| + |\mu(f_j) - \mu_n(f_j)| + |\mu_n(f_j) - \mu_n(f)| \\ &\leq |\mu(f_j) - \mu_n(f_j)| + \frac{2\|f\|_{\sup}}{j}. \end{aligned}$$

Haciendo $n \to \infty$ obtenemos que $\limsup_{n\to\infty} |\mu(f) - \mu_n(f)| \leq \frac{2\|f\|_{\sup}}{j}$. Esto vale para todo $j$, así que $\mu_n(f) \to \mu(f)$. Por lo tanto $\mu_n \Longrightarrow \mu$. □

### 1.1.3. Medida producto

Si $\{(X_i, \mathscr{B}_i)\}_{i\in I}$ es una familia de espacios medibles, definimos la *$\sigma$-álgebra producto* como la $\sigma$-álgebra $\mathscr{B} = \bigotimes_{i\in I} \mathscr{B}_i$ en $X = \prod_{i\in I} X_i$ generada por los conjuntos $A_i \times \prod_{j\neq i} X_j$, donde $A_i \in \mathscr{B}_i$. Llamamos *rectángulo medible* a todo conjunto de la forma

$$A = \prod_{j\in F} A_j \times \prod_{i\in I\setminus F} X_i,$$

donde $F \subseteq I$ es un conjunto finito y $A_j \in \mathscr{B}_j$ para todo $j \in F$. Es claro que el álgebra generada por los conjuntos $A_i \times \prod_{j\neq i} X_j$ considerados antes consiste de aquellos conjuntos que son unión finita de rectángulos medibles. Además, si $I$ es infinito definimos un *conjunto cilíndrico* como un conjunto de la forma $A = A_F \times \prod_{i\in I\setminus F} X_i$ donde $F \subseteq I$ es finito y $A_F \in \bigotimes_{j\in F} \mathscr{B}_j$.

Ahora procedemos a definir la medida producto. Veamos primero el caso finito. Sean $(X, \mathscr{B}, \mu)$, $(Y, \mathscr{C}, \nu)$ espacios de medida $\sigma$-finitos y consideremos el álgebra $\mathscr{A}$ generada por los rectángulos medibles en $X \times Y$. Definimos una función $\tau_0 : \mathscr{A} \to [0, +\infty]$ de la siguiente manera: si $A \in \mathscr{A}$ es unión disjunta de numerables rectángulos medibles $B_n \times C_n$, $n \in \mathbb{N}$, entonces

$$\tau_0(A) := \sum_{n=1}^\infty \mu(B_n)\nu(C_n).$$

Es rutinario ver que esta función está bien definida y es una pre-medida en $\mathscr{A}$. Luego por el teorema 1.1, $\tau_0$ se extiende de manera única a una medida $\tau$ en $\mathscr{B} \otimes \mathscr{C}$. Llamamos a $\tau$ la *medida producto* y la notamos $\tau = \mu \times \nu$. Podemos extender esta construcción a una medida producto en un espacio producto con finitos factores definiendo $\mu_1 \times \mu_2 \times \mu_3 = (\mu_1 \times \mu_2) \times \mu_3$, etc. Si los factores son $\sigma$-finitos entonces la medida resultante no depende del orden de los factores. En el caso de numerables factores solo consideramos espacios de probabilidad.

**Teorema 1.9.** *Si $\{(X_n, \mathscr{B}_n, \mu_n)\}_{n\in\mathbb{N}}$ una sucesión de espacios de probabilidad, entonces existe una única medida $\mu$ en la $\sigma$-álgebra $\mathscr{B} = \bigotimes_{n=1}^{\infty} \mathscr{B}_n$ tal que para todo conjunto cilíndrico de la forma $E = A \times \prod_{m>n} X_m$, con $A \in \mathscr{B}_1 \otimes \cdots \otimes \mathscr{B}_n$, se tiene*

$$\mu(E) = (\mu_1 \times \cdots \times \mu_n)(A).$$

*Demostración.* Ver [Hal50, Teo. 38.B] □

### 1.1.4. Esperanza condicional

Sea $\mu$ una medida en $(X, \mathscr{B})$ y $\nu$ una medida signada. Decimos que $\nu$ es *absolutamente continua* con respecto a $\mu$ (notado $\nu \ll \mu$) si $\nu(E) = 0$ para todo $E \in \mathscr{B}$ tal que $\mu(E) = 0$.

**Teorema 1.10** (Radon-Nikodym)**.** *Sean $\mu, \nu$ medidas signadas en un espacio medible $(X, \mathscr{B})$ y supongamos que $\mu$ es positiva. Si $\nu \ll \mu$ entonces existe $f \in L^1_\mu(X)$ tal que $\nu(B) = \int_B f\, d\mu$ para todo $B \in \mathscr{B}$. Cualquier otra función con estas propiedades es igual a $f$ $\mu$-c.t.p. Más aún, para toda $g \in L^1_\nu(X)$ se tiene que $gf \in L^1_\mu(X)$ y $\int g\, d\nu = \int gf\, d\mu$.*

*Demostración.* Ver [Fol99, §3.2]. □

El teorema de Radon-Nikodym nos permite definir la esperanza condicional de una función integrable con respecto a una $\sigma$-álgebra $\mathscr{F} \subseteq \mathscr{B}$. Si $f \in L^1_\mu(X)$, tomando parte real e imaginaria podemos suponer que $f$ toma valores reales. Entonces la fórmula $\nu(B) = \int_B f\, d\mu$ define una medida signada en $(X, \mathscr{B})$, que podemos restringir a $\mathscr{F}$. Como $\nu|_{\mathscr{F}} \ll \mu|_{\mathscr{F}}$, existe una única $g \in L^1(X, \mathscr{F}, \mu|_{\mathscr{F}})$ tal que $\int_A f\, d\mu = \int_A g\, d\mu$ para todo $A \in \mathscr{F}$. Llamamos a $g = \mathbb{E}(f|\mathscr{F})$ la *esperanza condicional* de $f$ con respecto a $\mathscr{F}$. Por linealidad podemos extender la definición a funciones a valores complejos.

Si $X$ es un espacio de probabilidad, entonces toda función $f \in L^2_\mu(X)$ pertenece a $L^1_\mu(X)$ ya que $\int |f|\, d\mu \leq \frac{1}{2}\int |f|^2\, d\mu + \frac{1}{2}$. Luego podemos mirar a $L^2_\mu(X)$ como un subespacio de $L^1_\mu(X)$, aunque con distinta norma.

**Proposición 1.11.** *Sea $(X, \mathscr{B}, \mu)$ un espacio de probabilidad y sea $\mathscr{F} \subseteq \mathscr{B}$ una $\sigma$-álgebra. Entonces la restricción del operador esperanza condicional $\mathbb{E}(-|\mathscr{F})$ a $L^2(X, \mathscr{B}, \mu)$ es igual a la proyección ortogonal $L^2(X, \mathscr{B}, \mu) \to L^2(X, \mathscr{F}, \mu)$.*

*Demostración.* Ver [Wal00, Teo. 0.12] □

Finalmente notemos que si $\mu(X) < \infty$ y $f : X \to \mathbb{R}$ es acotada, entonces $\|\mathbb{E}(f|\mathscr{F})\|_\infty \leq \|f\|_\infty$. En efecto, sea $M = \|f\|_\infty$ y sea $B = \{x : \mathbb{E}(f|\mathscr{F})(x) > M\} \in \mathscr{F}$. Si $\mu(B) > 0$ entonces

$$M\mu(B) < \int_B \mathbb{E}(f|\mathscr{F})\, d\mu = \int_B f\, d\mu \leq M\mu(B),$$

lo cual es absurdo. Por lo tanto $\mu(B) = 0$. Análogamente se demuestra que $\mathbb{E}(f|\mathscr{F}) \geq -M$ c.t.p. así que $\|\mathbb{E}(f|\mathscr{F})\|_\infty \leq M$.

### 1.1.5. Medida de Haar

Sea $G$ un grupo localmente compacto y Hausdorff. Una *medida de Haar a izquierda* (resp. a derecha) en $G$ es una medida de Radón $\mu$ que es invariante por traslaciones a izquierda (resp. a derecha), es decir, tal que $\mu(gA) = \mu(A)$ para todo $g \in G$ y para todo boreliano $A \subseteq G$. Si denotamos por $L_g$ al mapa $x \mapsto gx$, notemos que $\mu$ es invariante a izquierda si y solo si $(L_g)_*\mu = \mu$ para todo $g \in G$.

**Teorema 1.12** (Haar)**.** *Sea $G$ un grupo localmente compacto y Hausdorff. Entonces $G$ admite una medida de Haar a izquierda. Si $\lambda$ y $\mu$ son medidas de Haar a izquierda en $G$, entonces existe una constante $c > 0$ tal que $\mu = c\lambda$.*

*Demostración.* Ver [Fol15, §2.2]. □

Para varios grupos de Lie clásicos es sencillo encontrar su medida de Haar. Esto se debe al siguiente resultado:

**Proposición 1.13.** *Sea $G$ un grupo de Lie real o complejo cuya variedad subyacente es un abierto de $\mathbb{R}^N$ y tal que las traslaciones a izquierda son transformaciones afines, i.e. $xy = A_x y + b_x$ para todos $x, y \in G$, donde $A_x \in M_N(\mathbb{R})$ y $b_x \in \mathbb{R}^N$. Entonces $|\det(A_x)|^{-1}\, dx$ es una medida de Haar en $G$, donde $dx$ denota la medida de Lebesgue en $\mathbb{R}^N$.*

*Demostración.* Ver [Fol15, Teo. 2.21]. □

**Ejemplo 1.14.** Por ejemplo, $\mathrm{GL}_n(\mathbb{R})$ es un abierto de $\mathbb{R}^{n^2}$ y los mapas $y \mapsto xy$, $y \mapsto yx$ son $\mathbb{R}$-lineales para todo $x \in \mathrm{GL}_n(\mathbb{R})$. Usando la proposición anterior es fácil ver que la medida $|\det A|^{-n} dA$ es una medida de Haar a izquierda y a derecha en $\mathrm{GL}_n(\mathbb{R})$.

En el caso general de un grupo de Lie $G$ de dimensión real $n$, se puede construir una medida de Haar a izquierda usando una forma de volumen invariante a izquierda. Si $\omega_1, \dots, \omega_n$ son una base de 1-formas invariantes a izquierda, entonces $\omega = \omega_1 \wedge \cdots \wedge \omega_n$ es una $n$-forma invariante a izquierda que es nunca nula. En particular, $\omega$ determina una orientación en $G$. Además $\omega$ nos permite definir una funcional positiva $I$ en $C_c(G)$ dada por $I(f) = \int_G f\omega$. Por Riesz-Markov, existe una única medida de Radon $\mu$ tal que $I(f) = \int f\, d\mu$. Como $L_g^*\omega = \omega$ para todo $g \in G$, $\mu$ es una medida $G$-invariante a izquierda, ya que para toda $f \in C_c(G)$,

$$\begin{aligned}\int f\, d(L_g)_*\mu &= \int (f \circ L_g)\, d\mu = \int_G (f \circ L_g)\omega \\ &= \int_G (f \circ L_g) L_g^*\omega = \int_G L_g^*(f\omega) = \int_G f\omega = \int f\, d\mu.\end{aligned}$$

La anteúltima igualdad se debe a que $L_g$ preserva la orientación de $G$. Se sigue que $(L_g)_*\mu = \mu$ para todo $g \in G$, así que $\mu$ es invariante.

Sea $R_g : G \to G$ el mapa $x \mapsto xg$. Notemos que si $\mu$ es una medida de Haar a izquierda en $G$ entonces $(R_g)_*\mu$ también lo es. En efecto, como $R_g$ y $L_h$ conmutan para cualesquiera $g, h \in G$, tenemos

$$(L_h)_*(R_g)_*\mu = (R_g)_*(L_h)_*\mu = (R_g)_*\mu.$$

Por el teorema 1.12 existe una constante $\Delta_G(g^{-1}) > 0$ tal que $(R_g)_*\mu = \Delta_G(g^{-1})\mu$. La función $\Delta_G : G \to \mathbb{R}_{>0}$ es un morfismo de grupos topológicos y se denomina la *función modular* de $G$. Decimos que $G$ es *unimodular* si $\Delta_G \equiv 1$. Un ejemplo de grupo unimodular es $\mathrm{GL}_n(\mathbb{R})$, por conscuencia de la proposición 1.13.

Si $G$ es un grupo topológico que actúa por homeomorfismos en un espacio Hausdorff $X$, decimos que $X$ es un *espacio homogéneo* si para todo $x \in X$, el mapa $\pi_x : G \to X$, $\pi_x(g) = gx$ es continuo, abierto y sobreyectivo. Es fácil ver que bajo estas condiciones el estabilizador $H = \mathrm{Stab}_G(x)$ de cualquier punto es un subgrupo cerrado y que el mapa $G/H \to X$, $g \mapsto gx$ es un homeomorfismo que conmuta con la acción de $G$. En el caso de espacios homogéneos no siempre es posible definir una medida de Radón en $X$ que sea $G$-invariante.

**Teorema 1.15.** *Sea $G$ un grupo localmente compacto y Hausdorff y sea $H \leq G$ un subgrupo cerrado. Una condición necesaria y suficiente para que exista una medida de Radon $G$-invariante a izquierda $\mu$ en $G/H$ es que $\Delta_G|_H = \Delta_H$. En ese caso $\mu$ es única salvo por factor de proporcionalidad.*

*Demostración.* Ver [Nac65, §III.4]. □

Por ejemplo se dan las condiciones del teorema si $G$ y $H$ son grupos unimodulares. Esta medida también se llama medida de Haar.

**Lema 1.16.** *Sea $G$ un grupo localmente compacto y Hausdorff con una medida de Haar a izquierda $\lambda$. Entonces*

$$\int_G f(g)\,d\lambda(g) = \int_G f(g^{-1})\Delta_G(g^{-1})\,d\lambda(g). \tag{1.2}$$

*En particular, si $G$ es unimodular, entonces $\int_G f(g)\,d\lambda(g) = \int_G f(g^{-1})\,d\lambda(g)$.*

*Demostración.* Sea $\delta : G \to \mathbb{R}_{>0}$, $\delta(x) = \Delta(x^{-1})$ y sea $\nu$ la medida de Radon en $G$ definida por $\nu(f) = \int f(g^{-1})\delta(g)\,d\lambda(g)$ para toda $f \in C_c(G)$. Esta medida es $G$-invariante a izquierda ya que para todo $a \in G$,

$$\begin{aligned}\nu(f \circ L_a) &= \int f(ag^{-1})\delta(g)\,d\lambda(g) = \delta(a)\int f((ga^{-1})^{-1})\delta(ga^{-1})\,d\lambda(g)\\ &= \delta(a)\Delta(a)\int f(g^{-1})\delta(g)\,d\lambda(g) = \nu(f).\end{aligned}$$

Se sigue que $\nu = c\lambda$ para alguna constante $c > 0$. Para ver que $c = 1$, dado $\varepsilon > 0$, por continuidad de $\delta$ podemos tomar un entorno abierto precompacto $V$ de la identidad tal que $V = V^{-1}$ y $1 - \varepsilon < \delta(g) < 1 + \varepsilon$ para todo $g \in V$. En ese caso, tenemos $(1-\varepsilon)\lambda(V) < \nu(V) < (1+\varepsilon)\lambda(V)$. Como $\lambda(V) > 0$, esto implica que $|c - 1| < \varepsilon$. Por lo tanto $c = 1$. □

Sea $G$ un grupo localmente compacto y Hausdorff, y sea $H \leq G$ un subgrupo cerrado. Fijemos una medida de Haar a izquierda $\lambda_H$ en $H$. Definimos un mapa $I = I_H : C_c(G) \to C_c(G/H)$ como

$$If(gH) = \int_H f(gh)\,d\lambda_H(h). \tag{1.3}$$

La función $If$ está bien definida ya que si $yH = xH$, entonces $y = xh_0$ para algún $h_0 \in H$, lo que implica que

$$\int_H f(yh)\,d\lambda_H(h) = \int_H f(xh_0h)\,d\lambda_H(h) = \int_H f(xh)\,d\lambda_H(h)$$

por ser $\lambda_H$ invariante a izquierda. Claramente $If$ es continua y está soportada en $\pi(E)$ donde $E$ es el soporte de $f$ y $\pi : G \to G/H$ es la proyección canónica.

**Proposición 1.17.** *$I : C_c(G) \to C_c(G/H)$ es sobreyectiva.*

*Demostración.* Ver [Rag72, Lema 1.1]. □

**Lema 1.18.** *Sea $G$ un grupo localmente compacto y Hausdorff, y sea $H \leq G$ un subgrupo cerrado unimodular. Supongamos que $\lambda_H$ es una medida de Haar en $H$. Entonces hay una biyección $\alpha : \mathscr{M}^+(G/H) \to \mathscr{M}^+_H(G)$, $\nu \mapsto \mu$, donde $\mu$ es la única medida en $G$ tal que*

$$\int_G f\,d\mu = \int_{G/H} If(gH)\,d\nu(gH), \tag{1.4}$$

*para toda $f \in C_c(G)$.*

**Observación 1.19.** Este lema y la proposición 1.20 están sacados de [BO07, §8]. Sin embargo, los autores omitieron agregar que $H$ sea unimodular y en tal caso el enunciado del lema anterior es falso ya que junto con el teorema 1.15 implicaría que todo subgrupo de un grupo unimodular es unimodular. En efecto, si $G$ es unimodular, en particular su medida de Haar es $H$-invariante a derecha y luego induciría una medida en $G/H$ que es $G$-invariante a izquierda, es decir una medida de Haar. Por el teorema 1.15 esto implica que $\Delta_H = \Delta_G|_H$ así que $H$ sería unimodular. Un contraejemplo de esta afirmación es $G = \mathrm{GL}_2^+(\mathbb{R}) = \{g \in \mathrm{GL}_2(\mathbb{R}) : \det(g) > 0\}$ y $H = \left\{\left(\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right) : a > 0\right\}$.

*Demostración.* Si $\nu$ es una medida en $G/H$, entonces la medida $\mu$ definida por (1.4) es $H$-invariante a derecha ya que para todo $y \in H$, es

$$I(f \circ R_y)(gH) = \int_H f(ghy)\,d\lambda_H(h) = \int_H f(gh)\,d\lambda_H(h) = If(gH)$$

por ser $H$ unimodular. Vamos a construir la inversa de $\alpha$.

Dada $\mu \in \mathscr{M}_H^+(G)$, su preimagen por $\alpha$ es una medida $\nu$ tal que $\nu(If) = \mu(f)$ para toda $f \in C_c(G)$. Como $I : C_c(G) \to C_c(G/H)$ es sobreyectiva, $\nu$ está completamente determinada por esta condición. Para ver que $\nu$ está bien definida, supongamos que $f \in C_c(G)$ es tal que $If = 0$ y veamos que $\mu(f) = 0$. Si $\varphi \in C_c(G)$, entonces la función $F(g,h) = \varphi(g)f(gh) \in C(G \times H)$ tiene soporte compacto y en particular es integrable con respecto a $\mu \times \lambda_H$. Por Fubini, tenemos que

$$\begin{aligned}
0 &= \int_G \varphi(g) \left( \int_H f(gh)\, d\lambda_H(h) \right) d\mu(g) \\
&= \int_H \left( \int_G \varphi(g) f(gh)\, d\mu(g) \right) d\lambda_H(h) \\
&= \int_H \left( \int_G \varphi(gh^{-1}) f(g)\, d\mu(g) \right) d\lambda_H(h) \\
&= \int_G f(g) \left( \int_H \varphi(gh^{-1})\, d\lambda_H(h) \right) d\mu(g) \\
&= \int_G f(g) \left( \int_H \varphi(gh)\, d\lambda_H(h) \right) d\mu(g) = \int_G f(g) I\varphi(gH)\, d\mu(g).
\end{aligned}$$

Como $I$ es sobreyectiva, podemos elegir $\varphi \in C_c(G)$ tal que $I\varphi|_{\pi(E)} \equiv 1$, donde $E$ es el soporte de $f$. En ese caso, $I\varphi \circ \pi$ es igual a 1 en $\pi^{-1}\pi(E)$, así que $(I\varphi \circ \pi)f = f$. Por lo tanto $\int_G f\, d\mu = 0$, lo que implica que $\nu$ está bien definida. Es fácil ver que $\alpha$ y su inversa son continuas. □

**Proposición 1.20.** *Sea $G$ un grupo localmente compacto y Hausdorff, y sean $\Gamma, H \leq G$ subgrupos cerrados unimodulares. Fijemos medidas de Haar $\lambda_H$ y $\lambda_\Gamma$ en $H$ y $\Gamma$, respectivamente. Entonces hay homeomorfismos de los siguientes espacios provistos de la topología débil-*:*

$$\begin{array}{ccccccc}
\left\{ \begin{array}{c} \text{medidas de Radon} \\ H\text{-invariantes a} \\ \text{derecha en } \Gamma\backslash G \end{array} \right\} & \leftrightarrow & \left\{ \begin{array}{c} \text{medidas de Radon en } G, \\ H\text{-invariantes a derecha y} \\ \Gamma\text{-invariantes a izquierda} \end{array} \right\} & \leftrightarrow & \left\{ \begin{array}{c} \text{medidas de Radon} \\ \Gamma\text{-invariantes a} \\ \text{izquierda en } G/H \end{array} \right\} \\
\rho & \leftrightarrow & \mu & \leftrightarrow & \nu
\end{array}$$

*caracterizados por la siguiente propiedad: para toda $f \in C_c(G)$, se tiene*

$$\int_{\Gamma\backslash G} I_\Gamma f(\Gamma g)\, d\rho(\Gamma g) = \int_G f\, d\mu = \int_{G/H} I_H f(gH)\, d\nu(gH),$$

*donde $I_\Gamma f(\Gamma g) = \int_\Gamma f(\gamma g)\, d\lambda_\Gamma(\gamma)$ y $I_H f(gH) = \int_H f(gh)\, d\lambda_H(h)$.*

*Demostración.* Se deduce de aplicar el lema anterior dos veces, notando que bajo la correspondencia de este lema, la medidas de Radon en $G/H$ que son $\Gamma$-invariantes a izquierda se corresponden con medidas en $\mathscr{M}_H^+(G)$ que son $\Gamma$-invariantes a izquierda. □

## 1.2. Transformaciones invariantes

**Definición 1.21.** Sea $(X, \mathscr{B}, \mu)$ un espacio de medida y sea $T : X \to X$ una función medible.

(1) Decimos que un conjunto $A \in \mathscr{B}$ es *$T$-invariante* si $T^{-1}(A) = A$.

(2) Decimos que $\mu$ es *$T$-invariante* si $\mu(T^{-1}A) = \mu(A)$ para todo $A \in \mathscr{B}$, i.e. si $T_*\mu = \mu$. En ese caso también decimos que $T$ *preserva* a $\mu$ o que *preserva la medida*.

(3) Llamamos *sistema dinámico medible* a una tupla $(X, \mathscr{B}, \mu, T)$ donde $(X, \mathscr{B}, \mu)$ es un espacio de medida y $T : X \to X$ es medible y preserva a $\mu$. Además asumiremos que $\mu$ es una medida de probabilidad.

(4) Un *factor* de un sistema dinámico $(X, \mathscr{B}, \mu, T)$ es un sistema dinámico $(Y, \mathscr{D}, \nu, S)$ y una función medible $\phi : X \to Y$ tal que $\nu = \phi_*\mu$ y $\phi \circ T = S \circ \phi$.

Más adelante trabajaremos con acciones de grupo que preservan la medida. Si $G$ un grupo, una *acción* de $G$ en un espacio de medida $(X, \mathscr{B}, \mu)$ es un mapa $G \times X \to X$, $(g, x) \mapsto g.x$ tal que: $(i)$ $(gh).x = g.(h.x)$ para todos $g, h \in G$, $x \in X$, y $(ii)$ para todo $B \in \mathscr{B}$ y todo $g \in G$ tenemos que $\mu(B) = 0$ si y solo si $\mu(gB) = 0$. Si $G = \mathbb{R}$, la acción se denomina *flujo*. Decimos que $\mu$ es *$G$-invariante* si $\mu(gA) = \mu(A)$ para todo $A \in \mathscr{B}$ y para todo $g \in G$.

**Ejemplos.** (1) (Rotaciones del círculo) Sea $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ provisto de la topología cociente y de la medida de Lebesgue $\mu = \pi_*\lambda$ donde $\pi : [0,1] \to \mathbb{R}/\mathbb{Z}$ es la proyección al cociente y $\lambda$ es la medida de Lebesgue en $\mathbb{R}$. Entonces la transformación

$$R_\alpha : \mathbb{T} \to \mathbb{T}, \qquad R_\alpha(x) = x + [\alpha]$$

preserva la medida, para todo $\alpha \in \mathbb{R}$ (donde $[\alpha] = \pi(\alpha)$). Para ver esto, consideremos el diagrama conmutativo

$$\begin{array}{ccc} [0,1] & \xrightarrow{T_\alpha} & [0,1] \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \mathbb{T} & \xrightarrow{R_\alpha} & \mathbb{T} \end{array}$$

donde $T_\alpha(x) = x + \alpha - \lfloor x + \alpha \rfloor$. Es claro que $T_\alpha$ preserva la medida de Lebesgue en $\mathbb{R}$, luego

$$(R_\alpha)_*\mu = (R_\alpha)_*\pi_*\lambda = \pi_*(T_\alpha)_*\lambda = \pi_*\lambda = \mu.$$

(2) Más en general, si $G$ es un grupo localmente compacto y Hausdorff, entonces para todo $g \in G$ el mapa $T : G \to G$, $T(x) = gx$ preserva la medida de Haar a izquierda, por definición.

(3) Sea $G = \mathrm{PSL}_2(\mathbb{R})$. El *flujo geodésico* en $G$ es la acción a izquierda $\varphi : \mathbb{R} \times G \to G$ dada por $\varphi_t(g) = g \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$. Es claro que el flujo preserva la medida de Haar a derecha. En el capítulo siguiente veremos que $G$ es unimodular. Luego el flujo preserva la medida de Haar a izquierda.

(4) (Esquemas de Bernoulli) Sea $(\Omega, \mathscr{F}, \mu)$ un espacio de probabilidad. Usualmente se toma un conjunto finito $\Omega = \{w_1, \dots, w_n\}$ con $\mathscr{F} = \mathscr{P}(\Omega)$ y $\mu(\{w_i\}) = p_i$. Sea $X = \Omega^{\mathbb{N}}$ con la medida producto $\mu^{\mathbb{N}}$ y sea $S : X \to X$ dado por $S(x_1, x_2, \dots) = (x_2, x_3, \dots)$. Entonces $S$ preserva a $\mu^{\mathbb{N}}$. En efecto, para todo rectángulo medible $A = A_1 \times A_2 \times \cdots \times A_n \times \prod_{m \geq n+1} \Omega$, su preimagen es

$$S^{-1}(A) = \Omega \times A_1 \times \cdots \times A_n \times \prod_{m \geq n+2} \Omega$$

y ambos conjuntos tienen medida igual a $\mu(A_1) \dots \mu(A_n)$. Como los rectángulos generan la $\sigma$-álgebra producto, se sigue que $\mu^{\mathbb{N}}(S^{-1}B) = \mu^{\mathbb{N}}(B)$ para todo conjunto medible $B$. El sistema dinámico $(X, \mathscr{F}^{\otimes\mathbb{N}}, \mu^{\mathbb{N}}, S)$ se llama *esquema o shift de Bernoulli*. Análogamente es posible definir una versión inversible del sistema tomando como espacio $X = \Omega^{\mathbb{Z}}$. Para distinguirlos llamamos unilateral al primero y bilateral al último.

(5) Si $X$ es un espacio métrico compacto y $T : X \to X$ es una función continua, por el teorema de Krylov-Bogolyubov existe una medida de probabilidad de Radon $\mu$ que es $T$-invariante (ver [EW11, corolario 4.2]). Sin embargo, existen acciones de grupos que no admiten medidas invariantes.

**Proposición 1.22.** *Sea $T : X \to X$ una función medible en un espacio de medida $(X, \mathscr{B}, \mu)$. Son equivalentes:*

$(i)$ *$\mu$ es $T$-invariante.*

$(ii)$ *Para toda $f \in L^1_\mu(X)$ se tiene que $\int f \, d\mu = \int f \circ T \, d\mu$.*

*Demostración.* Supongamos que se cumple $(ii)$ y sea $B \in \mathscr{B}$ un conjunto medible. Si tomamos $f = \chi_B$ tenemos que $\mu(B) = \int \chi_B \, d\mu = \int \chi_B \circ T \, d\mu = \int \chi_{T^{-1}B} \, d\mu = \mu(T^{-1}B)$, así que $\mu$ es $T$-invariante.

Recíprocamente, si $T$ preserva a $\mu$ entonces la ecuación del punto $(ii)$ se verifica para $\chi_B$, para todo $B \in \mathscr{B}$ y luego por linealidad se verifica para toda función simple. Dada $f \in L^1_\mu(X)$, tras descomponerla en parte real e imaginaria y luego como $f = f_+ - f_-$ podemos suponer que $f \geq 0$. Entonces existe una sucesión $(\varphi_n)_{n\geq 1}$ de funciones simples tales que $\varphi_n \nearrow f$ puntualmente. Entonces $\varphi \circ T \nearrow f \circ T$ y por el teorema de convergencia monótona,

$$\int f \circ T \, d\mu = \lim_{n\to\infty} \int \varphi_n \circ T \, d\mu = \lim_{n\to\infty} \int \varphi_n \, d\mu = \int f \, d\mu. \qquad \square$$

**Teorema 1.23** (Teorema de recurrencia de Poincaré)**.** *Sea $(X, \mathscr{B}, \mu)$ un espacio de probabilidad, sea $T : X \to X$ una transformación que preserva a $\mu$ y sea $B \in \mathscr{B}$ un conjunto con medida no nula. Entonces casi todo punto $x \in B$ regresa a $B$ infinitas veces, es decir, el conjunto*

$$\{n \in \mathbb{N} : T^n(x) \in B\}$$

*es infinito para casi todo $x \in B$.*

El teorema sigue siendo válido si $\mu(B) = 0$, pero en ese caso el enunciado del teorema es trivial ya que el conjunto de puntos que no regresan a $B$ necesariamente tiene medida 0 (este conjunto es medible ya que es igual a $\liminf_{n\to\infty} B \setminus T^{-n}B = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} B \setminus T^{-m}B$). Tampoco es necesario que $X$ sea un espacio de probabilidad, solo hay que pedir $\mu(X) < \infty$ en cuyo caso la demostración es la misma. Sin embargo, el teorema no vale para espacios de medida infinita, como se ve tomando $X = \mathbb{R}$, $T(x) = x + 1$ y $B \subseteq \mathbb{R}$ cualquier conjunto acotado (medible).

*Demostración.* Sea $A = \{x \in B : T^n(x) \notin B, \forall n \geq 1\} = B \setminus \bigcup_{n\geq 1} T^{-n}(B)$. Entonces

$$T^{-m}(A) = T^{-m}(B) \setminus \bigcup_{n\geq m+1} T^{-n}(B)$$

así que $T^{-i}(A) \cap T^{-j}(A) = \varnothing$ para todo par de enteros no negativos $i \neq j$. Además todos los conjuntos $T^{-m}(A)$ tienen igual medida por ser $\mu$ $T$-invariante. Se sigue que $\mu(A) = 0$, ya que en caso contrario la unión $\bigcup_{m\geq 1} T^{-m}(A)$ tendría medida infinita, lo cual no es posible. Luego el conjunto $C_1 = B \setminus A \subseteq B$ tiene $\mu(C_1) = \mu(B)$ y verifica que todo punto $x \in C_1$ regresa a $B$ al menos una vez. Aplicando el mismo argumento a $T^2, T^3, \ldots$ hallamos conjuntos $C_n \subseteq B$, $n \geq 1$ con $\mu(C_n) = \mu(B)$ tales que para todo $x \in C_n$ existe $k \geq 1$ tal que $T^{nk}(x) \in B$. Entonces el conjunto $C = \bigcap_{n\geq 1} C_n$ tiene $\mu(C) = \mu(B)$ y todo punto $x \in C$ regresa a $B$ infinitas veces. $\square$

## 1.3. Transformaciones ergódicas

**Definición 1.24.** Sea $(X, \mathscr{B}, \mu)$ un espacio de medida y $T : X \to X$ una función medible. Decimos que $T$ es *ergódica* si para todo conjunto $T$-invariante $B \in \mathscr{B}$ se tiene $\mu(B) = 0$ o $\mu(X \setminus B) = 0$. En tal caso $\mu$ es *$T$-ergódica*. Análogamente, si $G$ es un grupo que actúa en $X$, decimos que la acción es *ergódica* si todo conjunto $G$-invariante $B \in \mathscr{B}$ verifica $\mu(B) = 0$ o $\mu(X \setminus B) = 0$.

Es decir, la ergodicidad nos dice que no es posible descomponer el sistema en sistemas más chicos de manera no trivial.

**Ejemplos.** (1) Sea $p \in X$ un punto y supongamos que $\mathscr{B} = \mathscr{P}(X)$ es el conjunto de partes y tomamos la medida de Dirac $\delta_p$. Entonces toda transformación $T : X \to X$ es ergódica ya que todos los conjuntos tienen medida 0 o 1.

(2) Sea $R_\alpha : \mathbb{T} \to \mathbb{T}$ la rotación del círculo vista en la sección anterior y supongamos que $\alpha$ es racional, digamos $\alpha = \frac{m}{n}$ con $m, n \in \mathbb{Z}$ coprimos. Entonces $R_\alpha$ no es ergódica para la medida de Lebesgue, ya que el conjunto $\bigcup_{k=0}^{n-1} \left[\frac{2k}{2n}, \frac{2k+1}{2n}\right]$ es $R_\alpha$-invariante y tiene medida $\frac{1}{2}$. Sin embargo, $R_\alpha$ posee medidas invariantes y ergódicas, por ejemplo $\mu = \frac{1}{n}\sum_{p\in\mathbb{T}[n]} \delta_p$ donde $\mathbb{T}[n] = \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\right\}$.

Para dar ejemplos no triviales de transformaciones ergódicas, necesitamos criterios de ergodicidad.

**Definición 1.25.** Sea $(X, \mathscr{B}, \mu)$ un espacio de medida y sea $T : X \to X$ medible.

(1) Un conjunto $B \in \mathscr{B}$ se dice *esencialmente $T$-invariante* si $\mu(B \triangle T^{-1}B) = 0$.

(2) Una función medible $f : X \to \mathbb{C}$ se dice *esencialmente $T$-invariante* si $f \circ T = f$ c.t.p.

**Teorema 1.26.** *Sea $(X, \mathscr{B}, \mu, T)$ un sistema dinámico. Son equivalentes:*

*(i) $T$ es ergódica.*

*(ii) Todo conjunto esencialmente $T$-invariante $B \in \mathscr{B}$ cumple $\mu(B) = 0$ o $\mu(B) = 1$.*

*(iii) Para todo $A \in \mathscr{B}$ tal que $\mu(A) > 0$ se tiene $\mu(\bigcup_{n=1}^{\infty} T^{-n}A) = 1$.*

*(iv) Para todo par de conjuntos $A, B \in \mathscr{B}$ de medida no nula, existe $n \geq 1$ tal que*

$$\mu(T^{-n}A \cap B) > 0.$$

*(v) Toda función $f : X \to \mathbb{C}$ esencialmente $T$-invariante es igual a una constante en casi todo punto.*

*(vi) Toda función $f \in L^p_\mu(X)$ esencialmente $T$-invariante es igual a una constante en casi todo punto.*

*Demostración.* $(i) \Longrightarrow (ii)$ Sea $B \in \mathscr{B}$ esencialmente $T$-invariante. Consideremos el conjunto

$$A = \limsup_{n\to\infty} T^{-n}B = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} T^{-m}B.$$

Afirmamos que $A \triangle B \subseteq \bigcup_{m=0}^{\infty} T^{-m}(B \triangle T^{-1}B)$. Sea $x \in B \setminus A$. Como $x \notin A$ existe un primer entero $n \geq 1$ tal que $x \notin T^{-n}B$. Si $m = n - 1$ entonces $x \in T^{-n}B \setminus T^{-(n+1)}B \subseteq T^{-n}B \triangle T^{-(n+1)}B = T^{-n}(B \triangle T^{-1}B)$. Análogamente se demuestra que $A \setminus B \subseteq \bigcup_{m=0}^{\infty} T^{-m}(B \triangle T^{-1}B)$. Luego

$$\mu(A \triangle B) \leq \sum_{m=0}^{\infty} \mu(T^{-m}(B \triangle T^{-1}B)) = 0$$

por $T$-invariancia de $\mu$ y porque $\mu(B \triangle T^{-1}B) = 0$. En particular $\mu(A) = \mu(B)$. Además notemos que

$$T^{-1}A = \bigcap_{n=2}^{\infty} \bigcup_{m=n}^{\infty} T^{-m}B = A.$$

Como $T$ es ergódica, $\mu(A) \in \{0, 1\}$ y luego $\mu(B) \in \{0, 1\}$.

$(ii) \Longrightarrow (iii)$ Sea $A \in \mathscr{B}$ un conjunto de medida positiva y sea $B = \bigcup_{n=1}^{\infty} T^{-n}A$. Notemos que $T^{-1}B = \bigcup_{n=2}^{\infty} T^{-n}A \subseteq B$. Como $\mu(T^{-1}B) = \mu(B)$, se sigue que $\mu(B \triangle T^{-1}B) = \mu(B \setminus T^{-1}B) = 0$. Por $(ii)$ se sigue que $\mu(B) \in \{0, 1\}$. Pero $B$ no puede tener medida 0 ya que $T^{-1}A \subseteq B$. Luego $\mu(B) = 1$, como queríamos ver.

$(iii) \Longrightarrow (iv)$ Supongamos que vale $(iii)$ y sean $A, B \in \mathscr{B}$ conjuntos de medida positiva. Como $\mu(\bigcup_{n=1}^{\infty} T^{-n}A) = 1$, tenemos que

$$0 < \mu(B) = \mu\left(B \cap \bigcup_{n=1}^{\infty} T^{-n}A\right) = \mu\left(\bigcup_{n=1}^{\infty} B \cap T^{-n}A\right) \leq \sum_{n=1}^{\infty} \mu(B \cap T^{-n}A).$$

Luego $\mu(B \cap T^{-n}A) > 0$ para algún $n \geq 1$.

$(iv) \implies (v)$ Notemos que $(iv)$ implica que para todo par de conjuntos $A, B \in \mathcal{B}$ de medida positiva que son esencialmente $T$-invariantes, su intersección tiene medida positiva. Esto se debe a que $\mu(A \triangle T^{-n}A) = 0$, puesto que

$$\begin{aligned} A \triangle T^{-n}A &= (A \triangle T^{-1}A) \triangle (T^{-1}A \triangle T^{-2}A) \triangle \cdots \triangle (T^{-n+1}A \triangle T^{-n}A) \\ &\subseteq \bigcup_{i=0}^{n-1} (T^{-i}A \triangle T^{-(i+1)}A) = \bigcup_{i=0}^{n-1} T^{-i}(A \triangle T^{-1}A) \end{aligned}$$

Sea $f : X \to \mathbb{C}$ una función medible y $T$-invariante. Tomando parte real e imaginaria podemos suponer que $f(X) \subseteq \mathbb{R}$. Dados $k \in \mathbb{Z}$ y $n \geq 0$, sea

$$A_{n,k} = \{x \in X : f(x) \in \left[\tfrac{k}{2^n}, \tfrac{k+1}{2^n}\right)\}.$$

Este conjunto es esencialmente $T$-invariante, ya que $A_{n,k} \triangle T^{-1}A_{n,k} \subseteq \{x \in X : f(x) \neq f \circ T(x)\}$. Fijemos $n \geq 0$ y sean $k \neq \ell$ enteros. Como $A_{n,k} \cap A_{n,\ell} = \varnothing$, por $(iv)$ alguno de los dos tiene medida 0. Luego para cada $n$ existe un único $k(n) \in \mathbb{Z}$ tal que $\mu(A_{n,k(n)}) \neq 0$ y como $X = \coprod_{k\in\mathbb{Z}} A_{n,k}$, se sigue que $\mu(A_{n,k(n)}) = 1$. Consideremos el conjunto

$$Y = \bigcap_{n=0}^{\infty} A_{n,k(n)},$$

que tiene medida 1. Como $f(Y)$ tiene diámetro 0, debe ser un punto y luego $f$ es igual a una constante en casi todo punto.

La implicación $(v) \implies (vi)$ es clara. Finalmente, supongamos que vale $(vi)$ y veamos que esto implica $(i)$. Sea $A \in \mathcal{B}$ un conjunto $T$-invariante. Entonces la función $f = \chi_A \in L^p_\mu(X)$ es $T$-invariante, así que por $(vi)$ $f$ es igual a una constante en casi todo punto. Como $f(X) \subseteq \{0, 1\}$ esta constante solo puede ser 0 o 1. En el primer caso $\mu(A) = 0$ y en el segundo $\mu(A) = 1$. Por lo tanto $T$ es ergódica. □

**Proposición 1.27.** *$R_\alpha : \mathbb{T} \to \mathbb{T}$ es ergódica con respecto a la medida de Lebesgue si y solo si $\alpha$ es irracional.*

*Demostración.* Ya vimos que $R_\alpha$ no puede ser ergódica si $\alpha \in \mathbb{Q}$. Supongamos que $\alpha$ no es racional y sea $f \in L^2(\mathbb{T})$ una función esencialmente $R_\alpha$-invariante. Entonces $f$ admite una expansión de Fourier $f(t) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi int}$ que converge en $L^2$, donde $c_n = \int_0^1 f(t)e^{-2\pi int}\,dt$. Como $f(t+\alpha) = f(t)$ c.t.p. esto implica que

$$c_n = \int_{\mathbb{T}} f(t+\alpha)e^{-2\pi int}\,dt = \int_{\mathbb{T}} f(t)e^{-2\pi in(t-\alpha)}\,dt = e^{2\pi in\alpha}c_n.$$

Como $\alpha$ es irracional, $e^{2\pi in\alpha} \neq 1$ para todo $n \neq 0$ y luego $c_n = 0$ para $n \neq 0$. Por lo tanto $f$ es constante c.t.p. así que $R_\alpha$ es ergódica. □

**Proposición 1.28.** *Todo shift de Bernoulli es ergódico.*

*Demostración.* Sea $S : X \to X$ el shift de Bernoulli unilateral asociado a un espacio de probabilidad $(\Omega, \mathcal{F}, \mu)$, descripto en la sección anterior. Sea $B \subseteq \Omega^{\mathbb{N}}$ medible y $S$-invariante. Dado $\varepsilon \in (0, 1)$, por el teorema de aproximación existe un conjunto $A$ que es unión finita de rectángulos medibles, tal que $\mu(A \triangle B) < \varepsilon$. En particular $|\mu(A) - \mu(B)| < \varepsilon$. Como $A$ es unión finita de rectángulos medibles, es de la forma $A = F \times \prod_{m\geq n+1} \Omega$ donde $F \subseteq \Omega^n$ es un conjunto medible. Entonces $S^{-n}(A) \setminus A = (X \setminus F) \times F \times \prod_{m\geq 2n+1} \Omega$ y luego es fácil ver que $\mu(S^{-n}A \setminus A) = \mu(A)\mu(X \setminus A)$. Ahora,

$$\begin{aligned} \mu(S^{-n}A \triangle A) &\leq \mu(S^{-n}A \triangle B) + \mu(B \triangle A) \\ &= \mu(S^{-n}A \triangle S^{-n}B) + \mu(B \triangle A) \\ &= \mu(A \triangle B) + \mu(B \triangle A) < 2\varepsilon. \end{aligned}$$

Luego

$$\begin{aligned}\mu(B)\mu(X\setminus B) &< (\mu(A)+\varepsilon)(\mu(X\setminus A)+\varepsilon)\\ &= \mu(A)\mu(X\setminus A)+\varepsilon[\mu(A)+\mu(X\setminus A)]+\varepsilon^2\\ &= \mu(S^{-n}A\setminus A)+\varepsilon+\varepsilon^2<4\varepsilon.\end{aligned}$$

Como esto vale para todo $\varepsilon\in(0,1)$, se sigue que $\mu(B)\mu(X\setminus B)=0$ y luego $\mu(B)\in\{0,1\}$. Por lo tanto $S$ es ergódico. El mismo argumento es válido para el shift bilátero. □

## 1.4. El teorema ergódico de Von Neumann

El primer teorema ergódico que consideraremos es el teorema ergódico de Von Neumann, que trata de la convergencia en $L^2_\mu$ de los promedios ergódicos

$$A_Nf(x)=\frac{1}{N}\sum_{n=0}^{N-1}f(T^nx), \tag{1.5}$$

donde $T^0$ es la identidad. Para todo $p\in[1,+\infty]$ la aplicación $U_T:L^p_\mu(X)\to L^p_\mu(X)$, $U_Tf=f\circ T$ está bien definida. En efecto, si $f,g:X\to\mathbb{C}$ son funciones medibles que representan el mismo elemento de $L^p_\mu(X)$, entonces

$$\{x:\ f\circ T(x)\neq g\circ T(x)\}=T^{-1}\{x:\ f(x)\neq g(x)\}$$

y luego $\mu(\{x:\ f\circ T(x)\neq g\circ T(x)\})=\mu(\{x:\ f(x)\neq g(x)\})=0$ ya que $T$ preserva la medida. Por el mismo motivo,

$$\|U_Tf\|_p=\left(\int_X|f\circ T|^p\,d\mu\right)^{1/p}=\left(\int_X|f|^p\,d\mu\right)^{1/p}=\|f\|_p<\infty$$

para toda $f\in L^p_\mu(X)$, así que $U_Tf\in L^p_\mu(X)$ (si $p=\infty$ es claro que $f\circ T$ es esencialmente acotada si $f$ lo es). Esto además demuestra que $U_T$ es un operador acotado, con $\|U_T\|=1$. El operador $U_T$ se suele denominar el *operador de Koopman*. Es claro que $A_N$ también es acotado como operador $A_N:L^p_\mu(X)\to L^p_\mu(X)$ por ser combinación lineal de potencias de $U_T$. Más precisamente,

$$\|A_N\|=\left\|\frac{1}{N}\sum_{n=0}^{N-1}U_T^n\right\|\leq\frac{1}{N}\sum_{n=0}^{N-1}\|U_T\|^n\leq 1,$$

así que $\|A_Nf\|_p\leq\|f\|_p$ para toda $f\in L^p_\mu(X)$.

En el caso $p=2$, podemos decir algo más de $U_T$. El hecho de que $\|U_Tf\|_2=\|f\|_2$ para toda $f$ implica que $U_T$ es una isometría, es decir, que cumple

$$\langle U_Tf,U_Tg\rangle=\langle f,g\rangle$$

para todo par de funciones $f,g\in L^2_\mu(X)$. Si además $T$ es inversible entonces $U_T\in\mathcal{U}(L^2_\mu(X))$, el grupo de operadores unitarios, ya que $U_T$ es una isometría inversible.

**Teorema 1.29** (Von Neumann)**.** *Sean $(X,\mathcal{B},\mu,T)$ un sistema dinámico y $P_T:L^2_\mu(X)\to L^2_\mu(X)$ la proyección ortogonal al subespacio cerrado*

$$S=\{f\in L^2_\mu(X):U_Tf=f\}.$$

*Entonces, para toda $f\in L^2_\mu(X)$ se tiene que*

$$\frac{1}{N}\sum_{n=0}^{N-1}U_T^nf\xrightarrow{L^2_\mu}P_Tf. \tag{1.6}$$

*Demostración.* Sea $E = \{U_T g - g : g \in L^2_\mu(X)\}$. Afirmamos que $E^\perp = S$. Si $f \in S$, entonces para toda $g \in L^2_\mu(X)$,

$$\langle f, U_T g - g\rangle = \langle f, U_T g\rangle - \langle f, g\rangle = \langle U_T f, U_T g\rangle - \langle f, g\rangle = 0,$$

así que $f \in E^\perp$. Recíprocamente, si $f \in E^\perp$ entonces

$$\langle g, f\rangle = \langle U_T g, f\rangle = \langle g, U_T^* f\rangle$$

para toda $g \in L^2_\mu(X)$, así que $f = U_T^* f$. Luego

$$\begin{aligned}\|U_T f - f\|_2^2 &= \|U_T f\|_2^2 - \langle U_T f, f\rangle - \langle f, U_T f\rangle + \|f\|_2^2 \\ &= 2\langle f, f\rangle - \langle U_T^* f, f\rangle - \langle f, U_T^* f\rangle = \langle f, f - U_T^* f\rangle + \langle f - U_T^* f, f\rangle = 0.\end{aligned}$$

Por lo tanto $f = U_T f$ y $f \in S$.

Como $S$ es un subespacio cerrado de $L^2_\mu(X)$, hay una descomposición $L^2_\mu(X) = S \oplus S^\perp$ y notemos que $S^\perp = (E^\perp)^\perp = \overline{E}$. Entonces por linealidad basta probar que se cumple (1.6) en $S$ y en $\overline{E}$. Es claro que (1.6) se cumple para $f \in S$ ya que $U_T^n f = f$ para todo $n$ en ese caso. Si $h = U_T g - g \in E$ entonces

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} U_T^n h \right\|_2 = \left\| \frac{1}{N} \sum_{n=0}^{N-1} (U_T^{n+1} g - U_T^n g) \right\|_2 = \frac{1}{N} \|U_T^N g - g\|_2 \to 0$$

y luego $\frac{1}{N}\sum_{n=0}^{N-1} U_T^n h \to 0 = P_T h$. Si en cambio $h \in \overline{E}$, podemos hallar una sucesión $(h_k)_{k\geq 1}$ en $E$ tal que $h = \lim_{k\to\infty} h_k$. Para todo $k$ tenemos que

$$\|A_N h\|_2 \leq \|A_N(h - h_k)\|_2 + \|A_N h_k\|_2.$$

Dado $\varepsilon > 0$ sea $k \in \mathbb{N}$ tal que $\|h - h_k\|_2 < \varepsilon$. Entonces $\|A_N(h - h_k)\|_2 \leq \|h - h_k\|_2 < \varepsilon$ para todo $N > 0$, luego

$$\limsup_{N\to\infty} \|A_N h\|_2 \leq \limsup_{N\to\infty} \|A_N(h - h_k)\|_2 + \limsup_{N\to\infty} \|A_N h_k\|_2 \leq \varepsilon.$$

Como esto vale para todo $\varepsilon > 0$, se sigue que $A_N h \to 0 = P_T h$, como queríamos ver. □

## 1.5. El teorema ergódico de Birkhoff

En esta sección probamos el teorema ergódico de Birkhoff, también llamado teorema ergódico puntual, que garantiza la convergencia c.t.p. de los promedios ergódicos $A_N f$, y mencionamos un par de aplicaciones.

**Teorema 1.30** (Birkhoff)**.** *Sea $(X, \mathcal{B}, \mu)$ un espacio de probabilidad y sea $T : X \to X$ una transformación que preserva medida. Entonces para toda $f \in L^1_\mu(X)$, se tiene que*

$$\frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) \longrightarrow \mathbb{E}(f|\mathcal{B}^T), \qquad \text{para } \mu\text{-casi todo } x \in X, \tag{1.7}$$

*donde $\mathcal{B}^T$ es la $\sigma$-álgebra de conjuntos esencialmente $T$-invariantes. Además $A_N f \xrightarrow{L^1_\mu} \mathbb{E}(f|\mathcal{B}^T)$ cuando $N \to \infty$.*

Si además $T$ es ergódica, $\mathbb{E}(f|\mathcal{B}^T)$ es una constante y luego $\mathbb{E}(f|\mathcal{B}^T) = \int f\, d\mu$ para $\mu$-casi todo $x$. Por lo tanto obtenemos el siguiente corolario:

**Corolario 1.31.** *Si $T$ es ergódica entonces*

$$\frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) \longrightarrow \int f\, d\mu, \qquad \text{para } \mu\text{-casi todo } x \in X. \tag{1.8}$$

Es decir, si $T$ es una transformación ergódica entonces para todo observable $f \in L^1_\mu(X)$, el promedio temporal del valor de $f$ en la órbita de $x$ se aproxima al promedio espacial de $f$ y esto nos dice que la órbita de casi todo punto $x \in X$ se distribuye uniformemente en $X$.

**Observación 1.32.** Recíprocamente, cualquier medida $\mu \in \mathscr{M}^1_T(X)$ que verifica (1.8) para toda $f \in L^1_\mu(X)$ es $T$-ergódica. En efecto, si aplicamos esto a una función $T$-invariante obtenemos que $f(x) \to \mu(f)$ para $\mu$-casi todo $x$, de modo que $f$ es constante $\mu$-c.t.p. Luego $\mu$ es ergódica por el punto $(vi)$ de 1.26.

Antes de demostrar el teorema, mencionamos un par de aplicaciones:

**Corolario 1.33.** *Si $T$ es ergódica, para todo conjunto $A \in \mathscr{B}$, se tiene que*

$$\lim_{N\to\infty} \frac{1}{N}|\{n \in \{1,\dots,N\} : T^n(x) \in A\}| \to \mu(A)$$

*para casi todo $x \in X$.*

Esto se deduce de aplicar el corolario 1.31 a $f = \chi_A$.

**Teorema 1.34** (Ley fuerte de los grandes números)**.** *Sean $(X_n)_{n\in\mathbb{N}}$ variables aleatorias reales i.i.d. en un espacio de probabilidad $(\Omega, \mathscr{F}, \mu)$ y supongamos que $\mathbb{E}|X_1| < \infty$. Entonces*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{} \mathbb{E}X_1 \quad \textit{casi seguro.} \tag{1.9}$$

*Demostración.* Consideremos el espacio $Y = \mathbb{R}^{\mathbb{N}}$ con la topología producto y el mapa $\theta : \Omega \to Y$, $\theta(\omega) = (X_1(\omega), X_2(\omega), \dots)$. Si le damos a $Y$ la $\sigma$-álgebra de Borel $\mathscr{B}_Y$, entonces $\theta$ es medible ya que para todo abierto básico $U$, es decir de la forma $U = U_1 \times \cdots \times U_n \times \prod_{m=n+1}^{\infty} \mathbb{R}$, su preimagen

$$\theta^{-1}(U) = X_1^{-1}(U_1) \cap \cdots \cap X_n^{-1}(U_n)$$

es medible. Como los abiertos básicos generan la $\sigma$-álgebra de Borel[1], la preimagen de cualquier Boreliano es medible. Luego podemos definir la medida $\nu = \theta_*\mu$ en $Y$. Además sea $\eta = (X_1)_*\mu$ la medida inducida por $X_1 : \Omega \to \mathbb{R}$. Dado un abierto $U \subseteq \mathbb{R}$ y $n \in \mathbb{N}$ definimos

$$U^{(n)} = \mathbb{R}^{n-1} \times U \times \prod_{m=n+1}^{\infty} \mathbb{R}.$$

Para todo abierto $U \subseteq \mathbb{R}$, tenemos que

$$\nu(U^{(n)}) = \mathbb{P}(X_n \in U) = \mathbb{P}(X_1 \in U) = \eta(U).$$

Como los abiertos $U^{(n)}$ generan la $\sigma$-álgebra $\mathscr{B}_Y$, se sigue que $\nu = \eta^{\mathbb{N}}$, la medida producto en $Y$. Ahora consideremos la transformación $T : Y \to Y$, $T(x_1, x_2, \dots) = (x_2, x_3, \dots)$. Esta transformación preserva la medida y por 1.28 es ergódica. Sea $f : Y \to \mathbb{R}$, $f(x_1, x_2, \dots) = x_1$. Esta función está en $L^1_\nu(Y)$ puesto que

$$\int_Y |f|\, d\nu = \int_Y |f|\, d\theta_*\mu = \int_\Omega |f \circ \theta|\, d\mu = \int_\Omega |X_1|\, d\mu = \mathbb{E}|X_1| < \infty.$$

Luego estamos en condiciones de aplicar el teorema ergódico de Birkhoff, que nos dice que

$$\frac{1}{N}\sum_{n=0}^{N-1} f(T^n(y)) = \frac{1}{N}\sum_{n=1}^{N} y_n \longrightarrow \int_Y f\, d\nu = \mathbb{E}X_1$$

para $\nu$-casi todo $y \in Y$. Es decir, si definimos $Z = \{y \in Y : \frac{1}{N}\sum_{n=1}^{N} y_n \not\to \mathbb{E}X_1\}$ entonces $\nu(Z) = \mu(\theta^{-1}Z) = 0$. Ahora, $\theta^{-1}(Z)$ es el conjunto de puntos $\omega \in \Omega$ tales que $\frac{1}{N}\sum_{n=1}^{N} X_n \not\to \mathbb{E}X_1$. □

[1] Notar que $Y$ es un espacio métrico ya que hay numerables factores en el producto $\prod_{n=1}^{\infty}\mathbb{R}$. Además $Y$ es separable ya que el subconjunto $E = \bigoplus_{n=1}^{\infty}\mathbb{Q}$ es denso en $Y$. Luego $Y$ es Lindelöf, de modo que todo abierto de $Y$ es unión a lo sumo numerable de abiertos básicos. Se sigue que los abiertos básicos generan $\mathscr{B}_Y$.

Ahora procedemos a probar el teorema, probando primero el teorema ergódico maximal 1.36. Para eso necesitamos el siguiente lema:

**Lema 1.35.** *Bajo las condiciones del teorema 1.30, sea $f \in L^1_\mu(X)$, sea $f_0 = 0$ y sea*

$$f_n = \sum_{i=0}^{n-1} f \circ T^i$$

*para $n \geq 1$. Sea $F_N(x) = \max\limits_{0\leq n\leq N} f_n(x)$. Entonces para todo $N \geq 1$ se tiene*

$$\int_{\{F_N>0\}} f\, d\mu \geq 0. \tag{1.10}$$

*Demostración.* Observemos primero que $F_N(x) \geq 0$ para todo $x$. Dado $N \geq 1$, para todo $n \leq N$ se tiene

$$F_N \circ T + f \geq f_n \circ T + f = f_{n+1}.$$

Luego $F_N \circ T + f \geq \max_{1\leq n\leq N} f_n$. Si $x \in P_N = \{y : F_N(y) > 0\}$ entonces $F_N(x) = \max_{1\leq n\leq N} f_n(x)$ y luego

$$F_N \circ T(x) + f(x) \geq F_N(x).$$

Se sigue que $f \geq F_N - F_N \circ T$ en $P_N$, así que

$$\begin{aligned}\int_{P_N} f\, d\mu &\geq \int_{P_N} F_N\, d\mu - \int_{P_N} F_N \circ T\, d\mu \\ &= \int_X F_N\, d\mu - \int_{P_N} F_N \circ T\, d\mu \\ &\geq \int_X F_N\, d\mu - \int_X F_N \circ T\, d\mu = 0\end{aligned}$$

ya que $T$ preserva la medida. □

**Teorema 1.36** (Teorema ergódico maximal)**.** *Sea $(X, \mathcal{B}, \mu)$ un espacio de probabilidad, sea $T : X \to X$ una transformación que preserva medida y sea $g \in L^1_\mu(X)$. Para cada $\alpha \in \mathbb{R}$, sea*

$$E_\alpha = \left\{ x \in X : \sup_{N\geq 1} A_N g > \alpha \right\}.$$

*Entonces*

$$\alpha\mu(E_\alpha) \leq \int_{E_\alpha} g\, d\mu. \tag{1.11}$$

*Además, $\alpha\mu(E_\alpha \cap A) \leq \int_{E_\alpha \cap A} g\, d\mu$ para todo conjunto $T$-invariante $A \in \mathcal{B}$.*

*Demostración.* Tomamos $f = g - \alpha$ y sea $F_N$ como en el lema 1.35. Entonces

$$F_N = \max_{0\leq n\leq N} nA_n f = \max_{0\leq n\leq N} n(A_n g - \alpha).$$

Se sigue que $F_N > 0$ si y solo si $A_n g > \alpha$ para algún $n \leq N$ y luego

$$E_\alpha = \bigcup_{N\geq 0} \{x \in X : F_N(x) > 0\}.$$

Notemos además que $F_N \leq F_{N+1}$ para todo $N$, así que los conjuntos $S_N = \{x \in X : F_N(x) > 0\}$ forman una sucesión creciente $S_0 \subseteq S_1 \subseteq S_2 \subseteq \ldots$. Entonces $f\chi_{S_N} \to f\chi_{E_\alpha}$ cuando $N \to \infty$ y por convergencia dominada,

$$\int_{E_\alpha} f\, d\mu = \lim_{N\to\infty} \int_{S_N} f\, d\mu \geq 0.$$

Luego $\int_{E_\alpha} g\, d\mu \geq \alpha\mu(E_\alpha)$. La segunda afirmación se deduce de aplicar el mismo argumento al sistema $(A, \mathcal{B}|_A, \frac{1}{\mu(A)}\mu|_A, T|_A)$. □

*Demostración de 1.30.* Consideremos las funciones

$$f^*(x) = \limsup_{N\to\infty} A_N f(x),$$

$$f_*(x) = \liminf_{n\to\infty} A_N f(x).$$

Afirmamos que estas funciones son $T$-invariantes. En efecto,

$$f^* \circ T(x) = \limsup_{N\to\infty} A_N f(T(x)) = \limsup_{N\to\infty} \frac{1}{N}\left(\sum_{n=0}^{N} f\circ T^n(x) - f(x)\right)$$
$$= \limsup_{N\to\infty}\left(\frac{N+1}{N} A_{N+1} f(x) - \frac{1}{N} f(x)\right) = f^*(x).$$

y una cuenta similar muestra que $f_* \circ T(x) = f_*(x)$. Para ver que el límite $\lim_{N\to\infty} A_N f$ converge c.t.p. basta ver que $f^*(x) = f_*(x)$ para $\mu$-casi todo $x \in X$. Ahora notemos que

$$\{x \in X : f^*(x) \neq f_*(x)\} = \bigcup_{\substack{\alpha,\beta\in\mathbb{Q}\\ \alpha<\beta}} \{x \in X : f_*(x) < \alpha < \beta < f^*(x)\}.$$

Luego basta probar que el conjunto $E_\alpha^\beta = \{x \in X : f_*(x) < \alpha < \beta < f^*(x)\}$ tiene medida 0 para todos $\alpha < \beta$. Fijemos $\alpha < \beta$ y notemos que $E_\alpha^\beta \subseteq E_\beta$, usando la notación del teorema 1.36. Aplicando este teorema al conjunto $A = E_\alpha^\beta$ obtenemos que $\int_{E_\alpha^\beta} f\,d\mu \geq \beta\mu(E_\alpha^\beta)$. Por un argumento similar aplicado a $-f$ (que cumple $(-f)^* = -f_*$) obtenemos $\int_{E_\alpha^\beta} f\,d\mu \leq \alpha\mu(E_\alpha^\beta)$. Combinando estas desigualdades resulta $(\beta-\alpha)\mu(E_\alpha^\beta) \leq 0$ y luego $\mu(E_\alpha^\beta) = 0$. Por lo tanto $\lim_{N\to\infty} A_N f$ converge en $\mu$-casi todo punto.

Sea $\overline{f}(x) = \lim_{N\to\infty} A_N f(x)$. Esta función está en $L^1_\mu(X)$ ya que por el lema de Fatou,

$$\|\overline{f}\|_1 = \int \lim_{N\to\infty} |A_N f|\,d\mu \leq \liminf_{N\to\infty} \int |A_N f|\,d\mu \leq \liminf_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \|f\circ T^n\|_1 = \|f\|_1. \tag{1.12}$$

Queda ver que $\|A_N f - \overline{f}\|_1 \to 0$. Dado $\varepsilon > 0$ podemos descomponer a $f$ como $f = f_0 + r$ donde $f$ es acotada y $r \in L^1_\mu(X)$ verifica $\|r\|_1 < \varepsilon$. Definimos $\overline{f_0}(x) = \lim_{N\to\infty} A_N f_0(x)$ y $\overline{r}(x) = \lim_{N\to\infty} A_N r(x)$. Como $f_0$ es acotada todas las funciones $|A_N f_0 - \overline{f_0}|$ están acotadas puntualmente por una misma constante (que no depende de $N$). Entonces, por convergencia dominada tenemos que $\|A_N f_0 - \overline{f_0}\|_1 \to 0$ cuando $N \to \infty$. Además $\|\overline{r}\|_1 \leq \|r\|_1 < \varepsilon$ por la misma cuenta que aparece en (1.12). Luego

$$\limsup_{N\to\infty} \|A_N f - \overline{f}\|_1 \leq \limsup_{N\to\infty} \|A_N(f-f_0)\|_1 + \limsup_{N\to\infty} \|A_N f_0 - \overline{f_0}\|_1 + \|\overline{f} - \overline{f_0}\|_1$$
$$= \limsup_{N\to\infty} \|A_N r\|_1 + \|\overline{r}\|_1 \leq 2\varepsilon.$$

Por lo tanto $\|A_N f - \overline{f}\|_1 \to 0$. Finalmente veamos que $\overline{f} = \mathbb{E}(f|\mathscr{B}^T)$ $\mu$-c.t.p. Es claro que $\overline{f}$ es $\mathscr{B}^T$-medible ya que es esencialmente $T$-invariante. Además, para cualquier $B \in \mathscr{B}^T$ tenemos que

$$\int_B f\circ T^k\,d\mu = \int \mathbb{1}_B(f\circ T^k)\,d\mu = \int (\mathbb{1}_B \circ T^k)(f\circ T^k)\,d\mu = \int \mathbb{1}_B f\,d\mu,$$

para todo $k$. Se sigue que $\int_B A_N f\,d\mu = \int_B f\,d\mu$ y pasando al límite, $\int_B \overline{f}\,d\mu = \int_B f\,d\mu$. Por lo tanto $\overline{f} = \mathbb{E}(f|\mathscr{B}^T)$ c.t.p. □

## 1.6. La descomposición ergódica

**Definición 1.37.** Sean $(X,\mathscr{B})$ e $(Y,\mathscr{D})$ dos espacios medibles. Llamamos *núcleo de probabilidad* a una aplicación $Y \to \mathscr{M}^1(X)$, $y \mapsto \mu_y$ tal que el mapa $y \mapsto \mu_y(f)$ es medible para toda función $f : X \to \mathbb{C}$ medible y acotada.

**Definición 1.38.** Decimos que un espacio medible $(X, \mathscr{B})$ es *regular* si existe una topología compacta metrizable $\tau$ en $X$ tal que $\mathscr{B}$ es la $\sigma$-álgebra de Borel asociada.

La condición de regularidad no es tan restrictiva como podría parecer a primera vista. De hecho, para todo par de espacios métricos $X$, $Y$ completos y separables de igual cardinalidad, existe un isomorfismo Borel entre $X$ e $Y$, es decir, una biyección $f : X \to Y$ tal que tanto $f$ como $f^{-1}$ son medibles Borel (ver [Gla03, Teo. 2.8]). Se sigue de esto que si $X$ es un espacio métrico completo y separable y $\mathscr{B}$ es su $\sigma$-álgebra de Borel, entonces $(X, \mathscr{B})$ es un espacio regular.

**Teorema 1.39** (Desintegración de medidas)**.** *Sean $(X, \mathscr{B}, \mu)$, $(Y, \mathscr{D}, \nu)$ dos espacios de probabilidad con $(X, \mathscr{B})$ regular y sea $\pi : X \to Y$ un factor. Entonces existe un núcleo de probabilidad $y \mapsto \mu_y$ tal que*

$$\int_X f(g \circ \pi)\, d\mu = \int_Y \left( \int_X f\, d\mu_y \right) g(y)\, d\nu(y), \tag{1.13}$$

*para todo par de funciones $f \in L^\infty_\mu(X)$, $g \in L^\infty_\nu(Y)$. Más aún, si $y \mapsto \mu'_y$ es otro núcleo de probabilidad con las mismas propiedades, entonces $\mu_y = \mu'_y$ para $\nu$-casi todo $y \in Y$.*

*Demostración.* Como $(X, \mathscr{B})$ es regular, podemos suponer que $X$ es un espacio métrico compacto y que $\mathscr{B}$ es la $\sigma$-álgebra de Borel. Veamos primero la unicidad. Supongamos que existen dos núcleos de probabilidad $y \mapsto \mu_y$, $y \mapsto \mu'_y$ con las propiedades del enunciado. Entonces

$$\int_Y (\mu_y(f) - \mu'_y(f)) g(y)\, d\nu(y) = 0$$

para todo par de funciones medibles y acotadas $f, g$. Tomando $g(y) = \mu_y(f) - \mu'_y(f)$ vemos que $\mu_y(f) = \mu'_y(f)$ para $\nu$-casi todo $y$. Como $X$ es un espacio métrico compacto, posee una base de abiertos numerable $\{U_n\}_{n\in\mathbb{N}}$. Tomando $f = \mathbb{1}_{U_n}$, obtenemos que $\mu_y(U_n) = \mu'_y(U_n)$ para todo $y$ afuera de un conjunto $Z_n \subseteq Y$ de medida 0. Como los $U_n$ generan $\mathscr{B}$, se sigue que $\mu_y = \mu'_y$ para todo $y \in Y \setminus \bigcup_{n=1}^\infty Z_n$.

Ahora demostramos la existencia. Sea $\pi^\# : L^2_\nu(Y) \to L^2_\mu(X)$ el operador $\pi^\#(g) = g \circ \pi$ y sea $\pi_\# : L^2_\mu(X) \to L^2_\nu(Y)$ su operador adjunto. Entonces

$$\int_X f(g \circ \pi)\, d\mu = \int_Y \pi_\#(f) g\, d\nu$$

para todo $f \in L^2_\mu(X)$ y todo $g \in L^2_\nu(Y)$. Si $f \in C(X)$, entonces

$$\|\pi_\# f\|_{L^\infty_\nu} = \sup_{\substack{g\in L^2_\nu(Y) \\ \|g\|_{L^1_\nu} \le 1}} \int_Y \pi_\#(f) g\, d\nu = \sup_{\substack{g\in L^2_\nu(Y) \\ \|g\|_{L^1_\nu} \le 1}} \int_X f(g \circ \pi)\, d\mu \le \|f\|_{\sup},$$

ya que $L^\infty_\nu(Y)$ es isomorfo al dual de $L^1_\nu(Y)$ y $L^2_\nu(Y)$ es denso en $L^1_\nu(Y)$. Luego $\pi_\# f \in L^\infty_\nu(Y)$ para toda $f \in C(X)$. Ahora, como $C(X)$ es separable podemos tomar un conjunto $\{f_n\}_{n\in\mathbb{N}} \subseteq C(X)$ numerable que genera un subespacio denso. Podemos suponer además que $f_1 \equiv 1$ y que los $f_n$ son linealmente independientes. Para cada $n \in \mathbb{N}$ elegimos un representante $\widetilde{\pi}_\# f_n \in \mathscr{L}^\infty_\nu(Y)$ de $\pi_\# f_n$. Para $n = 1$ tomamos $\widetilde{\pi}_\#(1) = 1$. Llamando $V \subseteq C(X)$ al subespacio vectorial sobre $\mathbb{Q}(i)$ generado por $\{f_n\}_{n\in\mathbb{N}}$, podemos extender linealmente la definición de $\widetilde{\pi}_\#$ a $V$. Es claro que la clase de $\widetilde{\pi}_\# f$ en $L^\infty_\nu(Y)$ es igual a $\pi_\# f$ para todo $f \in V$, y en particular existe un conjunto $Z_f \subseteq Y$ con $\nu(Z_f) = 0$ tal que $|\widetilde{\pi}_\# f(y)| \le \|f\|_{\sup}$ para todo $y \in Y \setminus Z_f$. Como $V$ es numerable, la unión $Z = \bigcup_{f\in V} Z_f$ tiene medida 0. Para cada $y \in Y \setminus Z$, la transformación lineal $\phi_y : V \to \mathbb{C}$, $\phi_y(f) = \widetilde{\pi}_\# f(y)$ es acotada. Como $V$ es denso en $C(X)$, es posible extender $\phi_y$ de manera única a una funcional $\mathbb{C}$-lineal en $C(X)$. Por el teorema de representación de Riesz-Markov, existe una medida de Radon finita $\mu_y$ en $X$ tal que

$$\phi_y(f) = \int_X f\, d\mu_y$$

para toda $f \in C(X)$. Notemos que $\mu_y$ es una medida de probabilidad ya que $\int_X 1\, d\mu_y = \phi_y(f_1) = 1$. Para $y \in Z$ podemos tomar $\mu_y = \mu_0$, donde $\mu_0 \in \mathscr{M}^1(X)$ es una medida fija. De esta manera

obtenemos una aplicación $y \mapsto \mu_y$. Para $f \in V$ es claro que el mapa $y \mapsto \mu_y(f)$ es medible ya que es igual a $\widetilde{\pi}_{\#} f$ en $Y \setminus Z$ y es constante en $Z$. Para una función $f \in C(X)$ general, sea $(h_k)_k \subseteq C(X)$ una sucesión que tiende a $f$ uniformemente. Entonces $\mu_y(h_k) \to \mu_y(f)$ para todo $y \in Y$, así que la función $y \mapsto \mu_y(f)$ es medible. Ahora notemos que el conjunto $\mathcal{F}$ de funciones $f \in L^\infty_\mu(X)$ tales que el mapa $y \mapsto \mu_y(f)$ es medible es cerrado bajo convergencia monótona, ya que si $(h_k)_k \subseteq \mathcal{F}$ es una sucesión creciente de funciones no negativas tales que $h_k \to h$ puntualmente, entonces $\mu_y(h_k) \to \mu_y(h)$, así que la función $y \mapsto \mu_y(h)$ también es medible. Dado un abierto $U \subseteq X$, existe una sucesión creciente de funciones $(h_k)_k \subseteq C(X)$ no negativas tales que $h_k \to \mathbb{1}_U$ puntualmente. Por ejemplo se puede tomar $h_k(x) = 1 - (1 - kd(x, U^c))_+$, donde $d$ es una distancia que induce la topología de $X$. Se sigue que $\mathbb{1}_U \in \mathcal{F}$ para todo abierto $U$ y luego $\mathbb{1}_E \in \mathcal{F}$ para todo $E \in \mathcal{B}$. Dada una función $f \in L^\infty_\mu(X)$ no negativa, existe una sucesión de funciones simples $\varphi_k \geq 0$ tales que $\varphi_k \nearrow f$ puntualmente. Entonces $f \in \mathcal{F}$, y luego $\mathcal{F} = L^\infty_\mu(X)$, ya que toda función en $L^\infty_\mu(X)$ es combinación lineal de funciones no negativas. Por lo tanto $y \mapsto \mu_y$ es un núcleo de probabilidad.

Finalmente, veamos que se cumple (1.13). Es claro que la ecuación vale si $f \in V$. Además es fácil ver que el conjunto $\mathcal{G}$ de funciones medibles $f : X \to \mathbb{C}$ que cumplen (1.13) para toda $g$ es cerrado bajo convergencia monótona y combinaciones lineales. Utilizando un argumento similar al anterior, obtenemos que (1.13) se verifica para toda $f \in L^\infty_\mu(X)$. □

Llamamos al núcleo de probabilidad del teorema anterior la *desintegración* de $\mu$ relativa a $\pi$.

**Observación 1.40.** (*a*) Supongamos $y \mapsto \mu_y$ es la desintegración respecto a un factor $\pi : X \to Y$. Si $g \in L^\infty_\nu(Y)$, entonces para $\nu$-casi todo $y \in Y$ se cumple que

$$g \circ \pi(x) = g(y), \qquad \text{para } \mu_y\text{-casi todo } x.$$

En efecto, dada una función $h \in L^\infty_\nu(Y)$, tenemos que

$$\int_Y \left( \int_X f(g \circ \pi)\, d\mu_y \right) h(y)\, d\nu(y) = \int_X f(g \circ \pi)(h \circ \pi)\, d\mu = \int_Y \left( \int_X f g(y)\, d\mu_y \right) h(y)\, d\nu(y).$$

Como esto vale para toda $h$ se sigue que

$$\int_X f(g \circ \pi)\, d\mu_y = \int_X f g(y)\, d\mu_y$$

para $\nu$-casi todo $y$. Tomando $f = g \circ \pi - g(y)$ obtenemos que $\int_X (g \circ \pi - g(y))^2\, d\mu_y = 0$ para $\nu$-casi todo $y$, de donde se deduce la afirmación.

(*b*) Notemos además que si $E \in \mathcal{B}$ cumple $\mu(E) = 0$, aplicando (1.13) a $f = \mathbb{1}_E$, $g \equiv 1$, obtenemos que

$$\mu(E) = \int_Y \mu_y(E)\, d\nu(y),$$

así que $\mu_y(E) = 0$ para $\nu$-casi todo $y \in Y$. En otras palabras, si una propiedad se cumple $\mu$-c.t.p. entonces se cumple $\mu_y$-c.t.p. para $\nu$-casi todo $y$.

**Teorema 1.41** (Descomposición ergódica)**.** *Sea $(X, \mathcal{B}, \mu, T)$ un sistema dinámico tal que $(X, \mathcal{B})$ es regular. Entonces existe un espacio de probabilidad $(Y, \mathcal{D}, \nu)$ y un núcleo de probabilidad $y \mapsto \mu_y$ que cumple:*

(*i*) *$\mu_y$ es $T$-invariante y ergódica para $\nu$-casi todo $y \in Y$.*

(*ii*) *$\mu = \int_Y \mu_y\, d\nu(y)$, es decir,*

$$\mu(f) = \int_Y \mu_y(f)\, d\nu(y)$$

*para toda $f \in L^\infty_\mu(X)$.*

*Demostración.* Sea $Y = X$, sea $\mathcal{B}^T \subseteq \mathcal{B}$ la $\sigma$-álgebra de conjuntos medibles $T$-invariantes y sea $\nu = \mu|_{\mathcal{B}^T}$. Entonces la identidad $\pi : X \to Y$ es un factor de $X$. Afirmamos que la desintegración $y \mapsto \mu_y$ de $\pi$ cumple las propiedades del enunciado.

En primer lugar observemos que toda $g \in L^\infty(Y, \mathscr{B}^T, \nu)$ cumple $g \circ T = g$. En efecto, dado $x \in X$, sea $\alpha = g(Tx)$. El conjunto $B = g^{-1}(\alpha) \in \mathscr{B}^T$ es $T$-invariante, así que $x \in T^{-1}B = B$ y luego $g(x) = \alpha = g(Tx)$. Esto implica que el mapa $y \mapsto T_*\mu_y$ es una desintegración ya que

$$\begin{aligned}\int_Y \left(\int_X f \, dT_*\mu_y\right) g(y)\, d\nu(y) &= \int_Y \left(\int_X f \circ T \, d\mu_y\right) g(y)\, d\nu(y)\\ &= \int_X (f \circ T) g \, d\mu = \int_X (f \circ T)(g \circ T)\, d\mu = \int_X fg \, d\mu.\end{aligned}$$

Por el teorema 1.39 deducimos que $\mu_y = T_*\mu_y$ para $\nu$-casi todo $y$.

Al igual que antes podemos suponer que $X$ es un espacio métrico compacto y que $\mathscr{B}$ es la $\sigma$-álgebra de Borel. Afirmamos que para toda $f \in C(X)$ fija, $A_N f \to \mu_y(f)$ $\mu_y$-c.t.p. cuando $N \to \infty$, para $\nu$-casi todo $y \in Y$. En efecto, sabemos por el teorema ergódico puntual que $A_N f \to \mathbb{E}(f|\mathscr{B}^T)$ $\mu$-c.t.p. y luego se da la convergencia $\mu_y$-c.t.p. para $\nu$-casi todo $y$ por 1.40.(*b*). Ahora, $\mathbb{E}(f|\mathscr{B}^T) \in L^\infty(Y, \mathscr{B}^T, \nu)$, así que por 1.40.(*a*), $\mathbb{E}(f|\mathscr{B}^T)$ es igual a una constante $\mu_y$-c.t.p. para $\nu$-casi todo $y$. Luego para $\nu$-casi todo $y \in Y$ los promedios $A_N f$ convergen $\mu_y$-c.t.p. a una constante y luego esa constante debe ser $\mu_y(f)$.

Notemos que la inclusión $C(X) \hookrightarrow L^1_\mu(X)$ es un operador acotado ya que $\int |f|\, d\mu \le \|f\|_{\sup}$ para toda $f \in C(X)$. Como $C(X)$ es separable existe un subconjunto $E \subseteq (X)$ denso y numerable. Como $C(X)$ es denso en $L^1_\mu(X)$, se sigue que $E$ es denso en $L^1_\mu(X)$. Para cada $f \in E$ existe un conjunto $Z_f \subseteq Y$ con $\nu(Z_f) = 0$ tal que $A_N f \to \mu_y(f)$ $\mu_y$-c.t.p. para todo $y \in Y \setminus Z_f$. Si definimos $Z = \bigcup_{f \in E} Z_f$ entonces $A_N f \to \mu_y(f)$ $\mu_y$-c.t.p. para todo $y \in Y \setminus Z$ y para toda $f \in E$. Es fácil ver que entonces lo mismo vale para toda $f \in L^1_\mu(X)$ por densidad. Por la observación 1.32 deducimos que $\mu_y$ es ergódica para todo $y \in Y \setminus Z$. La segunda propiedad se deduce de que $y \mapsto \mu_y$ es una desintegración. □

## 1.7. Entropía

La entropía de un sistema dinámico es un invariante que mide qué tan "impredecible" es un sistema. Fue introducida por Kolmogorov en 1958 para probar que distintos esquemas de Bernoulli no son isomorfos. La definición usada actualmente sin embargo es la que propuso Sinai en 1959. Uno de los problemas principales en la teoría ergódica es la clasificación de los sistemas dinámicos salvo isomorfismo, y uno de los mayores logros de la teoría ergódica en este sentido es el teorema de Ornstein, que dice que la entropía es un invariante completo para los esquemas de Bernoulli.

**Definición 1.42.** Sea $(X, \mathscr{B}, \mu)$ un espacio de probabilidad.

(1) Una *partición* es una colección $\mathscr{P}$ de elementos de $\mathscr{B}$ disjuntos cuya unión es $X$.

(2) Si $\mathscr{P}$ y $\mathscr{Q}$ son particiones, decimos que $\mathscr{Q}$ es un *refinamiento* de $\mathscr{P}$ (notado $\mathscr{Q} \succ \mathscr{P}$) si para todo $Q \in \mathscr{Q}$ existe $P \in \mathscr{P}$ tal que $\mu(Q \setminus P) = 0$. Decimos que $\mathscr{P}$ y $\mathscr{Q}$ son *equivalentes* (notado $\mathscr{P} \mathrel{\mathring{=}} \mathscr{Q}$) si $\mathscr{P} \succ \mathscr{Q}$ y $\mathscr{Q} \succ \mathscr{P}$.

(3) Si $\mathscr{P}$ y $\mathscr{Q}$ son particiones, su *supremo* es la partición

$$\mathscr{P} \vee \mathscr{Q} = \{P \cap Q : P \in \mathscr{P}, Q \in \mathscr{Q}\}.$$

(4) Si $\mathscr{P} = \{A_1, \dots, A_m\}$ es una partición finita, definimos la *entropía* de $\mathscr{P}$ como

$$H(\mathscr{P}) = -\sum_{i=1}^m \mu(A_i) \log \mu(A_i), \tag{1.14}$$

con la convención de que $x \log x = 0$ si $x = 0$. Ocasionalmente usaremos la notación $H_\mu(\mathscr{P})$ para especificar la medida.

(5) Si $\mathscr{P} = \{A_1, \dots, A_m\}$ y $\mathscr{Q} = \{B_1, \dots, B_n\}$ son particiones finitas, definimos la *entropía condicional* $H(\mathscr{P}|\mathscr{Q})$ como

$$H(\mathscr{P}|\mathscr{Q}) = -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(B_j)}. \tag{1.15}$$

En esta suma se deben omitir aquellos términos en los que $\mu(B_j) = 0$.

(6) Si $T : X \to X$ es una función medible, definimos

$$T^{-1}\mathscr{P} = \{T^{-1}(P) : P \in \mathscr{P}\}.$$

Si uno interpreta una partición $\mathscr{P} = \{A_1, \dots, A_n\}$ como una lista de los posibles resultados de un experimento, donde cada resultado $A_i$ tiene probabilidad $\mu(A_i)$ de ocurrir, entonces la entropía $H(\mathscr{P})$ es la esperanza de la función $-\log \mu(A_i)$, que mide la cantidad "información" que representa el conjunto $A_i$. Podemos justificar esto de la siguiente manera: si $A$ es un conjunto de medida $2^{-n}$ y particionamos a $X$ en $2^n$ conjuntos de igual medida, uno de los cuales es $A$, entonces hacen falta $n = -\log_2 \mu(A)$ dígitos binarios para especificar a $A$ entre los otros. Alternativamente, se puede ver que la función $H(p_1, \dots, p_n) = -\sum_{j=1}^{n} p_j \log p_j$ es la única (salvo por una constante multiplicativa) que cumple una serie de condiciones naturales [Wal00, Teo. 4.1]. Análogamente, la entropía condicional $H(\mathscr{P}|\mathscr{Q})$ representa la cantidad de información que uno obtiene al realizar el experimento asociado a $\mathscr{P}$ habiendo realizado antes el experimento asociado a $\mathscr{Q}$.

**Observación 1.43.** Si $\mathscr{P} = \{A_1, \dots, A_n\}$ una partición finita en un espacio de probabilidad $(X, \mathscr{B}, \mu)$, entonces podemos acotar su entropía por

$$H(\mathscr{P}) \geq -\log\left(\sum_{i=1}^{n} \mu(A_i)^2\right). \tag{1.16}$$

Más en general, vale que si $x_1, \dots, x_n \in (0,1)$ son tales que $x_1 + \cdots + x_n = 1$, entonces

$$-\sum_{i=1}^{n} x_i \log x_i \geq -\log\left(\sum_{i=1}^{n} x_i^2\right). \tag{1.17}$$

Esto se debe a que la función $f : (0,1) \to \mathbb{R}$, $f(x) = -\log x$ es convexa.

**Proposición 1.44.** *Sean $\mathscr{P}$, $\mathscr{Q}$, $\mathscr{R}$ particiones finitas de $X$.*

*(a) $H(\mathscr{P} \vee \mathscr{Q}) = H(\mathscr{Q}) + H(\mathscr{P}|\mathscr{Q})$.*
*(b) Si $\mathscr{P} \prec \mathscr{Q}$ entonces $H(\mathscr{P}) \leq H(\mathscr{Q})$ y $H(\mathscr{P}|\mathscr{R}) \leq H(\mathscr{Q}|\mathscr{R})$.*
*(c) $H(\mathscr{P}|\mathscr{Q}) \leq H(\mathscr{P})$ y $H(\mathscr{P} \vee \mathscr{Q}) \leq H(\mathscr{P}) + H(\mathscr{Q})$.*
*(d) Si $T : X \to X$ preserva medida entonces $H(T^{-1}\mathscr{P}) = H(\mathscr{P})$.*
*(e) $H(\mathscr{P}) \leq \operatorname{card}_\mu(\mathscr{P})$, donde $\operatorname{card}_\mu(\mathscr{P})$ es la cantidad de elementos de $\mathscr{P}$ con medida positiva.*

*Demostración.* (a) Supongamos que $\mathscr{P} = \{A_1, \dots, A_m\}$ y $\mathscr{Q} = \{B_1, \dots, B_n\}$. Entonces $\mathscr{P} \vee \mathscr{Q} = \{A_i \cap B_j : 1 \leq i \leq m, 1 \leq j \leq n\}$ así que[2]

$$\begin{aligned}
H(\mathscr{P} \vee \mathscr{Q}) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap B_j) \log \mu(A_i \cap B_j) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap B_j) \left(\log \frac{\mu(A_i \cap B_j)}{\mu(B_j)} + \log \mu(B_j)\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(B_j)} + \sum_{j=1}^{n} \log \mu(B_j) \sum_{i=1}^{n} \mu(A_i \cap B_j) \\
&= H(\mathscr{P}|\mathscr{Q}) + \sum_{j=1}^{n} \mu(B_j) \log \mu(B_j) = H(\mathscr{P}|\mathscr{Q}) + H(\mathscr{Q}).
\end{aligned}$$

[2]Notar que en la cuenta que sigue, aquellos términos en los que $\mu(A_i \cap B_j) = 0$ no contribuyen a la suma en ningún paso, así que podemos ignorarlos.

(*b*) Si $\mathscr{P} \prec \mathscr{Q}$ entonces $H(\mathscr{Q}) = H(\mathscr{P} \vee \mathscr{Q}) = H(\mathscr{P}) + H(\mathscr{Q}|\mathscr{P}) \geq H(\mathscr{P})$. Además,

$$H(\mathscr{P}|\mathscr{R}) = H(\mathscr{P} \vee \mathscr{R}) - H(\mathscr{R}) \leq H(\mathscr{Q} \vee \mathscr{R}) - H(\mathscr{R}) = H(\mathscr{Q}|\mathscr{R}).$$

(*c*) Sea $\phi : [0,1] \to \mathbb{R}$ la función dada por

$$\phi(x) = \begin{cases} -x \log x, & \text{si } x \neq 0, \\ 0, & \text{si } x = 0. \end{cases} \tag{1.18}$$

Esta función es cóncava en $(0,1]$ ya que $\phi''(x) = -\frac{1}{x} < 0$. Luego

$$\phi(tx + (1-t)y) \geq t\phi(x) + (1-t)\phi(y)$$

para todos $x, y \in (0,1]$ y $t \in [0,1]$. Notemos que la desigualdad anterior también vale si $y = 0$ ya que $\phi(tx) = -tx \log(tx) \geq -tx \log x = t\phi(x)$, puesto que $tx \leq x$. Por inducción tenemos que $\phi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \phi(x_i)$ siempre que $x_1, \ldots, x_n, p_1, \ldots, p_n \in [0,1]$ cumplen $p_1 + \cdots + p_n = 1$. Si $\mathscr{P} = \{A_1, \ldots, A_m\}$ y $\mathscr{Q} = \{B_1, \ldots, B_n\}$ son particiones, entonces

$$\begin{aligned} H(\mathscr{P}|\mathscr{Q}) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(B_j)} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mu(B_j) \phi\left(\frac{\mu(A_i \cap B_j)}{\mu(B_j)}\right) \\ &\leq \sum_{i=1}^{m} \phi\left(\sum_{j=1}^{n} \mu(B_j) \frac{\mu(A_i \cap B_j)}{\mu(B_j)}\right) \\ &= \sum_{i=1}^{m} \phi(\mu(A_i)) = H(\mathscr{P}). \end{aligned}$$

La segunda desigualdad se deduce de la primera y del punto (*a*).

(*d*) Se deduce del hecho que $\mu(T^{-1}A) = \mu(A)$ para todo elemento $A \in \mathscr{P}$.

(*e*) Podemos suponer que $\mathscr{P} = \{A_1, \ldots, A_m\}$ con $\mu(A_i) > 0$ para todo $i$ ya que los conjuntos de medida 0 no contribuyen a la entropía. Como la función $\phi$ definida antes es cóncava, tenemos que

$$\frac{1}{m} H(\mathscr{P}) = \frac{1}{m} \sum_{i=1}^{m} \phi(\mu(A_i)) \leq \phi\left(\frac{1}{m} \sum_{i=1}^{m} \mu(A_i)\right) = \phi(\tfrac{1}{m}) = \frac{1}{m} \log m.$$

Luego $H(\mathscr{P}) \leq \log m$. □

Definimos

$$\mathscr{P}^{(n)} = \bigvee_{k=0}^{n-1} T^{-k}\mathscr{P}. \tag{1.19}$$

**Definición 1.45.** Sea $(X, \mathscr{B}, \mu)$ un espacio de probabilidad y sea $T : X \to X$ una transformación que preserva medida. Definimos la *entropía* de $T$ con respecto a $\mu$ y una partición medible $\mathscr{P}$ como

$$h_\mu(T, \mathscr{P}) = \lim_{n\to\infty} \frac{1}{n} H(\mathscr{P}^{(n)}). \tag{1.20}$$

Veamos que este límite siempre existe:

**Lema 1.46.** *Si $(a_n)_{n\geq 0}$ es una sucesión de reales positivos tales que $a_{n+m} \leq a_n + a_m$ entonces $\lim_{n\to\infty} \frac{a_n}{n}$ existe y es igual a $\inf_{n\in\mathbb{N}} \frac{a_n}{n}$.*

*Demostración.* Fijemos $b \in \mathbb{N}$. Todo $n \in \mathbb{N}$ se puede expresar como $n = qb + r$ con $q \in \mathbb{N}_0$ y $r \in \{0, \dots, b-1\}$. Entonces

$$\frac{a_n}{n} \leq \frac{qa_b + a_r}{qb + r} = \frac{a_b + a_r/q}{b + r/q}$$

A medida que $n \to \infty$, $q = \lfloor \frac{n}{b} \rfloor$ también tiende a $\infty$ y luego

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_b}{b}.$$

Como $b$ es arbitrario, esto implica que $\limsup_{n\to\infty} \frac{a_n}{n} \leq \liminf_{n\to\infty} \frac{a_n}{n}$ y por lo tanto el límite existe. □

Notemos que $(H(\mathscr{P}^{(n)}))_{n\in\mathbb{N}}$ es una sucesión subaditiva ya que

$$H(\mathscr{P}^{(n+m)}) = H(\mathscr{P}^{(n)} \vee T^{-n}\mathscr{P}^{(m)}) \leq H(\mathscr{P}^{(n)}) + H(T^{-n}\mathscr{P}^{(m)}) = H(\mathscr{P}^{(n)}) + H(\mathscr{P}^{(m)}).$$

Entonces por el lema 1.46, $h_\mu(T, \mathscr{P})$ está bien definida. La *entropía* $h_\mu(T)$ del sistema dinámico $(X, \mathscr{B}, \mu, T)$ se define como

$$h_\mu(T) = \sup\{h_\mu(T, \mathscr{P}) : \mathscr{P} \text{ es una partición medible de } X\}. \tag{1.21}$$

Si pensamos a la transformación $T$ como el paso de un día, entonces $H(\mathscr{P}^{(n)})$ mide la cantidad de información que se obtiene de realizar el experimento asociado a la partición $\mathscr{P}$ en $n$ días consecutivos. Luego $h_\mu(T, \mathscr{P})$ se puede interpretar como la cantidad de información promedio por día que proporciona asintóticamente el experimento.

Ahora definimos la entropía de una subálgebra finita siguiendo a [Wal00]. Si $\mathscr{P}$ es una partición finita, entonces la $\sigma$-álgebra que genera $\mathscr{A} = \sigma(\mathscr{P})$ es un álgebra finita y los elementos de $\mathscr{P}$ son los conjuntos no vacíos de $\mathscr{A}$ que son minimales con respecto a la inclusión. Recíprocamente, dada un álgebra finita $\mathscr{A} \subseteq \mathscr{B}$, sus elementos minimales no vacíos forman una partición finita $\mathscr{P} = \mathscr{P}(\mathscr{A})$ tal que $\sigma(\mathscr{P}) = \mathscr{A}$. Luego hay una correspondencia 1-1 entre particiones finitas y subálgebras finitas de $\mathscr{B}$. En particular podemos definir la entropía de una subálgebra como $H(\mathscr{A}) = H(\mathscr{P}(\mathscr{A}))$, y análogamente $H(\mathscr{A}|\mathscr{C}) = H(\mathscr{P}(\mathscr{A})|\mathscr{P}(\mathscr{C}))$. Además definimos $\mathscr{A} \vee \mathscr{C}$ como la $\sigma$-álgebra generada por $\mathscr{A} \cup \mathscr{C}$ y $\mathscr{A}^{(n)} = \bigvee_{k=0}^{n-1} T^{-k}\mathscr{A}$. Finalmente, si $a < b$ son enteros entonces $\mathscr{A}^{[a,b]} := \bigvee_{k=a}^{b} T^{-k}\mathscr{A}$. La correspondencia entre particiones y álgebras respeta todas las operaciones y relaciones definidas. La ventaja de usar álgebras es que nos permite generalizar la entropía condicional. Sean $\mathscr{A}$ y $\mathscr{C}$ las álgebras asociadas a las particiones $\{A_1, \dots, A_m\}$ y $\{C_1, \dots, C_n\}$ respectivamente. Afirmamos que $H(\mathscr{A}|\mathscr{C}) = \int I_{\mathscr{A}|\mathscr{C}}\, d\mu$, donde

$$I_{\mathscr{A}|\mathscr{C}} := -\sum_{i=1}^{m} \chi_{A_i} \log \mathbb{E}(\chi_{A_i}|\mathscr{C}) \tag{1.22}$$

es la *función de información condicional* de $\mathscr{A}$ con respecto a $\mathscr{C}$. En efecto,

$$\begin{aligned}
\int I_{\mathscr{A}|\mathscr{C}}\, d\mu &= -\sum_{i=1}^{m} \int_{A_i} \log \mathbb{E}(\chi_{A_i}|\mathscr{C})\, d\mu \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} \int_{A_i\cap C_j} \log \mathbb{E}(\chi_{A_i}|\mathscr{C})\, d\mu \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} \int_{A_i\cap C_j} \log\left(\frac{1}{\mu(C_j)}\int_{C_j} \chi_{A_i}\, d\mu\right) d\mu \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} \mu(A_i \cap C_j) \log \frac{\mu(A_i \cap C_j)}{\mu(C_j)} = H(\mathscr{A}|\mathscr{C}).
\end{aligned}$$

Entonces tiene sentido definir $H(\mathscr{A}|\mathscr{C})$ en el caso que $\mathscr{C} \subseteq \mathscr{B}$ es una $\sigma$-álgebra arbitraria como $H(\mathscr{A}|\mathscr{C}) = \int I_{\mathscr{A}|\mathscr{C}}\, d\mu$, donde $I_{\mathscr{A}|\mathscr{C}}$ está definida por (1.22).

**Proposición 1.47.** *Sean $\mathscr{A}, \mathscr{C} \subseteq \mathscr{B}$ subálgebras finitas en un espacio de probabilidad $(X, \mathscr{B}, \mu)$ y supongamos que $T : X \to X$ preserva medida. Entonces*

(a) *$\mathscr{A} \subseteq \mathscr{C}$ implica $h_\mu(T, \mathscr{A}) \leq h_\mu(T, \mathscr{C})$.*

(b) *$h_\mu(T, \mathscr{C}) \leq h_\mu(T, \mathscr{A}) + H(\mathscr{C}|\mathscr{A})$.*

(c) *$h_\mu(T, T^{-1}\mathscr{A}) = h_\mu(T, \mathscr{A})$.*

(d) *$h_\mu(T, \mathscr{A}) = h_\mu(T, \mathscr{A}^{(k)})$ para todo $k \geq 1$.*

(e) *Si $T$ es inversible, entonces $h_\mu(T, \mathscr{A}) = h_\mu(T, \mathscr{A}^{[-k,k]})$ para todo $k \geq 1$.*

(f) *$h_\mu(T^k) = |k| h_\mu(T)$.*

La demostración de estas propiedades es rutinaria, y referimos al lector a [Wal00, Teoremas 4.12 y 4.13] para su demostración.

**Proposición 1.48.** *Sea $(X, \mathscr{B}, \mu)$ un espacio de probabilidad, sea $\{\mathscr{F}_n\}_{n\in\mathbb{N}}$ una sucesión creciente de sub-$\sigma$-álgebras de $\mathscr{B}$ y sea $\mathscr{F} = \bigvee_{n=1}^{\infty} \mathscr{F}_n$. Para toda $f \in L^2_\mu(X)$ se tiene*

$$\|\mathbb{E}(f|\mathscr{F}_n) - \mathbb{E}(f|\mathscr{F})\|_2 \longrightarrow 0. \tag{1.23}$$

*Demostración.* Sea $B \in \mathscr{F}$. Como $\mathscr{F}$ es la $\sigma$-álgebra generada por $\bigcup_{n=1}^{\infty} \mathscr{F}_n$, por el teorema de aproximación existen conjuntos $B_n \in \mathscr{F}_n$ tales que $\mu(B \triangle B_n) \to 0$. Como $\mathbb{E}(\chi_B|\mathscr{F}_n)$ es el elemento de $L^2(X, \mathscr{F}_n, \mu)$ más cercano a $\chi_B$ en norma, tenemos

$$\|\mathbb{E}(\chi_B|\mathscr{F}_n) - \chi_B\|_2 \leq \|\chi_{B_n} - \chi_B\|_2 = \mu(B \triangle B_n)^{1/2} \to 0.$$

Como las funciones $\chi_B$, $B \in \mathscr{F}$ generan $L^2(X, \mathscr{F}, \mu)$ esto implica que $\|\mathbb{E}(h|\mathscr{F}_n) - h\|_2 \to 0$ para toda $h \in L^2(X, \mathscr{F}, \mu)$. Si aplicamos esto a $h = \mathbb{E}(f|\mathscr{F})$ donde $f \in L^2(X, \mathscr{B}, \mu)$, obtenemos (1.23) ya que $\mathbb{E}(\mathbb{E}(f|\mathscr{F})|\mathscr{F}_n) = \mathbb{E}(f|\mathscr{F}_n)$. □

**Corolario 1.49.** *Bajo las condiciones de la proposición anterior, si $\mathscr{A}$ es una subálgebra finita de $\mathscr{B}$ entonces $H(\mathscr{A}|\mathscr{F}_n) \to H(\mathscr{A}|\mathscr{F})$.*

*Demostración.* Sea $\mathscr{P} = \{A_1, \ldots, A_m\}$ la partición asociada a $\mathscr{A}$. Por la proposición anterior, $\|\mathbb{E}(\chi_{A_i}|\mathscr{F}_n) - \mathbb{E}(\chi_{A_i}|\mathscr{F})\|_2 \to 0$ para todo $i$. En particular $\mathbb{E}(\chi_{A_i}|\mathscr{F}_n) \to \mathbb{E}(\chi_{A_i}|\mathscr{F})$ c.t.p. y luego $\phi \circ \mathbb{E}(\chi_{A_i}|\mathscr{F}_n) \to \phi \circ \mathbb{E}(\chi_{A_i}|\mathscr{F})$ c.t.p., donde $\phi$ es la función definida en (1.18). Además $\|\mathbb{E}(\chi_{A_i}|\mathscr{F}_n)\|_\infty \leq \|\chi_A\|_\infty = 1$. Por convergencia acotada, tenemos que

$$\int \mathbb{E}(\chi_{A_i}|\mathscr{F}_n) \log \mathbb{E}(\chi_{A_i}|\mathscr{F}_n)\, d\mu \longrightarrow \int \mathbb{E}(\chi_{A_i}|\mathscr{F}) \log \mathbb{E}(\chi_{A_i}|\mathscr{F})\, d\mu.$$

Finalmente notemos que $\int \mathbb{E}(\chi_{A_i}|\mathscr{F}_n) \log \mathbb{E}(\chi_{A_i}|\mathscr{F}_n)\, d\mu = \int \chi_{A_i} \log \mathbb{E}(\chi_{A_i}|\mathscr{F}_n)\, d\mu$ por 1.10, ya que $\log \mathbb{E}(\chi_{A_i}|\mathscr{F}_n)$ es una función medible con respecto a $\mathscr{F}_n$, y lo mismo vale reemplazando $\mathscr{F}_n$ por $\mathscr{F}$. Se sigue que

$$H(\mathscr{A}|\mathscr{F}_n) = -\sum_{i=1}^{m} \int \chi_{A_i} \log \mathbb{E}(\chi_{A_i}|\mathscr{F}_n)\, d\mu \longrightarrow -\sum_{i=1}^{m} \int \chi_{A_i} \log \mathbb{E}(\chi_{A_i}|\mathscr{F})\, d\mu = H(\mathscr{A}|\mathscr{F}). \qquad \square$$

**Teorema 1.50.** *Si $\mathscr{A} \subseteq \mathscr{B}$ es una subálgebra finita, entonces*

$$h_\mu(T, \mathscr{A}) = \lim_{n\to\infty} H\left(\mathscr{A} \middle| \bigvee_{k=1}^{n} T^{-k}\mathscr{A}\right) = H\left(\mathscr{A} \middle| \bigvee_{k=1}^{\infty} T^{-k}\mathscr{A}\right). \tag{1.24}$$

*Demostración.* La segunda igualdad se deduce del corolario 1.49. Luego basta probar la primera igualdad. Notemos que

$$\begin{aligned} H(\mathscr{A}^{(n+1)}) &= H\left(\mathscr{A} \vee T^{-1}\mathscr{A}^{(n)}\right) \\ &= H(T^{-1}\mathscr{A}^{(n)}) + H\left(\mathscr{A} \middle| T^{-1}\mathscr{A}^{(n)}\right) \\ &= H(\mathscr{A}^{(n)}) + H\left(\mathscr{A} \middle| T^{-1}\mathscr{A}^{(n)}\right). \end{aligned}$$

Entonces, por inducción

$$H(\mathscr{A}^{(n)}) = H(\mathscr{A}) + \sum_{j=1}^{n-1} H\left(\mathscr{A} \middle| T^{-1}\mathscr{A}^{(j)}\right). \tag{1.25}$$

Ahora dividimos por $n$ en (1.25) y hacemos $n \to \infty$. Como la sucesión $H\left(\mathscr{A} \middle| T^{-1}\mathscr{A}^{(n)}\right)$ tiende a $H\left(\mathscr{A} \middle| \bigvee_{k=1}^{\infty} T^{-k}\mathscr{A}\right)$, sus sumas de Cèsaro tienden al mismo límite. Por lo tanto $h_\mu(T, \mathscr{A}) = H(\mathscr{A} \big| \bigvee_{k=1}^{\infty} T^{-k}\mathscr{A})$. □

**Definición 1.51.** Sea $(X, \mathscr{B}, \mu, T)$ un sistema dinámico inversible y sea $\mathscr{A} \subseteq \mathscr{B}$ una $\sigma$-álgebra. Decimos que $\mathscr{A}$ es un *generador* (con respecto a $T$) si $\bigvee_{k\in\mathbb{Z}} T^k\mathscr{A} \doteq \mathscr{B}$.

El siguiente teorema da un método para calcular la entropía de un sistema dinámico, si uno puede hallar generadores:

**Teorema 1.52** (Kolmogorov-Sinai)**.** *Sea $T : X \to X$ una transformación inversible que preserva medida en un espacio de probabilidad $(X, \mathscr{B}, \mu)$ y sea $\mathscr{A} \subseteq \mathscr{B}$ una subálgebra finita. Si $\mathscr{A} \subseteq \mathscr{B}$ es un generador entonces $h_\mu(T) = h_\mu(T, \mathscr{A})$.*

*Demostración.* Sea $\mathscr{C} \subseteq \mathscr{B}$ un álgebra finita. Queremos ver que $h_\mu(T, \mathscr{C}) \leq h_\mu(T, \mathscr{A})$. Por 1.47, tenemos $h_\mu(T, \mathscr{C}) \leq h_\mu(T, \mathscr{A}^{[-n,n]}) + H(\mathscr{C}|\mathscr{A}^{[-n,n]}) = h_\mu(T, \mathscr{A}) + H(\mathscr{C}|\mathscr{A}^{[-n,n]})$. Como $\mathscr{A}$ es generador, $H(\mathscr{C}|\mathscr{A}^{[-n,n]}) \to H(\mathscr{C}|\mathscr{B})$. Ahora, $H(\mathscr{C}|\mathscr{B}) = 0$ ya que si $\mathscr{C} = \{C_1, \ldots, C_m\}$ entonces

$$H(\mathscr{C}|\mathscr{B}) = -\sum_{i=1}^{m} \int_{C_i} \log \mathbb{E}(\chi_{C_i}|\mathscr{B})\, d\mu = -\sum_{i=1}^{m} \int_{C_i} \log \chi_{C_i}\, d\mu = 0.$$

Luego, haciendo $n \to \infty$ resulta $h_\mu(T, \mathscr{C}) \leq h_\mu(T, \mathscr{A})$. Por lo tanto $h_\mu(T) = h_\mu(T, \mathscr{A})$. □

El teorema anterior nos permite calcular la entropía de un esquema de Bernoulli bilateral asociado a un espacio de probabilidad $(\Omega, \mathscr{F}, \mu)$ donde $\Omega$ es finito y $\mathscr{F} = \mathscr{P}(\Omega)$. Por ejemplo, si $\Omega = \{w_1, \ldots, w_r\}$ con $\mu(\{w_i\}) = p_i$, entonces la partición $\mathscr{P} = \{A_1, \ldots, A_r\}$, donde $A_i = \{x \in \Omega^{\mathbb{N}} : x_1 = w_i\}$, genera $\bigotimes_{n=1}^{\infty} \mathscr{F}$, así que $h_\mu(S) = h_\mu(S, \mathscr{P})$. Además es fácil ver que $H(\mathscr{P}^{(n)}) = nH(\mathscr{P})$ ya que las particiones $\mathscr{P}, S^{-1}\mathscr{P}, \ldots, S^{-n+1}\mathscr{P}$ son independientes entre sí. Luego

$$h_\mu(S) = H(\mathscr{P}) = -\sum_{j=1}^{r} p_j \log p_j.$$

**Teorema 1.53.** *Sea $(X, \mathscr{B}, \mu, T)$ un sistema dinámico medible y sea $\mathscr{A} \subseteq \mathscr{B}$ una subálgebra finita. Si*

$$\mu = \int_Y \mu_y\, d\nu(y)$$

*es la descomposición ergódica de $\mu$, entonces*

(a) $h_\mu(T, \mathscr{A}) = \int_Y h_{\mu_y}(T, \mathscr{A})\, d\nu(y)$.
(b) $h_\mu(T) = \int_Y h_{\mu_y}(T)\, d\nu(y)$.

*Demostración.* Ver [Gla03, Teo. 15.12]. □

# Capítulo 2

# El flujo geodésico

El flujo geodésico en variedades riemannianas es un sistema dinámico muy estudiado en teoría ergódica. El teorema de Duke en su segunda versión habla de la distribución de una serie de medidas en el campo tangente unitario de la curva modular $Y_0(1)$. Resulta que $Y_0(1)$ tiene una estructura natural de variedad riemanniana y que estas medidas son invariantes bajo el respectivo flujo geodésico. Esto nos permitirá utilizar los resultados de teoría ergódica vistos en el capítulo anterior para probar el teorema.

En primer lugar describiremos la conexión entre la equidistribución de las formas cuadráticas de discriminante $d$ y la equidistribución de $\mathcal{G}_d$[1] en $T^1Y_0(1)$. Esta conexión está basada en la proposición 1.20. Recordemos que $R_{\text{disc}}(d)$ es el conjunto de formas cuadráticas binarias enteras primitivas de discriminante $d$. El teorema de Skubenko habla de la distribución de $d^{-1/2}R_{\text{disc}}(d)$ en el espacio

$$V^+_{\text{disc}}(\mathbb{R}) = \{(a,b,c) \in \mathbb{R}^3 : b^2 - 4ac = 1\}. \tag{2.1}$$

Esta superficie es un hiperboloide de revolución. Identificamos a $V^+_{\text{disc}}(\mathbb{R})$ con el conjunto de formas cuadráticas reales $q(x,y) = ax^2 + bxy + cy^2$ de discriminante 1. El grupo $\mathrm{GL}_2(\mathbb{R})$ actúa a izquierda en $V^+_{\text{disc}}(\mathbb{R})$ vía

$$g \cdot q(x,y) = \frac{1}{\det(g)} q((x,y)g).$$

El subgrupo $\mathbb{R}^\times \mathrm{I}$ actúa trivialmente, así que la acción pasa al cociente $\mathrm{PGL}_2(\mathbb{R}) = \mathrm{GL}_2(\mathbb{R})/\mathbb{R}^\times$. Por comodidad escribiremos los elementos de $\mathrm{PGL}_2(\mathbb{R})$ y $\mathrm{PSL}_2(\mathbb{R})$ como matrices, sin hacer distinción notacional con un elemento de $\mathrm{GL}_2(\mathbb{R})$. Afirmamos que la acción es transitiva. En efecto, dada una forma cuadrática $q(x,y) = ax^2 + bxy + cy^2$ en $V^+_{\text{disc}}(\mathbb{R})$, como $\mathrm{disc}(q) > 0$ el polinomio $q(\,\cdot\,, 1)$ tiene raíces reales $\tau_1 < \tau_2$ y $q$ se factoriza como $q = a(x - \tau_1 y)(x - \tau_2 y)$. Si $a > 0$ entonces la matriz $g = \left(\begin{smallmatrix} a & 1 \\ -a\tau_2 & -\tau_1 \end{smallmatrix}\right)$ tiene determinante 1 y manda $q_0(x,y) = xy$ a $q$:

$$q_0\left((x,y)\begin{pmatrix} a & 1 \\ -a\tau_2 & -\tau_1 \end{pmatrix}\right) = a(x - \tau_2 y)(x - \tau_1 y) = q(x,y).$$

Si $a < 0$ entonces podemos tomar $g = \left(\begin{smallmatrix} a & 1 \\ -a\tau_1 & -\tau_2 \end{smallmatrix}\right)$ en cambio. Ahora calculemos el estabilizador $\mathrm{Stab}(q_0)$ de $q_0$ en $\mathrm{PGL}_2(\mathbb{R})$. Supongamos que $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PGL}_2(\mathbb{R})$ fija $q_0$. Multiplicando a la matriz que representa a $g$ por un escalar $\lambda > 0$, podemos suponer sin pérdida de generalidad que $ad - bc = \pm 1$. Si $ad - bc = 1$, entonces

$$g \cdot q_0(x,y) = q_0(ax + cy, bx + dy) = (ax + cy)(bx + dy) = abx^2 + (ad + bc)xy + cdy^2.$$

Luego $ab = cd = 0$ y $ad + bc = 1$. La última ecuación combinada con $ad - bc = 1$ implica que $ad = 1$ y $bc = 0$. Entonces $b = c = 0$ y $d = a^{-1}$, así que $g = \left(\begin{smallmatrix} a & 0 \\ 0 & 1/a \end{smallmatrix}\right)$ para algún $a \neq 0$. Si en cambio $ad - bc = -1$, entonces $g \cdot q_0(x,y) = -[abx^2 + (ad + bc)xy + cdy^2]$ y por un argumento análogo al anterior deducimos que $ad = -1$ y $bc = 0$, en cuyo caso $g = \left(\begin{smallmatrix} a & 0 \\ 0 & -1/a \end{smallmatrix}\right)$. Definimos

$$A = \left\{ \begin{pmatrix} a & 0 \\ 0 & \pm a^{-1} \end{pmatrix} : a \in \mathbb{R}^\times \right\} \subseteq \mathrm{GL}_2(\mathbb{R}) \tag{2.2}$$

[1] Este conjunto fue definido en la introducción, ver (3).

y por abuso de notación también denotamos por $A$ a su imagen en $G = \mathrm{PGL}_2(\mathbb{R})$. Hemos visto que $\mathrm{Stab}(q_0) = A$, así que $V^+_{\mathrm{disc}}(\mathbb{R}) = \mathrm{PGL}_2(\mathbb{R}).q_0 \cong \mathrm{PGL}_2(\mathbb{R})/A$.

Sea $d > 0$ tal que $d \equiv 0, 1 \pmod 4$. Observemos que $d^{-1/2} R_{\mathrm{disc}}(d) \subseteq V^+_{\mathrm{disc}}(\mathbb{R})$, así que a cada elemento $q \in R_{\mathrm{disc}}(d)$ con $d > 0$ le corresponde una coclase $g_q A \in \mathrm{PGL}_2(\mathbb{R})/A$, donde $g_q \in \mathrm{GL}_2(\mathbb{R})$ es una matriz tal que $g_q.q_0 = d^{-1/2} q$. El conjunto $d^{-1/2} R_{\mathrm{disc}}(d)$ es invariante bajo la acción de $\Gamma = \mathrm{PGL}_2(\mathbb{Z})$, así que es una unión disjunta de $\Gamma$-órbitas. Denotamos por $C(d)$ al conjunto de clases de equivalencia de $R_{\mathrm{disc}}(d)$ bajo la acción de $\Gamma$. Aunque $R_{\mathrm{disc}}(d)$ es infinito, veremos en la observación 3.19 que $C(d)$ es un conjunto finito.

El grupo $\mathrm{GL}_2(\mathbb{R})$ es unimodular, como vimos en el ejemplo 1.14. En particular su medida de Haar $\mu_0$ es $\mathbb{R}^\times$-invariante a derecha. Por 1.20, $\mu_0$ induce una medida $\nu_0$ en $\mathrm{GL}_2(\mathbb{R})/\mathbb{R}^\times$ que es $\mathrm{GL}_2(\mathbb{R})$-invariante a izquierda. A su vez $\mu_0$ induce una medida $\nu_1$ en $\mathbb{R}^\times \backslash \mathrm{GL}_2(\mathbb{R})$ que es $\mathrm{GL}_2(\mathbb{R})$-invariante a derecha. Ahora, como $\mathbb{R}^\times$ está en el centro de $\mathrm{GL}_2(\mathbb{R})$ estos dos espacios son isomorfos de manera canónica y es fácil ver que $\nu_0 = \nu_1$ bajo esta identificación. Por lo tanto $G$ también es unimodular.

Notemos que $A$ y $\Gamma$ son unimodulares, ya que $A$ es abeliano y $\Gamma$ es discreto[2]. Entonces por la proposición 1.20 hay una correspondencia entre medidas de Radon $\Gamma$-invariantes a izquierda en $G/A$ y medidas de Radon $A$-invariantes a derecha en $\Gamma\backslash G$. Usando este resultado, podemos trasladar el problema sobre la equidistribución de $d^{-1/2} R_{\mathrm{disc}}(d)$ en $V^+_{\mathrm{disc}}(\mathbb{R})$ a un problema de equidistribución de medidas en $\Gamma\backslash G$. Tomamos como medida de Haar $\lambda_\Gamma$ en $\Gamma$ la medida de contar y fijamos una medida de Haar a izquierda $\lambda_A$ en $A$. Además elegimos una medida de Haar $\mu_{\mathrm{disc}}$ para $V^+_{\mathrm{disc}}(\mathbb{R})$[3]. Como $G$ es unimodular, su medida de Haar a izquierda es $A$-invariante a derecha. La medida asociada en $\Gamma\backslash G$ es la medida de Haar $\mu_{\Gamma\backslash G}$, y la medida asociada en $G/A \cong V^+_{\mathrm{disc}}(\mathbb{R})$ es una medida de Haar a izquierda en $G/A$ y por eso difiere de $\mu_{\mathrm{disc}}$ solo en una constante multiplicativa. Entonces podemos normalizar a $\mu_{\Gamma\backslash G}$ de manera que se corresponda con $\mu_{\mathrm{disc}}$ bajo la correspondencia de la proposición 1.20.

Sea $\Gamma_q$ el estabilizador de $q$ en $\Gamma$. En vista del homeomorfismo $V^+_{\mathrm{disc}}(\mathbb{R}) = \mathrm{PGL}_2(\mathbb{R}).q_0 \cong \mathrm{PGL}_2(\mathbb{R})/A$, un elemento $\gamma \in \Gamma$ pertenece a $\Gamma_q$ si y solo si $\gamma g_q A = g_q A$, si y solo si $g_q^{-1} \gamma g_q \in A$. Se sigue que $\Gamma_q = g_q A g_q^{-1} \cap \Gamma$. Podemos escribir la medida asociada a $d^{-1/2} R_{\mathrm{disc}}(d)$ como

$$\lambda_d = \sum_{[q] \in C(d)} \sum_{\gamma \in \Gamma/\Gamma_q} \delta_{\gamma g_q A}.$$

Para cada $[q] \in C(d)$ la medida $\lambda_{[q]} = \sum_{\gamma \in \Gamma/\Gamma_q} \delta_{\gamma g_q A}$ es $\Gamma$-invariante a izquierda, así que le corresponde una medida $A$-invariante a derecha $\rho_{[q]}$ en $\Gamma\backslash G$. Afirmamos que $\rho_{[q]}$ es la medida de Haar $\lambda_{x_q A}$ del espacio homogéneo $x_q A$, donde $x_q = \Gamma g_q$. En efecto, usando la notación de 1.20 la medida $\mu$ en $G$ asociada a $\lambda_{[q]}$ está caracterizada por

$$\mu(\varphi) = \lambda_{[q]}(I_A \varphi) = \sum_{\gamma \in \Gamma/\Gamma_q} I_A \varphi(\gamma g_q A) = \sum_{\gamma \in \Gamma/\Gamma_q} \int_A \varphi(\gamma g_q a)\, d\lambda_A(a),$$

para toda $\varphi \in C_c(G)$. Ahora, la órbita $x_q A$ está en biyección con $\widetilde{\Gamma}_q \backslash A$ donde $\widetilde{\Gamma}_q$ es el estabilizador de $x_q$ en $A$. Es fácil ver que $\widetilde{\Gamma}_q = g_q^{-1} \Gamma g_q \cap A = g_q^{-1} \Gamma_q g_q$. Si tomamos un dominio fundamental[4] $S$ para la acción de $\widetilde{\Gamma}_q$ en $A$, entonces

$$\begin{aligned}\lambda_{[q]}(I_A \varphi) &= \sum_{\gamma \in \Gamma/\Gamma_q} \sum_{\alpha \in \widetilde{\Gamma}_q} \int_S \varphi(\gamma g_q \alpha a)\, d\lambda_A(a) = \sum_{\gamma \in \Gamma/\Gamma_q} \sum_{\beta \in \Gamma_q} \int_S \varphi(\gamma \beta g_q a)\, d\lambda_A(a) \\ &= \sum_{\gamma \in \Gamma} \int_S \varphi(\gamma g_q a)\, d\lambda_A(a) = \int_{x_q A} I_\Gamma \varphi(a)\, d\lambda_{x_q A}(a).\end{aligned}$$

Por 1.20 se sigue que $\rho_{[q]} = \lambda_{x_q A}$. Luego $\lambda_d$ se corresponde con $\rho_d = \sum_{[q]} \lambda_{x_q A}$, que está soportada en $\mathscr{G}_d := \bigcup_{[q]} x_q A$.

Luego vemos que si $\mu_d = \frac{1}{\mathrm{vol}(\mathscr{G}_d)} \rho_d$, entonces basta probar lo siguiente:

[2] Un grupo discreto es unimodular ya que su medida de Haar a izquierda es la medida de contar, que además es invariante a derecha.

[3] Es fácil ver que la medida $\mu_{\mathrm{disc}}$ definida en la introducción es efectivamente $G$-invariante.

[4] Daremos la definición más adelante, ver la sección 2.4

**Teorema 2.1.** *Cuando $d \to +\infty$ entre los enteros $d \equiv 0, 1 \pmod 4$, $\mu_d \xrightarrow{w*} \mu_{\Gamma\backslash G}$. Es decir, para toda $f \in C_c(\Gamma\backslash G)$ se tiene*

$$\frac{1}{\mathrm{vol}(\mathscr{G}_d)} \sum_{[q]\in C(d)} \int_{x_q A} f \, d\lambda_{x_q A} \to \mu_{\Gamma\backslash G}, \quad \text{cuando } d \to +\infty. \tag{2.3}$$

Por otro lado $\Gamma\backslash G \cong \mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R}) \cong T^1(\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H})$ como veremos más adelante, y resulta que la acción de $A$ en este espacio se puede interpretar de manera geométrica, como la acción del flujo geodésico, y las medidas $\mu_d$ son invariantes bajo este flujo. Por este motivo es necesario estudiar el flujo geodésico en $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$.

*Demostración de 0.1 asumiendo 2.1.* Supongamos que $\mu_d \xrightarrow{w*} \mu_{\Gamma\backslash G}$ y sea $f : G/A \to \mathbb{R}$ continua de soporte compacto. Por 1.17 el operador $I_A : C_c(G) \to C_c(G/A)$ es sobreyectivo, así que podemos elegir $h \in C_c(G)$ tal que $I_A h = f$. En ese caso, por 1.20 tenemos que

$$\frac{1}{\mathrm{vol}(\mathscr{G}_d)}\lambda_d(f) = \frac{1}{\mathrm{vol}(\mathscr{G}_d)}\lambda_d(I_A h) = \frac{1}{\mathrm{vol}(\mathscr{G}_d)}\rho_d(I_\Gamma h) = \mu_d(I_\Gamma h)$$

y luego

$$\frac{1}{\mathrm{vol}(\mathscr{G}_d)}\lambda_d(f) \to \mu_{\Gamma\backslash G}(I_\Gamma h) = \mu_{\mathrm{disc}}(I_A h) = \mu_{\mathrm{disc}}(f),$$

cuando $d \to +\infty$. Se sigue que si $\varphi_1, \varphi_2 : G/A \to \mathbb{R}$ son continuas de soporte compacto con $\mu_{\mathrm{disc}}(\varphi_2) \neq 0$, entonces

$$\frac{\lambda_d(\varphi_1)}{\lambda_d(\varphi_2)} = \frac{\lambda_d(\varphi_1)/\mathrm{vol}(\mathscr{G}_d)}{\lambda_d(\varphi_2)/\mathrm{vol}(\mathscr{G}_d)} \to \frac{\mu_{\mathrm{disc}}(\varphi_1)}{\mu_{\mathrm{disc}}(\varphi_2)}. \qquad \square$$

Para terminar esta sección, calcularemos la $\rho_d$-medida de las $A$-órbitas $x_q A$ en $\mathscr{G}_d$ y en particular veremos que tienen igual medida. Sea $K = \mathbb{Q}(\sqrt{d})$ y consideremos el morfismo de $\mathbb{Q}$-álgebras $i_0 : K \to M_2(\mathbb{R})$, $i_0(u + v\sqrt{d}) = u + v\sqrt{d}m_0$, donde

$$m_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Podemos extender $i_0$ a un morfismo $K \otimes_{\mathbb{Q}} \mathbb{R} \to M_2(\mathbb{R})$. El anillo $K \otimes \mathbb{R}$ es un $\mathbb{R}$-espacio vectorial con base $\{1\otimes 1, \xi = \sqrt{d}\otimes 1\}$. Su grupo de unidades consiste de los elementos $u+v\xi$ con $u^2 - dv^2 \neq 0$. De esto se deduce que $i_0((K \otimes \mathbb{R})^\times) \subseteq \mathrm{GL}_2(\mathbb{R})$, ya que

$$i_0(u + v\xi) = \begin{pmatrix} u + v\sqrt{d} & 0 \\ 0 & u - v\sqrt{d} \end{pmatrix}$$

tiene determinante no nulo. De hecho, la imagen de $(K \otimes \mathbb{R})^\times$ en $\mathrm{PGL}_2(\mathbb{R})$ está contenida en $A$. Sea $j_0 : (K \otimes \mathbb{R})^\times \to \mathrm{PGL}_2(\mathbb{R})$ el morfismo de grupos inducido por $i_0$ y sea $\mathcal{O}_d = \mathbb{Z}\left[\frac{d+\sqrt{d}}{2}\right] \subseteq K$. Claramente la imagen de $j_0$ es $A$. Afirmamos que $j_0(\mathcal{O}_d^\times) = \widetilde{\Gamma}_q$ para todo $q \in R_{\mathrm{disc}}(d)$. En efecto, si $q = ax^2 + bxy + cy^2$, podemos tomar $g_q = \left(\begin{smallmatrix} \frac{b+\sqrt{d}}{2c} & \frac{b-\sqrt{d}}{2c} \\ & \end{smallmatrix}\right)$. Entonces

$$g_q m_0 g_q^{-1} = \frac{1}{4c\sqrt{d}} \begin{pmatrix} b+\sqrt{d} & b-\sqrt{d} \\ 2c & 2c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2c & -b+\sqrt{d} \\ -2c & b+\sqrt{d} \end{pmatrix} = \frac{1}{\sqrt{d}} \begin{pmatrix} b & -2a \\ 2c & -b \end{pmatrix}.$$

Luego la matriz

$$g_q i_0 \left(\frac{d+\sqrt{d}}{2}\right) g_q^{-1} = \begin{pmatrix} \frac{b+d}{2} & -a \\ c & \frac{-b+d}{2} \end{pmatrix}$$

tiene coordenadas enteras, ya que $d = b^2 - 4ac \equiv b \pmod 2$. Se sigue que $g_q i_0(\beta) g_q^{-1} \in M_2(\mathbb{Z})$ para todo $\beta \in \mathcal{O}_d$. El mapa $\beta \mapsto g_q i_0(\beta) g_q^{-1}$ es un morfismo de anillos $\mathcal{O}_d \to M_2(\mathbb{Z})$, así que manda unidades en unidades, es decir, $g_q i_0(\mathcal{O}_d^\times) g_q^{-1} \subseteq \mathrm{GL}_2(\mathbb{Z})$. Luego $j_0(\mathcal{O}_d^\times) \subseteq g_q^{-1}\Gamma g_q \cap A = \widetilde{\Gamma}_q$. Para

ver la inclusión opuesta, sea $\alpha \in \widetilde{\Gamma}_q$. Entonces $\alpha$ está representado por una matriz de la forma $\left(\begin{smallmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{smallmatrix}\right)$ tal que

$$g_q \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix} g_q^{-1} \in \mathrm{GL}_2(\mathbb{Z}).$$

Podemos hallar $u, v \in \mathbb{R}$ tales que $u+v\sqrt{d} = \lambda_1$ y $u-v\sqrt{d} = \lambda_2$. Entonces $u^2 - dv^2 = \lambda_1\lambda_2 = \pm 1$, así que $u + v\xi \in (K \otimes \mathbb{R})^\times$. Por la elección de $u, v$ tenemos que $i_0(u+v\xi) = \left(\begin{smallmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{smallmatrix}\right)$, luego

$$g_q i_0(u+v\xi) g_q^{-1} = \begin{pmatrix} u+bv & -2av \\ 2cv & u-bv\end{pmatrix} \in M_2(\mathbb{Z}).$$

Tomando traza de esta matriz vemos que $u \in \frac{1}{2}\mathbb{Z}$ y luego $av, bv, cv \in \frac{1}{2}\mathbb{Z}$. Por hipótesis el máximo común divisor de $a, b, c$ es 1, así que $v \in \frac{1}{2}\mathbb{Z}$. Entonces

$$u + v\sqrt{d} = (u+bv) - 2v\frac{b+d}{2} + 2v\frac{d+\sqrt{d}}{2} \in \mathcal{O}_d.$$

Además $u + v\sqrt{d} \in \mathcal{O}_d^\times$ ya que $u^2 - dv^2 = \pm 1$. Por lo tanto $\alpha \in j_0(\mathcal{O}_d^\times)$.

Se sigue que $j_0$ induce un isomorfismo de grupos

$$\widehat{j_0} : \frac{(K\otimes\mathbb{R})^\times}{\mathbb{R}^\times \mathcal{O}_d^\times} \xrightarrow{\sim} A/\widetilde{\Gamma}_q,$$

donde identificamos a $\mathbb{R}^\times$ con su imagen en $(K\otimes\mathbb{R})^\times$ por el morfismo $x \mapsto 1 \otimes x$. Por otro lado, hay un isomorfismo $(K\otimes\mathbb{R})^\times \xrightarrow{\sim} (\mathbb{R}^\times)^2$, $u + v\xi \mapsto (u+v\sqrt{d}, u-v\sqrt{d})$ y el subgrupo $\mathbb{R}^\times$ tiene como imagen la diagonal $\{(x,x) : x \in \mathbb{R}^\times\}$, así que $(K\otimes\mathbb{R})^\times/\mathbb{R}^\times \cong \mathbb{R}^\times$. Sea $\phi : \mathcal{O}_d^\times \to \mathbb{R}^2$ dado por $\phi(u+v\sqrt{d}) = (\log|u+v\sqrt{d}|, \log|u-v\sqrt{d}|)$. La imagen de $\phi$ está contenida en el subespacio $V = \{(x,y) \in \mathbb{R}^2 : x + y = 0\}$. El *regulador* $\mathrm{Reg}(\mathcal{O}_d)$ de $\mathcal{O}_d$ se puede definir como el volumen de $V/\phi(\mathcal{O}_d^\times)$ con respecto a la medida de Lebesgue, que es una medida de Haar. Hay un epimorfismo de grupos $(K\otimes\mathbb{R})^\times/\mathbb{R}^\times \to V$ definido por

$$u + v\xi \mapsto \left(\frac{1}{2}\log\left|\frac{u+v\sqrt{d}}{u-v\sqrt{d}}\right|, -\frac{1}{2}\log\left|\frac{u+v\sqrt{d}}{u-v\sqrt{d}}\right|\right),$$

que coincide con $\phi$ en $\mathcal{O}_d$. Esto induce un epimorfismo $(K\otimes\mathbb{R})^\times/\mathbb{R}^\times\mathcal{O}_d^\times \to V/\phi(\mathcal{O}_d^\times)$, cuyo núcleo es de orden 2. Luego podemos elegir una medida de Haar en $(K\otimes\mathbb{R})^\times$ de modo que $(K\otimes\mathbb{R})^\times/\mathbb{R}^\times\mathcal{O}_d^\times$ tenga medida $\mathrm{Reg}(\mathcal{O}_d)$. El pushforward de esta medida bajo $\widehat{j_0}$ es una medida de Haar en $A/\widetilde{\Gamma}_q \cong x_q A$. Luego, multiplicando a $\rho_d$ por una constante (que no depende de $d$) podemos suponer que $\rho_d(x_q A) = \mathrm{Reg}(\mathcal{O}_d)$ para toda $q \in R_{\mathrm{disc}}(d)$. Notar finalmente que multiplicar a $\rho_d$ por una constante no modifica a $\mu_d$.

## 2.1. Variedades riemannianas

Recordemos que una *métrica riemanniana* en una variedad diferenciable $M$ es una aplicación $Q$ que le asigna a cada punto $p \in M$ un producto interno $Q_p = \langle\cdot\,,\cdot\rangle_p$ en $T_pM$ de manera suave, en el sentido de que $\langle X, Y\rangle$ es una función $C^\infty$ para todo par de campos suaves $X, Y \in \mathfrak{X}(M)$. Una *variedad riemanniana* es un par $(M, Q)$ donde $M$ es una variedad diferenciable y $Q$ es una métrica riemanniana.

Una métrica riemanniana nos permite definir longitudes y distancias (y ángulos, pero no los usaremos). El producto interno $\langle\cdot,\cdot\rangle_p$ determina una norma $\|v\|_p = \langle v, v\rangle_p^{1/2}$ en $T_pM$, para cada $p \in M$. La *longitud* de una curva suave a trozos $\gamma : [a,b] \to M$ se define como

$$L(\gamma) = \int_a^b \|\dot\gamma(t)\|_{\gamma(t)}\, dt. \tag{2.4}$$

A su vez, podemos definir una distancia en $M$ en términos de la longitud de curvas:

$$d_Q(p,q) = \text{ínf}\{L(\gamma) : \gamma \text{ es una curva suave a trozos que une a } p \text{ y } q\}. \tag{2.5}$$

Es fácil ver que si $M$ es una variedad riemanniana conexa, entonces todo par de puntos está conectado por una curva suave a trozos, así que $d_Q(p,q) < +\infty$. Además es claro que $d_Q$ cumple la desigualdad triangular, ya que si $\gamma_1$ es una curva que une $p$ y $q$ y $\gamma_2$ es una curva que une $q$ y $r$ entonces su concatenación es una curva que une $p$ y $r$, que es suave a trozos si $\gamma_1$ y $\gamma_2$ lo son.

**Teorema 2.2.** *Si $(M,Q)$ es una variedad riemanniana conexa entonces $d_Q$ es una distancia y la topología que induce en $M$ es igual a topología original de $M$.*

*Demostración.* Ver [Lee03, Teo. 13.29]. □

**Definición 2.3.** Una *conexión afín* en una variedad diferenciable $M$ es una aplicación

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M), \qquad (X,Y) \mapsto \nabla_X Y$$

que es $C^\infty(M)$-lineal en la primera variable, $\mathbb{R}$-lineal en la segunda y satisface la ecuación

$$\nabla_X(fY) = X(f)Y + f\nabla_X Y$$

para todo par de campos $X, Y \in \mathfrak{X}(M)$ y toda $f \in C^\infty(M)$. Decimos que una conexión $\nabla$ es *simétrica* si cumple

$$\nabla_X Y - \nabla_Y X = [X,Y]$$

para todos $X, Y \in \mathfrak{X}(M)$ y decimos que es *compatible* con una métrica riemanniana $Q$ si

$$X\langle Y,Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$$

para todos $X, Y, Z \in \mathfrak{X}(M)$.

**Teorema 2.4** (Levi-Civita)**.** *Sea $(M,Q)$ una variedad riemanniana. Entonces existe una única conexión afín $\nabla$ en $M$ que es simétrica y compatible con la métrica riemanniana.*

*Demostración.* Ver [CF92, Teo 2.3.6]. □

Si $\nabla$ es una conexión afín, es fácil ver usando un argumento con funciones bump que $(\nabla_X Y)_p$ solo depende de los valores de $X$ e $Y$ en un entorno de $p$. En particular si tomamos una carta $(U,\varphi)$ donde $\varphi = (x_1, \dots, x_n)$ y escribimos a $X$ e $Y$ en coordenadas $X = \sum_i a_i \partial_i$, $Y = \sum_i b_i \partial_i$ donde $\partial_i = \frac{\partial}{\partial x_i}$, entonces

$$\nabla_X Y = \sum_{i,j} a_i \nabla_{\partial_i}(b_j \partial_j) = \sum_{i,j} a_i \left[ \frac{\partial b_j}{\partial x_i}\partial_j + b_j \nabla_{\partial_i}(\partial_j) \right].$$

Expandiendo $\nabla_{\partial_i}(\partial_j) = \sum_k \Gamma_{ij}^k \partial_k$ tenemos que

$$\nabla_X Y = \sum_k \left( \sum_{i,j} a_i b_j \Gamma_{ij}^k + X(b_k) \right) \partial_k.$$

En particular se ve que $(\nabla_X Y)_p$ solo depende de $X_p$, $Y_p$ y de las derivadas $X_p(b_k)$. Los coeficientes $\Gamma_{ij}^k$ son los *símbolos de Christoffel* de la conexión.

**Definición 2.5.** Sea $I \subseteq \mathbb{R}$ un intervalo y sea $\gamma : I \to M$ una curva suave. Un *campo a lo largo de $\gamma$* es una función suave $V : I \to TM$ tal que $V(t) \in T_{\gamma(t)}M$ para todo $t \in I$. Denotamos por $\mathfrak{X}(\gamma)$ al conjunto de todos los campos a lo largo de $\gamma$. Decimos que un campo $\widetilde{V}$ *extiende* a un campo $V \in \mathfrak{X}(\gamma)$ si $\widetilde{V}_{\gamma(t)} = V(t)$ para todo $t \in I$.

**Proposición 2.6.** *Sea $M$ una variedad diferenciable con una conexión afín $\nabla$ y sea $\gamma : I \to M$ una curva suave. Entonces existe un único morfismo $D_t : \mathfrak{X}(\gamma) \to \mathfrak{X}(\gamma)$ que es $\mathbb{R}$-lineal y cumple*

*(i) $D_t(fV) = \frac{df}{dt}V + fD_t(V)$ para todo $V \in \mathfrak{X}(\gamma)$ y para toda $f : I \to \mathbb{R}$ suave.*

*(ii) Si $\widetilde{V} \in \mathfrak{X}(M)$ es un campo que extiende a $V$ entonces $D_t(V) = \nabla_{\dot\gamma(t)}\widetilde{V}$.*

*Este morfismo se denomina la* derivada covariante.

*Demostración.* Ver [CF92, Teo. 2.2.2]. □

Decimos que una curva suave $\gamma : I \to M$ es una *geodésica* con respecto a una conexión $\nabla$ si $D_t\dot{\gamma} = 0$ para todo $t \in I$. En coordenadas, esta condición se traduce en un sistema de ecuaciones diferenciales ordinarias de orden 2. En efecto, si $(U, \varphi)$ es una carta con coordenadas $(x_1, \dots, x_n)$, por las propiedades de la derivada covariante tenemos que

$$\begin{aligned} D_t\dot{\gamma} &= \sum_i D_t(\dot{\gamma}_i\partial_i) = \sum_i (\ddot{\gamma}_i\partial_i + \dot{\gamma}_i D_t(\partial_i)) \\ &= \sum_k \ddot{\gamma}_k\partial_k + \sum_i \dot{\gamma}_i \nabla_{\dot{\gamma}(t)}(\partial_i) \\ &= \sum_k \ddot{\gamma}_k\partial_k + \sum_{i,j} \dot{\gamma}_i\dot{\gamma}_j \nabla_{\partial_j}(\partial_i) \\ &= \sum_k \left(\ddot{\gamma}_k + \sum_{i,j} \dot{\gamma}_i\dot{\gamma}_j\Gamma_{ij}^k\right)\partial_k. \end{aligned}$$

Luego $\gamma$ es una geodésica si y solo si $\ddot{\gamma}_k + \sum_{i,j} \dot{\gamma}_i\dot{\gamma}_j\Gamma_{ij}^k = 0$ para todo $1 \le k \le n$, en toda carta $(U, \varphi)$.

**Proposición 2.7.** *Sea $\gamma : [a, b] \to M$ una curva suave a trozos y tal que $\|\dot{\gamma}(t)\| = c$ es constante en $[a, b]$. Si $L(\gamma) = d(\gamma(a), \gamma(b))$ entonces $\gamma$ es una geodésica.*

*Demostración.* Ver [CF92, Teo. 2.3.9]. □

**Definición 2.8.** Sean $(M_1, Q_1)$, $(M_2, Q_2)$ variedades riemannianas. Decimos que una función suave $\varphi : M_1 \to M_2$ es una *isometría riemanniana* si $\varphi^* Q_2 = Q_1$, es decir, si

$$\langle v, w\rangle_p = \langle d_p\varphi(v), d_p\varphi(w)\rangle_{\varphi(p)}$$

para todos $v, w \in T_p(M)$ y para todo $p \in M$.

## 2.2. El plano hiperbólico

El *semi-plano superior* se define como

$$\mathbb{H} = \{x + iy \in \mathbb{C} : y > 0\}. \tag{2.6}$$

Este espacio es un abierto de $\mathbb{C} \cong \mathbb{R}^2$ y luego tiene estructura de variedad diferenciable. Le damos a $\mathbb{H}$ la *métrica hiperbólica* $Q = \frac{1}{y^2}(dx^2 + dy^2)$.

Como $\mathbb{H}$ puede ser cubierto por una sola carta $(x, y)$, su fibrado tangente es trivial: hay un isomorfismo $\phi : T\mathbb{H} \xrightarrow{\sim} \mathbb{H} \times \mathbb{C}$ dado por $a\frac{\partial}{\partial x}\big|_z + b\frac{\partial}{\partial y}\big|_z \mapsto (z, a + ib)$. Este isomorfismo restringido a cada fibra induce un isomorfismo canónico $T_z\mathbb{H} \xrightarrow{\sim} \mathbb{C}$, para todo $z \in \mathbb{H}$. En particular $T_z\mathbb{H}$ tiene estructura de $\mathbb{C}$-espacio vectorial para todo $z$. Usando este isomorfismo, el producto interno en $T_z\mathbb{H}$ se puede expresar como

$$\langle v, w\rangle_z = \frac{1}{\mathrm{Im}(z)^2}\mathrm{Re}(v\overline{w}). \tag{2.7}$$

Ahora procedemos a describir la acción de $\mathrm{PSL}_2(\mathbb{R})$ en $\mathbb{H}$. El grupo $\mathrm{SL}_2(\mathbb{R})$ actúa en la esfera de Riemann $\mathbb{P}^1(\mathbb{C})$ via $g \cdot [v] = [gv]$ para todo $v \in \mathbb{C}^2$. Podemos mirar a $\mathbb{H}$ como subespacio de $\mathbb{P}^1(\mathbb{C})$ a través del mapa $z \in \mathbb{H} \mapsto [\left(\begin{smallmatrix} z \\ 1 \end{smallmatrix}\right)] \in \mathbb{P}^1(\mathbb{C})$. En ese caso, notemos que $g \cdot \mathbb{H} \subseteq \mathbb{H}$ para todo $g \in \mathrm{SL}_2(\mathbb{R})$. En efecto, si $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$, entonces

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \left[\begin{pmatrix} z \\ 1 \end{pmatrix}\right] = \left[\begin{pmatrix} az + b \\ cz + d \end{pmatrix}\right] = \left[\begin{pmatrix} \frac{az+b}{cz+d} \\ 1 \end{pmatrix}\right].$$

(Notemos que $cz+d \neq 0$ ya que tiene parte imaginaria no nula). Además,

$$\begin{aligned}\operatorname{Im}\left(\frac{az+b}{cz+d}\right) &= \operatorname{Im}\left(\frac{(az+b)(c\overline{z}+d)}{|cz+d|^2}\right)\\ &= \frac{\operatorname{Im}(ac|z|^2+adz+bc\overline{z}+bd)}{|cz+d|^2}\\ &= \frac{(ad-bc)\operatorname{Im}(z)}{|cz+d|^2} = \frac{\operatorname{Im}(z)}{|cz+d|^2} > 0, \end{aligned} \tag{2.8}$$

así que $g\cdot z \in \mathbb{H}$ para todo $z\in\mathbb{H}$. Por lo tanto esto define una acción de $\mathrm{SL}_2(\mathbb{R})$ en $\mathbb{H}$. Como $-\mathrm{I} = \left(\begin{smallmatrix}-1 & 0\\ 0 & -1\end{smallmatrix}\right)$ actúa trivialmente en $\mathbb{H}$, la acción pasa al cociente $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm\mathrm{I}\}$.

Esta acción a su vez induce una acción de $\mathrm{PSL}_2(\mathbb{R})$ en $T\mathbb{H} \cong \mathbb{H}\times\mathbb{C}$, $g\cdot(z,w) = (g\cdot z, d_z g(w))$. Notemos que para todo $g = \left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{R})$ la función $g(z) = \frac{az+b}{cz+d}$ es holomorfa en $\mathbb{H}$. Por las ecuaciones de Cauchy-Riemann esto implica que la matriz Jacobiana real de $g(z) = u(z)+iv(z)$ es de la forma

$$Jg(z) = \begin{pmatrix} u_x & u_y\\ v_x & v_y\end{pmatrix} = \begin{pmatrix} u_x & -v_x\\ v_x & u_x\end{pmatrix} \tag{2.9}$$

y luego bajo la identificación $T_z\mathbb{H}\cong\mathbb{C}$, tenemos que

$$d_z g(w) = g'(z)w = \frac{w}{(cz+d)^2}$$

para todo $w\in T_z\mathbb{H}$.

**Proposición 2.9.** *La acción de* $\mathrm{PSL}_2(\mathbb{R})$ *en* $\mathbb{H}$ *es isométrica y transitiva. El estabilizador* $\mathrm{Stab}(i)$ *de* $i$ *es* $\mathrm{PSO}(2)$.

*Demostración.* Sea $g = \left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{R})$, sea $z = x+iy\in\mathbb{H}$ y sean $v,w\in T_z\mathbb{H}$. Por (2.8) tenemos que

$$\begin{aligned}\langle d_z g(v), d_z g(w)\rangle_{g(z)} &= \left(\frac{y}{|cz+d|^2}\right)^{-2} \operatorname{Re}\left(d_z g(v)\overline{d_z g(w)}\right)\\ &= \frac{|cz+d|^4}{y^2}\operatorname{Re}\left(\frac{v\overline{w}}{|cz+d|^4}\right) = \frac{1}{y^2}\operatorname{Re}(v\overline{w}) = \langle v,w\rangle_z.\end{aligned}$$

Luego $g$ actúa isométricamente en $\mathbb{H}$. Para probar que la acción es transitiva, basta ver que para todo $z=x+iy\in\mathbb{H}$ existe $g\in\mathrm{PSL}_2(\mathbb{R})$ tal que $g(i)=z$. Y efectivamente la matriz $\left(\begin{smallmatrix}\sqrt{y} & x/\sqrt{y}\\ 0 & 1/\sqrt{y}\end{smallmatrix}\right)$ manda $i$ a $x+iy$.

Queda probar la última afirmación. Es claro que $\mathrm{PSO}(2)\subseteq\mathrm{Stab}(i)$ ya que

$$\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix} i = \frac{i\cos\theta-\sin\theta}{\cos\theta+i\sin\theta} = i.$$

Recíprocamente, supongamos que $g = \left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{R})$ cumple $g(i)=i$. Por (2.8) se tiene que $\operatorname{Im} g(i) = |ci+d|^{-2} = 1$, es decir, $|ci+d|=1$ y luego existe $\theta\in[0,2\pi)$ tal que $c=\sin\theta$, $d=\cos\theta$. Como $g(i)=i$, tenemos que $ai+b = i(ci+d) = -c+di$, así que $a=d$ y $b=-c$. Se sigue que

$$g = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix} \in \mathrm{PSO}(2). \qquad \square$$

Como $\mathrm{PSL}_2(\mathbb{R})$ actúa por isometrías, su acción en $T\mathbb{H}$ preserva el subespacio

$$T^1\mathbb{H} := \{(z,v)\in T\mathbb{H} : \|v\|_z = 1\}$$

que es llamado el *fibrado tangente unitario*. La misma definición es válida para cualquier variedad riemanniana $(M,Q)$ reemplazando $\mathbb{H}$ por $M$. Notemos que hay un difeomorfismo $T^1\mathbb{H} \xrightarrow{\sim} \mathbb{H}\times S^1$, $(z,v)\mapsto (z, \frac{v}{\operatorname{Im}(z)})$.

**Proposición 2.10.** *La acción de* $\mathrm{PSL}_2(\mathbb{R})$ *en* $T^1\mathbb{H}$ *es simplemente transitiva.*

*Demostración.* Dado $(z,v) \in T^1\mathbb{H}$, veamos que existe $g \in \mathrm{PSL}_2(\mathbb{R})$ tal que $g.(z,v) = (i,i)$. De esto se deduce que la acción es transitiva. Por la proposición anterior existe $g \in \mathrm{PSL}_2(\mathbb{R})$ tal que $g(z) = i$. Si $h = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$ con $\theta \in \mathbb{R}$, entonces $h(i) = i$ y

$$d_i h(w) = \frac{w}{(\cos\theta + i\sin\theta)^2} = \frac{w}{\cos(2\theta) + i\sin(2\theta)}.$$

A medida que $\theta$ varía en $\mathbb{R}$, el denominador toma cualquier valor en $S^1$. Luego existe $\theta \in \mathbb{R}$ tal que $d_i h \circ d_z g(v) = i$. Se sigue que $hg.(z,v) = (i,i)$.

Para probar la transitividad simple basta ver que el estabilizador de $(i,i)$ es la identidad. Si $g \in \mathrm{PSL}_2(\mathbb{R})$ fija $(i,i)$, entonces $g \in \mathrm{PSO}(2)$, digamos $g = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. Como $d_i g(i) = i(\cos\theta + i\sin\theta)^{-2}$, se sigue que $\cos\theta = \pm 1$ y $\sin\theta = 0$. Luego $g = \pm I$ es la identidad. □

Luego hay un isomorfismo $\mathrm{PSL}_2(\mathbb{R}) \xrightarrow{\sim} T^1\mathbb{H}$ dado por $g \mapsto g.(i,i)$.

**Proposición 2.11.** *Las geodésicas en* $\mathbb{H}$ *son los semicírculos cuyo centro yace en el eje* $x$ *y las semirrectas que son paralelas al eje* $y$*.*

*Demostración.* Veamos primero que las semirrectas paralelas al eje $y$ son geodésicas. Supongamos que $\gamma : [0,\ell] \to \mathbb{H}$ es una curva que une los puntos $z = ia$ y $w = ib$. Cambiando la orientación de $\gamma$ podemos suponer que $b > a$. Entonces

$$L(\gamma) = \int_0^\ell \frac{\sqrt{\dot{x}^2(t) + \dot{y}^2(t)}}{y(t)}\, dt \geq \int_0^\ell \frac{\dot{y}(t)}{y(t)}\, dt = \log b - \log a.$$

Por otro lado, si uno elige $\ell = \log b - \log a$ y $\gamma(t) = iae^t$ entonces se da la igualdad en la ecuación anterior, lo que implica que $L(\gamma) = d(z,w)$. Además

$$\|\dot{\gamma}(t)\| = \frac{|\dot{y}(t)|}{y(t)} = \frac{ae^t}{ae^t} = 1,$$

así que por la proposición 2.7, $\gamma$ es una geodésica.

Pasando al caso general, supongamos que $\gamma : [t_0,t_1] \to \mathbb{H}$ es una geodésica con $\gamma(t_0) = z$, $\dot{\gamma}(t_0) = v$. Como la acción de $\mathrm{PSL}_2(\mathbb{R})$ en $T^1\mathbb{H}$ es transitiva, podemos tomar $g \in \mathrm{PSL}_2(\mathbb{R})$ tal que $g(i) = z$ y $d_i g(\dot{\gamma}(i)) = v$. Sea $\gamma_0(t) = ie^t$ la geodésica del párrafo anterior con $a = 1$. Como $g$ es una isometría inversible, $\beta = g \circ \gamma_0$ cumple que $\|\dot{\beta}(t)\| = \|\dot{\gamma}_0(t)\| = 1$ para todo $t$ y

$$L(\beta) = d(\gamma(t_0), \gamma(t_1)) = d(\beta(t_0), \beta(t_1)).$$

Luego $\beta$ es una geodésica por 2.7. Ahora, esta geodésica está determinada por $\beta(t_0)$ y $\dot{\beta}(t_0)$ ya que $\beta$ cumple una ecuación diferencial ordinaria de orden 2. Como los valores de $\beta$ y su derivada en $a$ coinciden con los de $\gamma$, esto implica que $\gamma(t) = \beta(t) = g \circ \gamma_0(t)$.

Supongamos que $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Si $c = 0$, entonces $d = a^{-1}$ y $g(z) = a^2 z + ab$. En ese caso $g \circ \gamma_0(t) = ia^2 e^t + ab$ es una recta vertical. Si en cambio $d = 0$, entonces $b = -1/c$ y

$$g(z) = \frac{a}{c} - \frac{1}{c^2 z}.$$

Luego $g \circ \gamma(t) = \frac{a}{c} - ic^{-2}e^{-t}$ también es una recta vertical. Finalmente, si $cd \neq 0$, afirmamos que $g(i\mathbb{R}_{>0})$ está contenido en una circunferencia con centro $\frac{ad+bc}{2cd}$. En efecto, dado $t > 0$, tenemos

$$\begin{aligned} \frac{ait+b}{cit+d} - \frac{ad+bc}{2cd} &= \frac{(ait+b)2cd - (cit+d)(ad+bc)}{2cd(cit+d)} \\ &= \frac{cit(ad-bc) + d(bc-ad)}{2cd(cit+d)} = \frac{cit-d}{2cd(cit+d)}. \end{aligned}$$

La norma del último término es $(2cd)^{-1}$, que no depende de $t$. Luego la imagen de $g \circ \gamma_0$ está contenida en el semicírculo de centro $\frac{ad+bc}{2cd}$ y radio $(2cd)^{-1}$. Notemos además que a medida que $t \to -\infty$, $g \circ \gamma_0(t)$ tiende a $b/d$ y a medida que $t \to +\infty$, $g \circ \gamma_0(t)$ tiende a $a/c$. Estos son el punto inicial y final del semicírculo. Como la imagen de $g \circ \gamma_0$ es conexa, debe ser todo el semicírculo. □

Por la demostración se ve que cualquier geodésica en $\mathbb{H}$ se puede extender para estar definida en todo $\mathbb{R}$. Las variedades riemannianas con esta propiedad se dicen *geodésicamente completas*.

**Proposición 2.12.** *La distancia hiperbólica entre dos puntos $z, w \in \mathbb{H}$ está dada por*

$$d(z,w) = \log\left(\frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}\right). \tag{2.10}$$

*Demostración.* Llamemos $\delta(z,w)$ al lado derecho de 2.10. Afirmamos que $\delta(g(z), g(w)) = \delta(z,w)$ para todo $g \in \mathrm{PSL}_2(\mathbb{R})$. En efecto, es claro que $\delta$ es invariante bajo $z \mapsto z+t$ con $t \in \mathbb{R}$, y es invariante bajo $z \mapsto -1/z$ ya que

$$\frac{|1/z - 1/\overline{w}| + |1/z - 1/w|}{|1/z - 1/\overline{w}| - |1/z - 1/w|} = \frac{(|z-\overline{w}| + |z-w|)/|wz|}{(|z-\overline{w}| - |z-w|)/|wz|} = \frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}.$$

Como veremos en la siguiente sección, $\mathrm{SL}_2(\mathbb{R})$ está generado por $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ y $\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$ con $t \in \mathbb{R}$. Luego la afirmación es válida.

Si $z = ia$ y $w = ib$, con $a < b$, vimos en la demostración de la proposición anterior que $\gamma : [0, \log(b/a)] \to \mathbb{H}$, $\gamma(t) = iae^t$ es una geodésica entre $z$ y $w$ con longitud $\log(b/a)$. En particular $d(ia, ib) = \log(b/a)$, mientras que

$$\log\left(\frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}\right) = \log\left(\frac{|a+b| + |a-b|}{|a+b| - |a-b|}\right) = \log\left(\frac{2b}{2a}\right) = d(ia, ib).$$

En el caso general, aplicando una transformación afín $s \mapsto as + t$ podemos llevar $z$ a $i$, y luego podemos aplicar una transformación en $\mathrm{PSO}(2)$ para llevar a $w$ a la recta $\mathrm{Re}(s) = 0$. Luego existe $g \in \mathrm{PSL}_2(\mathbb{R})$ tal que $g(z) = i$ y $g(w) = ib$ con $b > 0$. Intercambiando $z$ y $w$ podemos suponer que $b > 1$. Como $s \mapsto g(s)$ es una isometría, tenemos que

$$d(z,w) = d(i, ib) = \delta(i, ib) = \delta(z,w). \qquad \square$$

**Corolario 2.13.** *Sean $z, w \in \mathbb{H}$ tales que $|\mathrm{Re}(z-w)| \leq 1$ y $\mathrm{Im}(z) \geq \mathrm{Im}(w) \geq 1$. Entonces*

$$\log\left(\frac{\mathrm{Im}(z)}{\mathrm{Im}(w)}\right) \leq d(z,w) \leq \log\left(4\frac{\mathrm{Im}(z)}{\mathrm{Im}(w)}\right). \tag{2.11}$$

*Demostración.* Sean $r = \mathrm{Im}(z)$, $s = \mathrm{Im}(w)$. Tenemos que

$$\frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|} = \frac{(|z-\overline{w}| + |z-w|)^2}{|z-\overline{w}|^2 - |z-w|^2}$$

y puesto que $\mathrm{Re}(z-\overline{w}) = \mathrm{Re}(z-w)$, es

$$|z-\overline{w}|^2 - |z-w|^2 = \mathrm{Im}(z-\overline{w})^2 - \mathrm{Im}(z-w)^2 = (r+s)^2 - (r-s)^2 = 4rs. \tag{2.12}$$

Por otro lado, como $|\mathrm{Re}(z-w)| \leq 1$,

$$|z-\overline{w}| + |z-w| \leq (r+s+1) + (r-s+1) = 2(r+1) \leq 4r.$$

Luego $d(z,w) \leq \log\left(\frac{16r^2}{4rs}\right) = \log\left(\frac{4r}{s}\right)$. La primera desigualdad se obtiene acotando

$$|z-\overline{w}| + |z-w| \geq (r+s) + (r-s) = 2r$$

lo que junto con (2.12) nos da

$$d(z,w) \geq \log\left(\frac{4r^2}{4rs}\right) = \log\left(\frac{r}{s}\right). \qquad \square$$

**Lema 2.14.** *La medida $d\mu_0 = y^{-2}dx\,dy$ en $\mathbb{H}$ es invariante bajo la acción de $\mathrm{PSL}_2(\mathbb{R})$.*

*Demostración.* Sea $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{R})$. Por (2.9) tenemos que $\det Jg(z) = u_x^2 + v_x^2 = |g'(z)|^2 = |cz+d|^{-4}$. Luego, por el teorema de cambio de variable,

$$\int_{\mathbb{H}} f(g^{-1}\cdot z)\frac{dx\,dy}{y^2} = \int_{\mathbb{H}} \frac{f(z)}{|cz+d|^4}\frac{|cz+d|^4}{y^2}dx\,dy = \int_{\mathbb{H}} f(z)\frac{dx\,dy}{y^2},$$

para toda $f \in C_c(\mathbb{H})$. Se sigue que $(L_g^{-1})_*\mu_0 = \mu_0$, así que $\mu_0$ es invariante. □

El *flujo geodésico* de una variedad riemanniana geodésicamente completa $(M, Q)$ es el flujo $g_t : T^1M \to T^1M$ en el fibrado tangente unitario definido de la siguiente forma: si $(p, v) \in T^1M$ entonces $g_t(p, v) = (\gamma_v(t), \dot\gamma_v(t))$ donde $\gamma_v$ es la geodésica en $M$ con $\gamma_v(0) = p$, $\dot\gamma_v(0) = v$.

**Proposición 2.15.** *El flujo geodésico en $T^1\mathbb{H} \cong \mathrm{PSL}_2(\mathbb{R})$ es de la forma $g_t(x) = xa_t^{-1}$ donde*

$$a_t = \begin{pmatrix} e^{-t/2} & 0 \\ 0 & e^{t/2} \end{pmatrix}. \tag{2.13}$$

*Demostración.* Como vimos en la demostración de la proposición 2.11 la curva $\gamma(t) = ie^t$ es la geodésica que comienza en $i$ con $\dot\gamma(0) = i$. Y es inmediato verificar que $(\gamma(t), \dot\gamma(t)) = a_t^{-1}.(i, i)$.

Ahora sea $(p, v) \in T^1\mathbb{H}$ un punto cualquiera y sea $g \in \mathrm{PSL}_2(\mathbb{R})$ tal que $(p, v) = g.(i, i)$. Como $g$ actúa isométricamente, $\beta(t) = g \circ \gamma(t)$ es la geodésica que pasa por $p$ con vector tangente $v$. Luego $g_t(p, v) = (\beta(t), \dot\beta(t)) = (ga_t^{-1}.i, g_*(a_t^{-1})_*(v)) = ga_t^{-1}.(i, i)$, que bajo el isomorfismo $T^1\mathbb{H} \cong \mathrm{PSL}_2(\mathbb{R})$ se corresponde con $ga_t^{-1}$. □

## 2.3. La descomposición de Iwasawa

El principal resultado de este capítulo es la ergodicidad del flujo geodésico en $\mathrm{PSL}_2(\mathbb{Z})\backslash G$. Para probarlo necesitaremos un par de resultados sobre la estructura topológica y algebraica de $\mathrm{PSL}_2(\mathbb{R})$. Todos ellos se deducen de la *descomposición de Iwasawa* de $\mathrm{SL}_2(\mathbb{R})$. En general, si $G$ es un grupo de Lie conexo semisimple entonces existen un subgrupo compacto maximal $K$, un subgrupo abeliano $A$ y un subgrupo nilpotente $N$, tales que $G$ es difeomorfo a $K \times A \times N$ [Hel78, §VI.5]. En el caso de $\mathrm{SL}_2(\mathbb{R})$ este resultado se puede probar directamente.

Definimos los siguientes subgrupos de $\mathrm{SL}_2(\mathbb{R})$:

$$K = \left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} : \theta \in \mathbb{R} \right\},$$

$$N = \left\{ \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} : s \in \mathbb{R} \right\}, \qquad N^- = \left\{ \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} : t \in \mathbb{R} \right\}$$

y $A$ es el grupo definido en (2.2). También denotamos por $A$, $N$, $N^-$ a la imagen del respectivo grupo en $\mathrm{PSL}_2(\mathbb{R})$. Este abuso de notación no es tan grave ya que la proyección de estos subgrupos en el cociente es inyectiva.

**Proposición 2.16.** *El mapa $\phi : N \times A \times K \to \mathrm{SL}_2(\mathbb{R})$, $\phi(n, a, k) = nak$ es un homeomorfismo.*

*Demostración.* Es claro que $\phi$ es continua. Veamos primero que $\phi$ es sobreyectiva. Sea $g \in \mathrm{SL}_2(\mathbb{R})$ y sea $x + iy = g(i)$. Entonces

$$g(i) = x + iy = \begin{pmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{pmatrix} i = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{y} & 0 \\ 0 & 1/\sqrt{y} \end{pmatrix} i.$$

Llamando $n = \left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right)$, $a = \left(\begin{smallmatrix} \sqrt{y} & 0 \\ 0 & 1/\sqrt{y} \end{smallmatrix}\right)$, tenemos que $(na)^{-1}g.i = i$ así que $(na)^{-1}g = k$ para algún $k \in K$. Luego $g = nak$ está en la imagen de $\phi$.

Para ver inyectividad, supongamos que $g \in \mathrm{SL}_2(\mathbb{R})$ es la imagen de $(n, a, k)$ bajo $\phi$, donde $n = \left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right)$ y $a = \left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right)$ con $r > 0$. Entonces

$$g(i) = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r & 0 \\ 0 & 1/r \end{pmatrix} i = s + r^2 i.$$

Luego las entradas de $n$ y $a$ quedan determinadas por las de $g$. Además, $k = (na)^{-1}g$, así que $n$, $a$ y $k$ están determinados por $g$. Luego $\phi$ es una biyección. Finalmente, $\phi^{-1}$ es continua ya que $n$, $a$ y $k$ dependen de manera continua de $g$. En efecto,

$$n = \begin{pmatrix} 1 & \mathrm{Re}(g.i) \\ 0 & 1 \end{pmatrix}, \qquad a = \begin{pmatrix} \mathrm{Im}(g.i)^{1/2} & 0 \\ 0 & \mathrm{Im}(g.i)^{-1/2} \end{pmatrix}$$

y $k = (na)^{-1}g$. Por lo tanto $\phi$ es un homeomorfismo. □

**Proposición 2.17.** *No hay morfismos continuos no triviales* $\mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}$. *En particular* $\mathrm{SL}_2(\mathbb{R})$ *y* $\mathrm{PSL}_2(\mathbb{R})$ *son grupos unimodulares.*

*Demostración.* Sea $f : \mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}$ un morfismo de grupos continuo. Como $\mathrm{SL}_2(\mathbb{R}) = NAK$, basta ver que $f(K) = f(A) = f(N) = 1$ para ver que $f$ es trivial. En primer lugar, como $K \cong S^1$ es compacto, $f(K)$ es un subgrupo compacto de $\mathbb{R}$ y luego $f(K) = 1$. Dado $\left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right) \in A$, como

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r & 0 \\ 0 & 1/r \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 1/r & 0 \\ 0 & r \end{pmatrix}$$

esto implica que $f\left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right) = -f\left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right)$ y luego $f\left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right) = 0$. Finalmente, para ver que $f(N) = 1$ usamos la identidad

$$\begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 4x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Aplicando $f$ a ambos lados, obtenemos que $f\left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right) = f\left(\begin{smallmatrix} 1 & 4x \\ 0 & 1 \end{smallmatrix}\right) = 4f\left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right)$, luego $f\left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right) = 0$. En particular la función modular $\Delta_{\mathrm{SL}_2(\mathbb{R})} : \mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}_{>0}$ debe ser trivial ya que $\mathbb{R} \cong \mathbb{R}_{>0}$ como grupos topológicos, así que $\mathrm{SL}_2(\mathbb{R})$ es unimodular. Análogamente, $\Delta_{\mathrm{PSL}_2(\mathbb{R})} : \mathrm{PSL}_2(\mathbb{R}) \to \mathbb{R}_{>0}$ es trivial ya que la composición $\mathrm{SL}_2(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R}) \xrightarrow{\Delta} \mathbb{R}_{>0}$ lo es. □

**Lema 2.18.** *Los subgrupos* $N, N^- \leq \mathrm{SL}_2(\mathbb{R})$ *generan* $\mathrm{SL}_2(\mathbb{R})$. *En particular,* $\mathrm{SL}_2(\mathbb{R})$ *está generado por* $N$ *y* $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$.

*Demostración.* Sea $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$. Podemos suponer que $c \neq 0$ reemplazando a $g$ por $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right) g$. Notemos que $N = \{h \in \mathrm{SL}_2(\mathbb{R}) : he_1 = e_1\}$ donde $e_1 = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$. Dados $s, t \in \mathbb{R}$, tenemos que

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ t \end{pmatrix} = \begin{pmatrix} 1+st \\ t \end{pmatrix}.$$

Como $c \neq 0$, existe $s \in \mathbb{R}$ tal que $1 + cs = a$. Para tal $s$, se tiene $\left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right) e_1 = \left(\begin{smallmatrix} a \\ c \end{smallmatrix}\right) = ge_1$. Luego $\left(\begin{smallmatrix} 1 & 0 \\ -c & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & -s \\ 0 & 1 \end{smallmatrix}\right) g = \left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right)$ para algún $x \in \mathbb{R}$, y por lo tanto $g$ pertenece al subgrupo generado por $N \cup N^-$. La última afirmación se deduce de que $N^-$ está en el subgrupo generado por $N$ y $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ ya que

$$\begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{-1}.$$

□

## 2.4. El flujo geodésico en superficies hiperbólicas

Sea $\Gamma \leq \mathrm{PSL}_2(\mathbb{R})$ un subgrupo discreto. Como $\Gamma$ actúa por isometrías en $\mathbb{H}$, se le puede dar al cociente una estructura de variedad riemanniana de modo que la proyección $\pi : \mathbb{H} \to \Gamma\backslash\mathbb{H}$ sea una isometría local. En efecto, dado $q \in \Gamma\backslash\mathbb{H}$, si elegimos $p \in \pi^{-1}(q)$ entonces $d_p\pi : T_p\mathbb{H} \to T_q(\Gamma\backslash\mathbb{H})$ es un isomorfismo y esto nos permite definir una forma bilineal positiva definida en $T_q(\Gamma\backslash\mathbb{H})$ de la siguiente forma:

$$\widetilde{Q}_q(v, w) = Q_p(d_p\pi^{-1}(v), d_p\pi^{-1}(w)).$$

Esta forma no depende del punto $p$ elegido ya que $\mathrm{PSL}_2(\mathbb{R})$ actúa por isometrías. Es fácil ver que la métrica $\widetilde{Q}$ así definida es suave y luego determina una estructura de variedad riemanniana en $\Gamma\backslash\mathbb{H}$. El objetivo de esta sección es probar que el flujo geodésico en $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$ es ergódico. Probaremos este resultado para retículos de $\mathrm{PSL}_2(\mathbb{R})$ ya que esto no hace más difícil la demostración.

Sea $G$ un grupo de Lie. Una métrica riemanniana $\langle\cdot,\cdot\rangle$ en $G$ se dice *invariante a izquierda* si $L_g$ es una isometría riemanniana para todo $g \in G$. Las métricas $G$-invariantes a izquierda están en correspondencia con los productos internos en $\mathfrak{g} = T_eG$ donde $e \in G$ es la identidad, ya que $\langle\cdot,\cdot\rangle_e$ determina $\langle\cdot,\cdot\rangle_g$ para todo $g \in G$. Si una métrica en $G$ es invariante a izquierda, la distancia $d = d_G$ asociada es invariante a izquierda, es decir verifica que

$$d(gx, gy) = d(x, y), \qquad \text{para todos } x, y, g \in G.$$

Sea $B_r^G$ la bola de radio $r$, centrada en la identidad de $G$.

**Proposición 2.19.** *Sea $G$ un grupo lineal cerrado y $\Gamma \leq G$ un subgrupo discreto. Entonces para todo $x \in X = \Gamma\backslash G$ existe un número $r > 0$ tal que el mapa $B_r^G \to B_r^X(x)$, $g \mapsto xg$ es una isometría biyectiva. Más aún, para todo compacto $K \subseteq X$ podemos elegir $r$ de modo que la propiedad anterior valga para todo $x \in K$.*

*Demostración.* Ver [EW11, Prop. 9.14]. □

El número $r > 0$ de la proposición anterior es llamado un *radio de inyectividad* de $x$. Cabe aclarar que esta no es la definición usual de radio de inyectividad en geometría riemanniana.

**Definición 2.20.** (1) Si $G$ es un grupo localmente compacto y Hausdorff, un *retículo* en $G$ es un subgrupo discreto $\Gamma \leq G$ tal que $G/\Gamma$ admite una medida de Borel $G$-invariante finita.

(2) Supongamos que $G$ actúa en un espacio $X$ localmente compacto y Hausdorff por homeomorfismos. Un *dominio fundamental* para la acción de $G$ es un abierto conexo $F \subseteq X$ tal que

   (*i*) $\bigcup_{g\in G} g\overline{F} = X$.

   (*ii*) $F \cap gF = \varnothing$ para todo $g \in G \setminus \{1\}$.

(3) Un *grupo Fuchsiano* es un subgrupo discreto de $\mathrm{PSL}_2(\mathbb{R})$.

En la definición de retículo uno puede tomar equivalentemente el espacio de coclases a derecha $\Gamma\backslash G$ ya que el mapa $G/\Gamma \to \Gamma\backslash G$, $g\Gamma \mapsto \Gamma g^{-1}$ es un homeomorfismo y luego induce una biyección entre medidas de Borel finitas $G$-invariantes a izquierda en $G/\Gamma$ y medidas de Borel finitas $G$-invariantes a derecha en $\Gamma\backslash G$.

**Proposición 2.21.** *Si $\Gamma \leq \mathrm{PSL}_2(\mathbb{R})$ es un retículo, entonces $T^1(\Gamma\backslash\mathbb{H}) \cong \Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$. Bajo este isomorfismo, el flujo geodésico en $T^1(\Gamma\backslash\mathbb{H})$ se corresponde con el flujo $g_t : \Gamma\backslash\mathrm{PSL}_2(\mathbb{R}) \to \Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$,*

$$g_t(\Gamma x) = \Gamma x a_t^{-1},$$

*donde $a_t$ es la matriz definida en (2.13).*

*Demostración.* Ver [BM00, Prop. II.3.4] □

**Lema 2.22.** *El conjunto*

$$F = \{z \in \mathbb{H} : |\mathrm{Re}(z)| < \tfrac{1}{2},\ |z| > 1\} \tag{2.14}$$

*es un dominio fundamental para la acción de $\mathrm{PSL}_2(\mathbb{Z})$ en $\mathbb{H}$.*

*Demostración.* Sea $\tau \in \mathbb{H}$. Como $\mathbb{Z} \oplus \mathbb{Z}\tau$ es una retículo en $\mathbb{C}$, existe $\min_{(c,d)\in\mathbb{Z}^2-\{0\}} |c\tau + d|$ y luego existe $\max_{g\in\mathrm{SL}_2(\mathbb{Z})} |\mathrm{Im}(g\cdot\tau)|$ por (2.8). Supongamos que $g_0 \in \mathrm{SL}_2(\mathbb{Z})$ maximiza $|\mathrm{Im}(g\cdot\tau)|$ y sea $z = g_0 \cdot \tau$. Componiendo a $g_0$ con $\left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)$ para algún $m \in \mathbb{Z}$ podemos suponer que $|\mathrm{Re}(z)| \leq \frac{1}{2}$. Si $z \notin \overline{F}$, entonces $|z| < 1$, en cuyo caso $\mathrm{Im}(-1/z) = \mathrm{Im}(z/|z|^2) > \mathrm{Im}(z)$, lo cual es absurdo. Por lo tanto $z \in \overline{F}$ y luego $F$ cumple la condición $(i)$ de la definición.

Para ver la segunda condición, supongamos que $z, w \in F$ son elementos distintos tales que $w = g(z)$ para algún $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$. Podemos suponer que $\mathrm{Im}(z) \leq \mathrm{Im}(w) = \mathrm{Im}(z)/|cz+d|^2$. Entonces

$$|c|.\mathrm{Im}(z) \leq |cz + d| \leq 1. \tag{2.15}$$

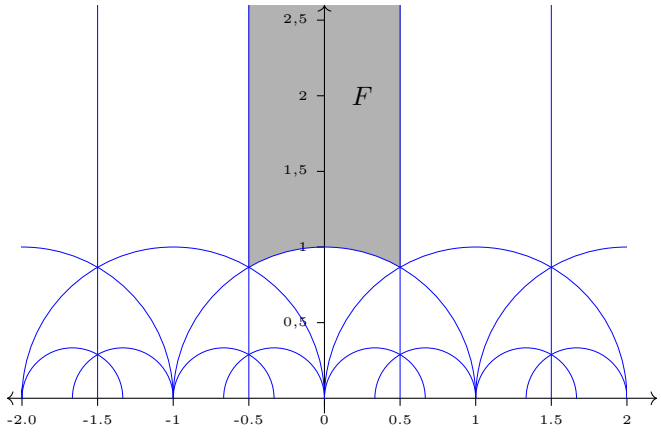


Figura 2.1: El dominio fundamental de $\mathrm{PSL}_2(\mathbb{Z})$

Como $z \in F$, se tiene $\mathrm{Im}(z) > \sqrt{3}/2$ y luego $|c| < 2/\sqrt{3}$. Además, si $c = 0$ entonces $a = d = \pm 1$ y $w = z \pm b$, absurdo ya que $|z - w| < 1$. Por lo tanto $c = \pm 1$ y por (2.15), $|z \pm d| \leq 1$. Si $d \neq 0$ entonces $|d \pm \mathrm{Re}(z)| > 1 - \frac{1}{2} = \frac{1}{2}$, así que

$$|z \pm d|^2 = \mathrm{Im}(z)^2 + (d \pm \mathrm{Re}(z))^2 > \tfrac{3}{4} + \tfrac{1}{4} = 1,$$

que es imposible. Se sigue que $d = 0$ y luego $|z| \leq 1$. Pero esto contradice $z \in F$. Por lo tanto $F$ es un dominio fundamental. □

Se sigue del lema anterior que el conjunto

$$S = \{(z, v) \in T^1\mathbb{H} : \, z \in F\} \tag{2.16}$$

es un dominio fundamental para la acción de $\mathrm{PSL}_2(\mathbb{Z})$ en $\mathrm{PSL}_2(\mathbb{R})$. En efecto, dado un punto $(z, v) \in T^1\mathbb{H}$, podemos llevarlo a $\overline{S}$ por medio de la acción de un elemento $g \in \mathrm{PSL}_2(\mathbb{Z})$ tal que $g(z) \subseteq \overline{F}$. Además $S \cap gS = \varnothing$ para todo $g \neq 1$ ya que para todo $p \in S$ las componentes $z$ de $p$ y $g \cdot p$ son distintas.

**Lema 2.23.** *La medida $d\mu = y^{-2}\, dx\, dy\, d\theta$ en $T^1\mathbb{H} \cong \mathbb{H} \times S^1$ es $\mathrm{PSL}_2(\mathbb{R})$-invariante a izquierda, donde $d\theta$ es la medida de Haar en $S^1$.*

*Demostración.* Recordemos que hay un difeomorfismo $T^1\mathbb{H} \cong \mathbb{H} \times S^1$, $(x + iy, v) \mapsto (x + iy, \xi)$ donde $\xi = v/y$. Bajo esta identificación, la acción asociada de $\mathrm{PSL}_2(\mathbb{R})$ en $\mathbb{H} \times S^1$ es de la forma

$$g.(z, \xi) = \left(g(z), \frac{v}{(cz + d)^2}\right) = \left(g(z), \frac{|cz + d|^2}{(cz + d)^2}\xi\right).$$

En particular vemos que para $z \in \mathbb{H}$ fijo, la acción de $g$ en la segunda coordenada es una rotación. Además, la medida $\mu$ coincide con la medida producto $\mu_0 \times \lambda$, donde $\mu_0$ es la medida del lema 2.14 y $\lambda$ es la medida de Haar de $S^1$. Si $f \in C_c(\mathbb{H} \times S^1)$, por Fubini tenemos que

$$\begin{aligned}
\int_{\mathbb{H}\times S^1} f(g.(z,\xi))\, d\mu &= \int_{\mathbb{H}} \int_{S^1} f\left(g(z), \frac{|cz+d|^2}{(cz+d)^2}\xi\right)\, d\lambda(\xi) d\mu_0(z) \\
&= \int_{\mathbb{H}} \int_{S^1} f\left(g(z), \xi\right)\, d\lambda(\xi)\, d\mu_0(z) \\
&= \int_{S^1} \int_{\mathbb{H}} f(g(z), \xi)\, d\mu_0(z)\, d\lambda(\xi) \\
&= \int_{S^1} \int_{\mathbb{H}} f(z, \xi)\, d\mu_0(z)\, d\lambda(\xi) = \int_{\mathbb{H}\times S^1} f\, d\mu.
\end{aligned}$$

Por lo tanto $\mu$ es $\mathrm{PSL}_2(\mathbb{R})$-invariante. □

**Proposición 2.24.** $\mathrm{PSL}_2(\mathbb{Z})$ *es un retículo de* $\mathrm{PSL}_2(\mathbb{R})$.

*Demostración.* Sea $G = \mathrm{PSL}_2(\mathbb{R})$ y sea $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$. Como $G$ es unimodular, su medida de Haar a izquierda $\mu$ es bi-invariante, así que por 1.20 se corresponde con una medida de Radon $\nu$ en $\Gamma\backslash G$ que es $G$-invariante a derecha. Estas medidas se relacionan por la ecuación

$$\int_G f\,d\mu = \int_{\Gamma\backslash G} \sum_{\gamma\in\Gamma} f(\gamma x)\,d\nu(\Gamma x),$$

para toda $f \in C_c(G)$. Sea $I : C_c(G) \to C_c(\Gamma\backslash G)$ el mapa $I(f)(\Gamma x) = \sum_{\gamma\in\Gamma} f(\gamma x)$ definido en la sección 1.1. Dada una función $h \in C_c(\Gamma\backslash G)$, podemos elegir $f \in C_c(G)$ tal que $I(f) = h$. Como $G = \bigcup_{\gamma\in\Gamma} \gamma\overline{S}$ y $\mu(\partial S) = 0$, tenemos que

$$\begin{aligned}\mu(f) &= \sum_{\gamma\in\Gamma}\int_{\gamma S} f(g)\,d\mu(g)\\ &= \sum_{\gamma\in\Gamma}\int_{S} f(\gamma g)\,d\mu(g)\\ &= \int_S \sum_{\gamma\in\Gamma} f(\gamma g)\,d\mu(g) = \int_S h(\Gamma g)\,d\mu(g).\end{aligned}$$

Luego $\nu(h) = \int_S (h\circ\pi)\,d\mu$ para toda $h \in C_c(\Gamma\backslash G)$, donde $\pi : G \to \Gamma\backslash G$ es la proyección al cociente. Se sigue que $\nu = (\pi|_S)_*\mu$, y por el lema anterior, $\mu$ es proporcional a $y^{-2}dx\,dy\,d\theta$. Entonces, salvo por un factor de proporcionalidad la medida de $\Gamma\backslash G$ es

$$\nu(\Gamma\backslash G) = \mu(S) = \int_F \frac{1}{y^2}\,dx\,dy = \int_{-1/2}^{1/2}\int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\,dy\,dx = \frac{\pi}{3} < +\infty.$$

Luego $\Gamma$ es un retículo. □

**Lema 2.25.** *Sea* $g = \left(\begin{smallmatrix} r & 0 \\ 0 & 1/r \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$ *con* $r > 0$. *Entonces*

(a) *Si* $h \in N$ *y* $r < 1$ *entonces*

$$\lim_{n\to\infty} g^n h g^{-n} = 1.$$

(b) *Si* $h \in N^-$ *y* $r > 1$ *entonces*

$$\lim_{n\to\infty} g^n h g^{-n} = 1.$$

*Demostración.* Supongamos que $h = \left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right) \in N$. Entonces

$$\begin{pmatrix} r^n & 0 \\ 0 & r^{-n} \end{pmatrix}\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}\begin{pmatrix} r^{-n} & 0 \\ 0 & r^n \end{pmatrix} = \begin{pmatrix} 1 & r^{2n}s \\ 0 & 1 \end{pmatrix}.$$

Si $r \in (0,1)$ entonces $r^{2n}s \to 0$ cuando $n \to \infty$ y luego $g^n h g^{-n} \to 1$. Análogamente, si $h = \left(\begin{smallmatrix} 1 & 0 \\ s & 1 \end{smallmatrix}\right)$ entonces $g^n h g^{-n} = \left(\begin{smallmatrix} 1 & 0 \\ sr^{-2n} & 1 \end{smallmatrix}\right) \to 1$ siempre que $r > 1$. □

**Definición 2.26.** Sea $G$ un grupo localmente compacto y Hausdorff. Una *representación unitaria* de $G$ es un par $(\pi, \mathscr{H})$ donde $\mathscr{H}$ es un espacio de Hilbert y $\pi : G \to \mathscr{U}(\mathscr{H})$ es un morfismo de $G$ en el grupo de operadores unitarios, que es SOT-continua, i.e. para todo $v \in \mathscr{H}$, el mapa $g \mapsto \pi(g)v$ es continuo.

**Lema 2.27** (Mautner)**.** *Sea* $G$ *un grupo localmente compacto y Hausdorff, y sea* $(\pi, \mathscr{H}_\pi)$ *una representación unitaria de* $G$. *Supongamos que* $g, h \in G$ *verifican*

$$\lim_{n\to\infty} g^n h g^{-n} = 1.$$

*Entonces todo vector* $v \in \mathscr{H}_\pi$ *que es fijado por* $g$ *también es fijado por* $h$.

*Demostración.* Para todo $n \geq 1$ tenemos

$$\|\pi(h)v - v\| = \|\pi(h)\pi(g^{-n})v - \pi(g^{-n})v\| = \|\pi(g^n h g^{-n})v - v\|.$$

Como $g^n h g^{-n} \to 1$, se sigue que $\pi(g^n h g^{-n})v \to v$. Por lo tanto $\pi(h)v = v$. □

Sea $\Gamma \leq \mathrm{PSL}_2(\mathbb{R})$ un retículo. Como $\mathrm{PSL}_2(\mathbb{R})$ es unimodular, su medida de Haar a izquierda $\lambda$ es bi-invariante. Por la proposición 1.20, $\lambda$ se corresponde con una medida $\mathrm{PSL}_2(\mathbb{R})$-invariante a derecha $\mu$ en $X = \Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$.

El grupo $\mathrm{SL}_2(\mathbb{R})$ actúa a derecha en $X$ por traslación. Esta acción induce una representación unitaria $(\pi, L^2_\mu(X))$ dada por $\pi(g)f(x) = f(xg)$. El operador $\pi(g)$ es unitario ya que $\mu$ es invariante a derecha y $\pi(g)$ es inversible. Para ver que $\pi$ es SOT-continua, supongamos que $g_n \to g$ en $\mathrm{SL}_2(\mathbb{R})$. Si $f \in C_c(X)$, entonces $f(xg_n) \to f(xg)$ puntualmente ya que la acción de $G$ en $X$ es continua. Por convergencia acotada se sigue que $\pi(g_n)f \to \pi(g)f$ en $L^2_\mu(X)$. Como $C_c(X)$ es denso en $L^2_\mu(X)$, lo mismo vale para toda $f \in L^2_\mu(X)$.

**Teorema 2.28.** *Sea $\Gamma \leq \mathrm{PSL}_2(\mathbb{R})$ un retículo. Entonces todo $g \in A \setminus \{1\}$ actúa ergódicamente en $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$. En particular, el flujo geodésico en $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$ es ergódico.*

*Demostración.* Sea $\mathscr{H} = L^2(\Gamma\backslash\mathrm{PSL}_2(\mathbb{R}))$ y sea $\pi : \mathrm{SL}_2(\mathbb{R}) \to \mathscr{U}(\mathscr{H})$ la representación unitaria asociada a la acción de $\mathrm{SL}_2(\mathbb{R})$ en $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$. Si $f \in \mathscr{H}$ es una función $g$-invariante para algún $g \in A$, por los lemas 2.25 y 2.27 se sigue que $\pi(N)f = f$ y $\pi(N^-)f = f$. Como $N$ y $N^-$ generan a $\mathrm{SL}_2(\mathbb{R})$, la función $f$ es esencialmente $\mathrm{SL}_2(\mathbb{R})$-invariante y luego constante c.t.p. ya que la acción de $\mathrm{SL}_2(\mathbb{R})$ es transitiva. □

# Capítulo 3

# Formas cuadráticas binarias

## 3.1. Formas cuadráticas y espacios cuadráticos

*En esta sección, $K$ denotará un cuerpo de característica distinta de 2, a menos que se especifique lo contrario.*

**Definición 3.1.** Sea $K$ un cuerpo de caraterística $\neq 2$. Un *espacio cuadrático* sobre $K$ es un par $(V, q)$ que consiste de un $K$-espacio vectorial $V$ de dimensión finita y una forma cuadrática $q$ en $V$, es decir, una función $q : V \to K$ que cumple

(*i*) $q(\lambda v) = \lambda^2 q(v)$, para todo $\lambda \in K$ y todo $v \in V$,

(*ii*) la función $\phi : V \times V \to K$ dada por

$$\phi(x, y) = q(x + y) - q(x) - q(y)$$

es una forma bilineal simétrica.

Decimos que $q$ es *$n$-aria* si $\dim(V) = n$. Usualmente notaremos la forma bilineal como $\langle \cdot, \cdot \rangle$.

Notemos que la función $q$ está completamente determinada por $\phi$ ya que $q(x) = \frac{1}{2}\phi(x, x)$. Además, si elegimos una base $e_1, \ldots, e_n$ de $V$ como $K$-espacio vectorial, esto determina un isomorfismo $\rho : K^n \xrightarrow{\sim} V$, $\rho(a_1, \ldots, a_n) = \sum_i a_i e_i$ y puesto que $\phi$ es bilineal,

$$q \circ \rho(a_1, \ldots, a_n) = q\left(\sum_{i=1}^{n} a_i e_i\right) = \frac{1}{2} \sum_{1 \leq i,j \leq n} a_i a_j \phi(e_i, e_j). \tag{3.1}$$

Entonces $q \circ \rho$ es una forma cuadrática en el sentido usual, es decir un polinomio homogéneo de grado 2 en $K$. Recíprocamente, si $f \in K[x_1, \ldots, x_n]$ es un polinomio homogéneo de grado 2 entonces $q = f \circ \rho^{-1}$ cumple las condiciones $(i)$ y $(ii)$ de la definición anterior.

Otra forma de caracterizar a $q$ es por medio de su matriz asociada. Si fijamos una base $e_1, \ldots, e_n$ de $V$, entonces existe una única matriz $M_q \in M_n(K)$ simétrica tal que

$$q(x_1 e_1 + \cdots + x_n e_n) = \frac{1}{2} x^t M_q x,$$

donde $x = (x_1, \ldots, x_n)^t$, para cualquier elección de $x_i$. De hecho $M_q = (\langle e_i, e_j \rangle)_{i,j}$. Definimos el *discriminante* de $q$ con respecto a esta base como $\operatorname{disc}(q) = \det M_q$. Si $q$ es una forma cuadrática binaria con coeficientes en $K$, tomando la base canónica de $K^2$ recuperamos la noción de discriminante definida anteriormente.

Un *morfismo de espacios cuadráticos* $f : (V_1, q_1) \to (V_2, q_2)$ es una transformación lineal $f : V_1 \to V_2$ tal que $q_1 = q_2 \circ f$. Llamamos *embedding* de espacios cuadráticos a un morfismo inyectivo.

El *grupo ortogonal* de $q$, notado $\mathrm{O}(q)$, es el grupo de automorfismos $K$-lineales $T : V \to V$ que preservan a $q$, es decir, tales que $q \circ T = q$. A su vez definimos $\mathrm{SO}(q)$ como el subgrupo de

elementos de $\mathrm{O}(q)$ de determinante 1. También notamos a estos grupos como $\mathrm{O}_q(K)$ y $\mathrm{SO}_q(K)$ si queremos especificar el cuerpo $K$.

Decimos que $q$ es *no degenerada*, o que $V$ es *regular*, si para todo $x \in V$ no nulo existe $y \in V$ tal que $\langle x, y\rangle \neq 0$. Notemos que $q$ es no degenerada si y solo si $\mathrm{disc}(q) \neq 0$. En efecto, $q$ determina una transformación lineal $\beta : V \to V^* = \mathrm{Hom}_K(V, K)$ dada por $x \mapsto \langle x, -\rangle$. Sea $e_1, \ldots, e_n$ la base con respecto a la cual calculamos el discriminante y sea $e_1^*, \ldots, e_n^*$ la base dual en $V^*$. Entonces la matriz de $\beta$ con respecto a estas bases es $M_q$ ya que

$$\langle e_i, x\rangle = \left\langle e_i, \sum_{j=1}^{n} e_j^*(x) e_j \right\rangle = \sum_{j=1}^{n} \langle e_i, e_j\rangle e_j^*(x)$$

para todo $x \in V$. Por definición $q$ es no degenerada si y solo si $\beta$ es un monomorfismo. Como $V$ tiene dimensión finita, esto es equivalente a pedir $\det M_q \neq 0$. Notemos ademas que en ese caso $\beta$ es un isomorfismo. En particular, para toda funcional lineal $\psi \in V^*$ existe $v \in V$ tal que $\psi = \langle v, -\rangle$.

Decimos que un subespacio $S \subseteq V$ es *regular* si $q|_S$ es no degenerada. Además definimos el *complemento ortogonal* de $S$ como $S^\perp = \{y \in V : \langle x, y\rangle = 0, \text{ para todo } x \in S\}$. Supongamos que $V$ es regular y sea $S$ un subespacio. Afirmamos que $\dim(S^\perp) = \dim(V) - \dim(S)$. En efecto, $S^\perp$ es el núcleo del morfismo $V \xrightarrow{\beta} V^* \xrightarrow{\rho} S^*$, donde $\rho$ denota la restricción a $S$. El morfismo $\rho \circ \beta$ es un epimorfismo ya que $\rho$ y $\beta$ lo son. Entonces tenemos una sucesión exacta corta

$$0 \longrightarrow S^\perp \longrightarrow V \xrightarrow{\rho\beta} S^* \longrightarrow 0$$

de donde se deduce que $\dim V = \dim S^* + \dim S^\perp = \dim S + \dim S^\perp$.

**Proposición 3.2.** *Sea $(V, q)$ un espacio cuadrático y sea $S \subseteq V$ un subespacio. Entonces*

(a) *Si $S$ es regular, entonces $V = S \oplus S^\perp$.*
(b) *Supongamos que $V$ es regular. Entonces $S$ es regular si y solo si $S^\perp$ es regular.*

*Demostración.* (a) Dado $v \in V$, consideremos la funcional lineal $\psi_v : S \to K$, $\psi_v(x) = \langle v, x\rangle$. Como $S$ es regular, existe $s \in S$ tal que $\psi_v(x) = \langle s, x\rangle$ para todo $x \in S$. Se sigue que $v - s \in S^\perp$ y luego $v = s + (v - s) \in S + S^\perp$. Por otro lado, si $w \in S \cap S^\perp$, esto significa que la funcional $\langle w, -\rangle$ se anula en $S$. Como $S$ es regular, esto solo puede pasar si $w = 0$. Por lo tanto $V = S \oplus S^\perp$.

(b) Supongamos que $S$ es regular. Por el punto $(a)$, tenemos $V = S \oplus S^\perp$. Dado $y \in S^\perp$ existe $x \in V$ tal que $\langle x, y\rangle \neq 0$. Escribiendo $x = s + w$ con $s \in S$, $w \in S^\perp$, obtenemos que $\langle x, y\rangle = \langle w, y\rangle \neq 0$. Se sigue que $S^\perp$ es regular. Recíprocamente, si $S$ no es regular, entonces por definición existe $y \in S \cap S^\perp$ no nulo. Claramente $S \subseteq (S^\perp)^\perp$, así que $S^\perp \cap (S^\perp)^\perp \neq 0$. Luego por el punto $(a)$, $S^\perp$ no puede ser regular. □

**Teorema 3.3** (Teorema de Extensión de Witt)**.** *Sean $(V_1, q_1)$ y $(V_2, q_2)$ espacios cuadráticos isométricos, sea $S \subseteq V_1$ un subespacio y supongamos que hay un embedding $f : (S, q_1) \to (V_2, q_2)$. Entonces $f$ se extiende a una isometría $\overline{f} : (V_1, q_1) \to (V_2, q_2)$.*

*Demostración.* Ver [Cas78, Teo. 4.1] o [Ser73, §IV.1.5]. □

Un elemento $v$ de un espacio cuadrático $(V, q)$ se dice *isotrópico* si $v \neq 0$ y $q(v) = 0$. Decimos que $q$ es *isotrópica* si es no degenerada y $V$ contiene algún elemento isotrópico. En caso contrario, $q$ se dice *anisotrópica*. Un ejemplo de espacio isotrópico es un *plano hiperbólico*, que se define como un espacio $(H, q)$ de dimensión 2 que posee una base $\{u_1, u_2\}$ tal que $q(u_1) = q(u_2) = 0$ y $\langle u_1, u_2\rangle = 0$. El siguiente lema muestra que el plano hiperbólico es en cierto sentido el único ejemplo de espacio isotrópico no trivial.

**Lema 3.4.** *Todo espacio cuadrático isotrópico $(V, q)$ contiene un plano hiperbólico.*

*Demostración.* Sea $u_1 \in V$ un elemento no nulo tal que $q(u_1) = 0$. Como $q$ es no degenerada, existe $w \in V$ tal que $\langle u_1, w\rangle \neq 0$. Multiplicando a $w$ por una constante podemos suponer que $\langle u_1, w\rangle = 1$. Ahora tomamos $u_2 = w + \lambda u_1$, donde $\lambda = -q(w)$. Entonces

$$q(u_2) = q(w) + \lambda^2 q(u_1) + \lambda\langle w, u_1\rangle = q(w) + \lambda = 0.$$

Además $\langle u_1, u_2\rangle = \langle u_1, w\rangle + \lambda\langle u_1, u_1\rangle = 1$. Por lo tanto el subespacio $H = Ku_1 + Ku_2$ es un plano hiperbólico. □

Finalmente, damos un par de definiciones sobre formas cuadráticas binarias. Diremos que una forma cuadrática $q$ sobre $\mathbb{Q}$ es *entera* si sus coeficientes son enteros. Decimos que dos formas cuadráticas binarias enteras $q_1, q_2$ son *equivalentes* si existe $g \in \mathrm{GL}_2(\mathbb{Z})$ tal que $q_2 = q_1 \circ g$, y decimos que son *propiamente equivalentes* si podemos tomar $g \in \mathrm{SL}_2(\mathbb{Z})$ en cambio. Llamamos *retículo cuadrático* a un par $(L, q)$ donde $L$ es un $\mathbb{Z}$-módulo libre de rango finito y $q$ es una forma cuadrática en $L \otimes_{\mathbb{Z}} \mathbb{Q}$.

A cada forma $q = ax^2 + bxy + cy^2$ con $a \neq 0$ le podemos asignar el número complejo

$$\tau_q = \frac{-b + \sqrt{d}}{2a},$$

que es una raíz del polinomio $f(x) = q(x, 1)$, donde $d$ es el discriminante de $q$. Llamamos a $\tau_q$ la *raíz principal* de $q$. Si $d < 0$ y $a > 0$, esta raíz tiene parte imaginaria positiva, es decir, $\tau_q \in \mathbb{H}$. El conjunto $\{\tau_q : q \in R_{\mathrm{disc}}(d)\}$ se suele denominar el conjunto de *puntos de Heegner* de discriminante $d < 0$.

## 3.2. Órdenes en cuerpos cuadráticos

Una de las razones por las que el estudio de las formas cuadráticas tiene mucha relevancia en teoría de números es la correspondencia entre clases de formas cuadráticas binarias enteras de discriminante $d$ fijo, y cierto grupo de clases de ideales en un anillo $\mathcal{O}_d \subseteq K$. El objetivo principal de esta sección es demostrar esta correspondencia.

**Definición 3.5.** Un *cuerpo de números* es un cuerpo $K$ que es una extensión finita de $\mathbb{Q}$. Un *cuerpo cuadrático* es una extensión de $\mathbb{Q}$ de grado 2.

Decimos que un elemento $\alpha \in K$ es un *entero algebraico* si el subanillo $\mathbb{Z}[\alpha] \subseteq K$ es finitamente generado como $\mathbb{Z}$-módulo. Si $\alpha$ es un entero algebraico, como las potencias $\alpha^n$ con $n \geq 0$ forman un conjunto de generadores de $\mathbb{Z}[\alpha]$ como $\mathbb{Z}$-módulo, existe un entero $N \geq 0$ tal que $\mathbb{Z}[\alpha] = \sum_{n=0}^{N} \mathbb{Z}\alpha^n$. En particular, $\alpha^{N+1}$ se escribe como combinación $\mathbb{Z}$-lineal de $1, \alpha, \ldots, \alpha^N$ y luego $\alpha$ es raíz de un polinomio mónico con coeficientes enteros. Recíprocamente, si $\alpha$ es raíz de un polinomio mónico $f \in \mathbb{Z}[x]$ de grado $d > 0$, es fácil ver que $\mathbb{Z}[\alpha] = \sum_{n=0}^{d-1} \mathbb{Z}\alpha^n$ es finitamente generado. En conclusión estas dos condiciones son equivalentes.

Observemos que si $\alpha, \beta \in K$ son enteros algebraicos entonces $\mathbb{Z}[\alpha, \beta]$ es un $\mathbb{Z}$-módulo finitamente generado. En efecto, si $\mathbb{Z}[\alpha] = \mathbb{Z} + \mathbb{Z}\alpha + \cdots + \mathbb{Z}\alpha^r$ y $\mathbb{Z}[\beta] = \mathbb{Z} + \mathbb{Z}\beta + \cdots + \mathbb{Z}\beta^s$ entonces todo monomio $\alpha^m\beta^n$ se escribe como combinación lineal de elementos $\alpha^j\beta^k$ con $0 \leq j \leq r$, $0 \leq k \leq s$, y luego $\mathbb{Z}[\alpha, \beta]$ está generado por el conjunto $\{\alpha^j\beta^k : 0 \leq j \leq r, 0 \leq k \leq s\}$ que es finito. En particular $\alpha + \beta$ y $\alpha\beta$ son enteros algebraicos ya que $\mathbb{Z}[\alpha + \beta] \subseteq \mathbb{Z}[\alpha, \beta]$ y $\mathbb{Z}[\alpha\beta] \subseteq \mathbb{Z}[\alpha, \beta]$. Se sigue que el conjunto $\mathcal{O}_K$ de enteros algebraicos en $K$ es un anillo, y lo llamamos el *anillo de enteros* de $K$. Se puede ver que $\mathcal{O}_K$ es un $\mathbb{Z}$-módulo libre de rango $[K : \mathbb{Q}]$ (ver [Neu99, §I.2]). En el caso de cuerpos cuadráticos, esto se deducirá de la proposición 3.6.

Sea $K$ un cuerpo cuadrático. Por la teoría de cuerpos, toda extensión algebraica de $\mathbb{Q}$ puede ser embebida como subcuerpo de $\mathbb{C}$, ya que $\mathbb{C}$ es algebraicamente cerrado. Luego podemos suponer por simplicidad que $K \subseteq \mathbb{C}$. Tomemos un elemento $\xi \in K - \mathbb{Q}$. Como $[K : \mathbb{Q}] = 2$, los elementos $1, \xi, \xi^2$ son linealmente dependientes y luego existe un polinomio $ax^2 + bx + c \in \mathbb{Q}[x]$ que se anula en $\xi$. Tomando $d = b^2 - 4ac$ tenemos que

$$4a(ax^2 + bx + c) = (2ax + b)^2 + 4ac - b^2 = (2ax + b)^2 - d.$$

Evaluando en $\xi$ obtenemos que $(2a\xi+b)^2 = d$, así que $2a\xi+b = \pm\sqrt{d}$, donde $\sqrt{d} = i|d|^{1/2}$ si $d < 0$. En particular $d$ no es un cuadrado perfecto ya que $2a\xi+b \notin \mathbb{Q}$. Se sigue que $\mathbb{Q}(\sqrt{d}) \subseteq K$ y se debe dar la igualdad ya que ambos cuerpos tienen grado 2 sobre $\mathbb{Q}$. Notemos que $\mathbb{Q}(\sqrt{m^2 d}) = \mathbb{Q}(\sqrt{d})$ para cualquier $m \in \mathbb{Q}$ no nulo. Luego podemos suponer que $d \in \mathbb{Z}$. De hecho, dividiendo a $d$ por el mayor cuadrado entero $m^2$ que lo divide, podemos suponer que $d$ es libre de cuadrados.

Si $\sigma : K \to K$ es un morfismo de cuerpos, es fácil ver que $\sigma$ fija a $\mathbb{Q}$ puntualmente. Luego

$$\sigma(\sqrt{d})^2 = \sigma\left(\sqrt{d}^2\right) = \sigma(d) = d$$

así que $\sigma(\sqrt{d}) = \pm\sqrt{d}$. En el primer caso $\sigma(a + b\sqrt{d}) = a + b\sqrt{d}$ para cualquier par de racionales $a, b$, así que $\sigma$ es la identidad. En el segundo caso

$$\sigma(a + b\sqrt{d}) = a - b\sqrt{d}$$

para todos $a, b \in \mathbb{Q}$. Esto caracteriza a $\sigma$ ya que $\{1, \sqrt{d}\}$ es una base de $\mathbb{Q}(\sqrt{d})$ como $\mathbb{Q}$-espacio vectorial. De ahora en adelante notamos por $\beta \mapsto \beta'$ a este último automorfismo de $K$. Es el único automorfismo de $K$ que no es la identidad. Es claro que un elemento $\beta \in K$ cumple $\beta' = \beta$ si y solo si $\beta \in \mathbb{Q}$. Para cada $\beta \in K$ llamamos a $\beta'$ el *conjugado* de $\beta$. La norma y traza de un elemento $\beta \in K$ están dadas por $N(\beta) = \beta\beta'$ y $\mathrm{tr}(\beta) = \beta + \beta'$ respectivamente. Recordemos que $N(\beta)$ y $\mathrm{tr}(\beta)$ son el determinante y la traza de la transformación lineal $L_\beta : K \to K$, $x \mapsto \beta x$, respectivamente. En particular, $N(\beta)$ es una función multiplicativa en $\beta$, $\mathrm{tr}(\beta)$ es aditiva y ambas tienen imagen en $\mathbb{Q}$.

Si $L \subseteq K$ es un retículo en un cuerpo cuadrático, definimos su *discriminante* $\Delta(L)$ de la siguiente manera: si $\omega_1, \omega_2$ es una base de $L$ como $\mathbb{Z}$-módulo entonces

$$\Delta(L) = \begin{vmatrix} \omega_1 & \omega_1' \\ \omega_2 & \omega_2' \end{vmatrix}^2 = (\omega_1\omega_2' - \omega_1'\omega_2)^2.$$

Esta expresión es independiente de la base elegida, ya que si $\eta_1, \eta_2$ es otra base, existe una matriz $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z})$ tal que

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}.$$

En ese caso

$$\begin{pmatrix} \eta_1 & \eta_1' \\ \eta_2 & \eta_2' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 & \omega_1' \\ \omega_2 & \omega_2' \end{pmatrix}$$

y tomando determinante en ambos lados vemos que el discriminante calculado con respecto a las dos bases es el mismo. Observemos además que $\Delta(L) \in \mathbb{Q}$ ya que es invariante bajo conjugación. Notamos el discriminante de $\mathcal{O}_K$ como $d_K$ y lo llamamos el *discriminante de* $K$. Dados elementos $\alpha_1, \ldots, \alpha_n \in K$ notamos al $\mathbb{Z}$-módulo generado por $\alpha_1, \ldots, \alpha_n$ como $[\alpha_1, \ldots, \alpha_n]$.

**Proposición 3.6.** *Sea $m \neq 0, 1$ un entero libre de cuadrados y sea $K = \mathbb{Q}(\sqrt{m})$.*

(a) *Si $m \equiv 2, 3 \pmod 4$ entonces $\mathcal{O}_K = \mathbb{Z}[\sqrt{m}]$ y $d_K = 4m$.*

(b) *Si $m \equiv 1 \pmod 4$ entonces $\mathcal{O}_K = \mathbb{Z}\left[\frac{1+\sqrt{m}}{2}\right]$ y $d_K = m$.*

*Demostración.* Ver [IR90, §13.1]. □

Llamamos *discriminante fundamental* a un entero $d$ que es el discriminante de algún cuerpo cuadrático $K$. Por la proposición anterior, vemos que $d \in \mathbb{Z}$, $d \neq 0, 1$ es un discriminante fundamental si y solo si $d \equiv 1 \pmod 4$ y es libre de cuadrados, o bien $4 \mid d$ y $\frac{d}{4} \equiv 2, 3 \pmod 4$ es libre de cuadrados.

**Corolario 3.7.** *Si $K$ es un cuerpo cuadrático entonces $\mathcal{O}_K = \mathbb{Z}\left[\frac{d_K+\sqrt{d_K}}{2}\right]$.*

**Definición 3.8.** Un *orden* en un cuerpo de números $K$ es un subanillo $\mathcal{O} \subseteq K$ que es un $\mathbb{Z}$-módulo libre y que genera a $K$ como espacio vectorial sobre $\mathbb{Q}$.

Un ejemplo de orden es el anillo de enteros $\mathcal{O}_K$ de $K$. De hecho este es el orden más grande posible: en efecto, si $\mathcal{O} \subseteq K$ es un orden, para todo $\xi \in \mathcal{O}$ el anillo $\mathbb{Z}[\xi] \subseteq \mathcal{O}$ es finitamente generado como $\mathbb{Z}$-módulo así que $\xi$ debe ser un entero algebraico. Esto demuestra que $\mathcal{O} \subseteq \mathcal{O}_K$.

Si $\mathcal{O}$ es un orden entonces cualquier ideal $I \subseteq \mathcal{O}$ no nulo debe ser un retículo. En efecto, $I$ es un subgrupo del $\mathbb{Z}$-módulo libre $\mathcal{O}$, así que es un $\mathbb{Z}$-módulo libre también. Si tomamos $\alpha \in I$ no nulo, entonces el mapa $\mathcal{O} \to I$, $x \mapsto \alpha x$ es un monomorfismo ya que $\mathcal{O}$ es un dominio, lo que implica que el rango de $I$ es mayor o igual al de $\mathcal{O}$. Se sigue que $I$ tiene rango máximo $[K : \mathbb{Q}]$. En particular $\mathcal{O}/I$ es un grupo finito. Definimos la *norma* de un ideal $I \neq 0$ como $N(I) = |\mathcal{O}/I|$. El siguiente lema relaciona la norma con el discriminante:

**Lema 3.9.** $N(I) = \left|\frac{\Delta(I)}{\Delta(\mathcal{O})}\right|^{1/2}$.

*Demostración.* Ver [Neu99, Prop. I.2.12]. □

**Definición 3.10.** Sea $K$ un cuerpo de números y sea $\mathcal{O}$ un orden de $K$. El *conductor* de $\mathcal{O}$ se define como

$$\mathfrak{f} = \{x \in K : x\mathcal{O}_K \subseteq \mathcal{O}\}. \tag{3.2}$$

Notemos en primer lugar que $\mathfrak{f} \subseteq \mathcal{O}$ ya que $1 \in \mathcal{O}_K$. Además $\mathfrak{f}$ es un ideal de $\mathcal{O}_K$ ya que si $x \in \mathfrak{f}$ y $\alpha \in \mathcal{O}_K$ entonces $x\alpha\mathcal{O}_K \subseteq x\mathcal{O}_K \subseteq \mathcal{O}$. En particular es un ideal de $\mathcal{O}$.

**Lema 3.11.** *Sea $\mathcal{O}$ un orden en un cuerpo cuadrático $K$ de discriminante $d_K$, y sea $\alpha = \frac{d_K+\sqrt{d_K}}{2}$. Entonces existe un entero $f > 0$ tal que*

$$\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K = [1, f\alpha].$$

*Más aún, en ese caso el conductor de $\mathcal{O}$ es $f\mathcal{O}_K$ y $\operatorname{disc}(\mathcal{O}) = f^2 d_K$.*

*Demostración.* Consideremos el índice $f = (\mathcal{O}_K : \mathcal{O})$ . Este índice es finito ya que tanto $\mathcal{O}_K$ como $\mathcal{O}$ son $\mathbb{Z}$-módulos libres de rango 2. Es claro que $f\mathcal{O}_K \subseteq \mathcal{O}$ y luego $\mathbb{Z} + f\mathcal{O}_K \subseteq \mathcal{O}$. Por otro lado, $\mathcal{O}_K = [1, \alpha]$ y $\mathbb{Z} + f\mathcal{O}_K = [1, f\alpha]$ por el corolario 3.7, así que

$$(\mathcal{O}_K : \mathbb{Z} + f\mathcal{O}_K) = f = (\mathcal{O}_K : \mathcal{O}).$$

Se sigue que $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$. El discriminante de este orden se puede computar usando la base $\{1, f\alpha\}$:

$$\operatorname{disc}(\mathcal{O}) = \begin{vmatrix} 1 & 1 \\ f\alpha & f\alpha' \end{vmatrix}^2 = f^2 \begin{vmatrix} 1 & 1 \\ \alpha & \alpha' \end{vmatrix}^2 = f^2 d_K.$$

Finalmente, veamos que el conductor $\mathfrak{f}$ de $\mathcal{O}$ es $f\mathcal{O}_K$. Como $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$, es claro que $f\mathcal{O}_K \subseteq \mathfrak{f}$. Recíprocamente, sea $\beta = m + n\alpha \in \mathfrak{f}$, donde $m, n \in \mathbb{Z}$. Como $\beta \in \mathcal{O} = [1, f\alpha]$, tenemos que $f \mid n$. Además, $\beta\alpha \in \mathcal{O}$. Si $x^2 + bx + c$ es el polinomio minimal de $\alpha$, entonces

$$\begin{aligned} \beta\alpha &= m\alpha + n\alpha^2 = m\alpha - n(b\alpha + c) \\ &= -nc + (m - nb)\alpha \in [1, f\alpha]. \end{aligned}$$

Se sigue que $f \mid m - nb$ y luego $f \mid m$. Por lo tanto $\beta \in f\mathcal{O}_K$. □

En vista del lema anterior también se suele llamar conductor al entero $f = (\mathcal{O}_K : \mathcal{O})$.

**Corolario 3.12.** *Para todo entero $d \equiv 0, 1 \pmod 4$, $d \neq 0, 1$, el orden $\mathcal{O}_d = \mathbb{Z}\left[\frac{d+\sqrt{d}}{2}\right]$ es el único orden cuadrático de discriminante $d$*

*Demostración.* Dado $d \in \mathbb{Z}$ como en el enunciado, se puede escribir como $d = mr^2$ con $m, r \in \mathbb{Z}$ y $m$ libre de cuadrados. Entonces el cuerpo $K = \mathbb{Q}(\sqrt{m})$ tiene discriminante $d_K = m$ o $d_K = 4m$. Si $d_K = 4m$, entonces $m \equiv 2, 3 \pmod 4$. Como $d \equiv 0, 1 \pmod 4$, necesariamente $r$ debe ser par en este caso. Se sigue que en ambos casos $d = f^2 d_K$ para algún entero $f > 0$.

Por otro lado notemos que $d$ se factoriza de manera única como $d = f^2 d_0$, donde $d_0$ es un discriminante fundamental y $f > 0$. Se sigue que si $\mathcal{O}$ es un orden de discriminante $d$ en un cuerpo cuadrático $K$, entonces $d_0 = d_K$ y $\mathcal{O}$ tiene conductor $f$. En particular, dos órdenes con discriminante $d$ deben estar contenidos en el mismo cuerpo cuadrático y tienen igual conductor, así que deben ser iguales, por el lema 3.11. □

**Definición 3.13.** Sea $R$ un dominio íntegro con cuerpo de fracciones $K$. Un *ideal fraccional* es un $R$-submódulo $I \subseteq K$ para el cual existe $r \in R \setminus \{0\}$ tal que $rI \subseteq R$. Decimos que un ideal fraccional $I$ es *inversible* si existe un ideal fraccional $J \subseteq K$ tal que $IJ = R$, y decimos que $I$ es *propio* si verifica que

$$\{x \in K : xI \subseteq I\} = R.$$

Llamamos a los ideales $I \subseteq R$ *ideales integrales* para diferenciarlos de los ideales fraccionarios.

Si el anillo es un orden $\mathcal{O}$ en un cuerpo de números $K$, entonces un ideal fraccional es lo mismo que un $\mathcal{O}$-submódulo de $K$ finitamente generado. Esto se debe a que si $J$ es un ideal fraccional y $r \in \mathcal{O} \setminus \{0\}$ es tal que $rJ \subseteq \mathcal{O}$, entonces $J$ es isomorfo a $rJ$, que es un grupo finitamente generado y en particular es finitamente generado como $\mathcal{O}$-módulo. Notemos que todo ideal principal $\alpha\mathcal{O}$ con $\alpha \in K$ no nulo es inversible ya que tiene como inverso a $\alpha^{-1}\mathcal{O}$. Otro ejemplo de ideal inversible es el ideal $[3, 1+\sqrt{-5}]$ en $\mathbb{Z}[\sqrt{-5}]$, ya que

$$[3, 1+\sqrt{-5}] \cdot \left[1, \frac{1-\sqrt{-5}}{3}\right] = \mathbb{Z}[\sqrt{-5}].$$

En el caso del anillo de enteros $\mathcal{O}_K$ (o más en general un dominio de Dedekind), todo ideal fraccional no nulo es inversible. Lo mismo no vale para el resto de los órdenes, por ejemplo se puede ver que el ideal $I = [2, 2i]$ de $\mathcal{O} = \mathbb{Z}[2i]$ no es inversible.

Dado un elemento $\tau$ en un cuerpo de números $K$ con polinomio minimal $p \in \mathbb{Q}[x]$, existe un único racional $\lambda > 0$ tal que $\lambda p$ es un polinomio con coeficientes enteros coprimos. Llamamos a $\lambda p$ el *polinomio primitivo* de $\tau$.

**Lema 3.14.** *Sea $K = \mathbb{Q}(\tau)$ un cuerpo cuadrático y sea $ax^2 + bx + c = 0$ el polinomio primitivo de $\tau$. Entonces $[1, \tau]$ es un ideal fraccional propio del orden $\mathcal{O} = [1, a\tau]$.*

*Demostración.* Notemos primero que $\mathcal{O}$ es un orden ya que $(a\tau)^2 = -a(b\tau + c) = -b(a\tau) - bc \in \mathcal{O}$. Para probar que $[1, \tau]$ es un ideal fraccionario de $\mathcal{O}$ y que es propio debemos ver que

$$\{\beta \in K : \beta[1, \tau] \subseteq [1, \tau]\} = [1, a\tau].$$

Sea $\beta = m + n\tau \in K$, donde $m, n \in \mathbb{Q}$. Entonces $\beta[1, \tau] \subseteq [1, \tau]$ si y solo si $\beta, \beta\tau \in [1, \tau]$. Ahora, $\beta \in [1, \tau]$ si y solo si $m, n \in \mathbb{Z}$, mientras que

$$\beta\tau = m\tau + n\tau^2 = m\tau - \frac{n}{a}(b\tau + c) = \frac{-cn}{a} + \tau\left(\frac{-bn}{a} + m\right).$$

Luego $\beta\tau \in [1, \tau]$ si y solo si $a \mid bn$ y $a \mid cn$. Como $\operatorname{mcd}(a, b, c) = 1$, esto es equivalente a que $a \mid n$. Por lo tanto $\beta[1, \tau] \subseteq [1, \tau]$ si y solo si $\beta \in [1, a\tau]$, como queríamos ver. □

**Lema 3.15.** *Sea $\mathcal{O}$ un orden en un cuerpo cuadrático $K$ y sea $I$ un ideal fraccional de $\mathcal{O}$. Entonces $I$ es inversible si y solo si $I$ es propio.*

*Demostración.* Supongamos primero que $I$ es inversible y sea $J$ el ideal fraccional tal que $IJ = \mathcal{O}$. Es claro que $xI \subseteq I$ para todo $x \in \mathcal{O}$. Supongamos que $x \in K$ satisface $xI \subseteq I$. Entonces $x\mathcal{O} = xIJ = IJ = \mathcal{O}$ y en particular $x \in \mathcal{O}$. Por lo tanto $I$ es propio.

Recíprocamente, supongamos que $I$ es propio y sea $\{\alpha, \beta\}$ una base de $I$ como $\mathbb{Z}$-módulo. Entonces $I = \alpha[1, \tau]$, donde $\tau = \beta/\alpha$. Sea $ax^2 + bx + c \in \mathbb{Z}[x]$ el polinomio primitivo de $\tau$. Por el lema 3.14, $[1, \tau]$ es un ideal fraccional propio del orden $[1, a\tau]$. Luego

$$\mathcal{O} = \{x \in K : x[1, \tau] \subseteq [1, \tau]\} = [1, a\tau].$$

Notemos que $\mathcal{O}_K' = \mathcal{O}_K$ y que como $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$ para algún $f > 0$, tenemos que $\mathcal{O}' = \mathcal{O}$. Además $I' = \alpha'[1, \tau']$ es un ideal fraccional propio de $\mathcal{O}' = \mathcal{O}$. Afirmamos que $II' = \frac{N(\alpha)}{a}\mathcal{O}$. En efecto,

$$\begin{aligned}\frac{a}{N(\alpha)}II' &= a[1, \tau, \tau', \tau\tau'] = a[1, \tau, \tau + \tau', \tau\tau'] \\ &= [a, a\tau, -b, c] = [1, a\tau] = \mathcal{O},\end{aligned}$$

ya que $\operatorname{mcd}(a, b, c) = 1$. Por lo tanto $I$ es inversible. □

A diferencia de lo que sucede en el anillo de enteros $\mathcal{O}_K$, en un orden la norma de ideales no es multiplicativa en general. Por ejemplo, si $\mathcal{O} = \mathbb{Z}[2i]$ y tomamos el ideal $I = [2, 2i]$, entonces $N(I) = 2$ pero $N(I^2) = N([4, 4i]) = 8 \neq N(I)^2$. Sin embargo, la multiplicatividad sigue valiendo si uno se restringe a los ideales inversibles:

**Lema 3.16.** *Sea $\mathcal{O}$ un orden en un cuerpo cuadrático. Entonces*

(a) $N(\alpha\mathcal{O}) = |N(\alpha)|$ *para todo $\alpha \in \mathcal{O}$ no nulo.*
(b) *Si $I \subseteq \mathcal{O}$ es un ideal propio entonces $II' = N(I)\mathcal{O}$.*
(c) *Si $I, J \subseteq \mathcal{O}$ son ideales propios entonces $N(IJ) = N(I)N(J)$.*

*Demostración.* (a) Sea $L_\alpha : K \to K$ la transformación $\mathbb{Q}$-lineal $L_\alpha(x) = \alpha x$. Entonces es bien sabido que $|\mathcal{O}/\alpha\mathcal{O}| = |\det(L_\alpha)| = |N(\alpha)|$.

(b) Veamos primero que $N(\alpha I) = |N(\alpha)|N(I)$ para todo $\alpha \in \mathcal{O}$ y todo ideal $I \subseteq \mathcal{O}$. Para ver esto notemos que hay una sucesión exacta corta

$$0 \longrightarrow \alpha\mathcal{O}/\alpha I \longrightarrow \mathcal{O}/\alpha I \longrightarrow \mathcal{O}/\alpha\mathcal{O} \longrightarrow 0,$$

lo que implica que $|\mathcal{O}/\alpha I| = |\mathcal{O}/\alpha\mathcal{O}| \cdot |\alpha\mathcal{O}/\alpha I|$. Por otro lado, la multiplicación por $\alpha$ nos da un isomorfismo $\mathcal{O}/I \xrightarrow{\sim} \alpha\mathcal{O}/\alpha I$, así que $N(\alpha I) = N(\alpha\mathcal{O})N(I) = |N(\alpha)|N(I)$ por el punto anterior.

Ahora, dado un ideal propio $I \subseteq \mathcal{O}$, lo podemos escribir como $I = \alpha[1, \tau]$ con $\alpha, \tau \in K$. Sea $ax^2 + bx + c$ el polinomio primitivo de $\tau$. Por el lema 3.14, $I$ es un ideal fraccional propio del orden $[1, a\tau]$. Luego

$$\mathcal{O} = \{x \in K : xI \subseteq I\} = [1, a\tau]$$

ya que $I$ es propio. Es claro que el ideal $[a, a\tau]$ tiene índice $a$ en $[1, a\tau]$, así que

$$N(I) = N(\frac{\alpha}{a}[a, a\tau]) = \frac{N(\alpha)}{a^2}N([a, a\tau]) = \frac{N(\alpha)}{a}.$$

Entonces, como vimos en el lema anterior, es

$$II' = \frac{N(\alpha)}{a}\mathcal{O} = N(I)\mathcal{O}.$$

(c) Usando el punto anterior, tenemos que

$$N(IJ)\mathcal{O} = IJI'J' = N(I)\mathcal{O}N(J)\mathcal{O} = N(I)N(J)\mathcal{O}.$$

Por lo tanto $N(IJ) = N(I)N(J)$. □

Por el lema podemos extender la definición de norma a ideales fraccionales inversibles, de la siguiente manera: si $I = \lambda J$ con $J \subseteq \mathcal{O}$ un ideal integral, entonces $N(I) := N(\lambda)N(J)$. La buena definición de la norma se deduce de 3.16.(c).

**Proposición 3.17.** *Sean $q_1(x, y) = a_1x^2 + b_1xy + c_1y^2$, $q_2(x, 2) = a_2x^2 + b_2xy + c_2y^2$ dos formas cuadráticas binarias enteras y supongamos que existe $g = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ tal que $q_2 = q_1 \circ g$. Sean*

$$\tau_1 = \frac{-b_1 + \sqrt{d}}{2a_1}, \qquad \tau_2 = \frac{-b_2 + \sqrt{d}}{2a_2}$$

*donde $d$ es el discriminante común de ambas formas. Entonces*

(a) $a_2 = a_1N(-r\tau_1 + p)$.
(b) $\tau_2 = g^{-1}.\tau_1$.

*Demostración.* Como $\tau_2$ es raíz del polinomio $q_2(x,1)$, podemos factorizar a $q_2$ como

$$q_2(x,y) = a_2(x-\tau_2 y)(x-\tau_2' y).$$

Similarmente, $q_1(x,y) = a_1(x-\tau_1 y)(x-\tau_1' y)$, así que

$$\begin{aligned} q_1 \circ g(x,y) &= q_1(px+qy, rx+sy) \\ &= a_1\left[(px+qy)-\tau_1(rx+sy)\right]\left[(px+qy)-\tau_1'(rx+sy)\right] \\ &= a_1\left[x(-r\tau_1+p)-y(s\tau_1-q)\right]\left[x(-r\tau_1'+p)-y(s\tau_1'-q)\right]. \end{aligned}$$

Se sigue que $\tau_3 = (s\tau_1-q)/(-r\tau_1+p) = g^{-1}.\tau_1$ es una raíz de $q_2$, y por tanto es igual a $\tau_2$ o $\tau_2'$. Supongamos que $\tau_3 = \tau_2'$ y sea $\lambda = -r\tau_1'+p$. Mirando el coeficiente de $x^2$ en $q_2 = q_1 \circ g$, tenemos

$$a_2 = a_1 N(\lambda) = a_1 N(\lambda'). \tag{3.3}$$

Esto demuestra $(a)$. Por otro lado, como $s\tau_1' - q = \lambda\tau_2$,

$$\begin{pmatrix} s & -q \\ -r & p \end{pmatrix}\begin{pmatrix} \tau_1 & \tau_1' \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \lambda'\tau_2' & \lambda\tau_2 \\ \lambda' & \lambda \end{pmatrix} = \begin{pmatrix} \tau_2' & \tau_2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \lambda' & 0 \\ 0 & \lambda \end{pmatrix}.$$

Tomando determinante obtenemos que

$$\tau_1 - \tau_1' = (\tau_2' - \tau_2)N(\lambda),$$

luego

$$N(\lambda) = \frac{\tau_2'-\tau_2}{\tau_1-\tau_1'} = -\frac{a_2}{a_1}.$$

Esto se contradice con (3.3). Por lo tanto $g^{-1}.\tau_1 = \tau_2$. □

Ahora estamos en condiciones de demostrar el principal resultado de esta sección, que es la proposición 3.18. El conjunto $\mathcal{I}(\mathcal{O})$ de ideales fraccionarios no nulos forma un grupo con el producto de ideales. Además hay un morfismo de grupos $K^\times \to \mathcal{I}(\mathcal{O})$ dado por $x \mapsto x\mathcal{O}$. Si denotamos por $\mathcal{P}$ a la imagen de este morfismo, el grupo $\mathrm{Pic}(\mathcal{O}) = \mathcal{I}(\mathcal{O})/\mathcal{P}$ se denomina el *grupo de clase* o *grupo de Picard* de $\mathcal{O}$. El grupo que nos interesa en esta sección es el cociente $\mathrm{Pic}^+(\mathcal{O}) = \mathcal{I}(\mathcal{O})/\mathcal{P}^+$, donde $\mathcal{P}^+$ es el grupo de ideales principales no nulos generados por un elemento de norma positiva. El grupo $\mathrm{Pic}^+(\mathcal{O})$ se llama el *grupo de clases angostas* (*narrow class group* en inglés).

Supongamos que $K$ es un cuerpo cuadrático y sea $L \subseteq K$ un retículo. Una base $\{\omega_1, \omega_2\}$ de $L$ como $\mathbb{Z}$-módulo se dice *bien ordenada* si $\omega_1'\omega_2 - \omega_1\omega_2' > 0$. Recordemos que $C(d)$ es el cociente de $R_{\mathrm{disc}}(d)$ por la acción de $\mathrm{PGL}_2(\mathbb{Z})$.

**Proposición 3.18.** *Sea $\mathcal{O}$ un orden de discriminante $d > 0$ en un cuerpo cuadrático $K$. Entonces hay una biyección $\psi : C(d) \to \mathrm{Pic}^+(\mathcal{O})$, definida de la siguiente manera: dada una forma cuadrática $q(x,y) = ax^2+bxy+cy^2 \in R_{\mathrm{disc}}(d)$, la imagen de $[q]$ es la clase del ideal*

$$I_q = \eta\left[a, \frac{-b+\sqrt{d}}{2}\right] = a\eta[1,\tau], \tag{3.4}$$

*donde $\eta = \eta(q)$ es cualquier elemento de $K^\times$ tal que $N(\eta)$ tiene igual signo que $a$. El inverso de $\psi$ está caracterizado por*

$$\psi^{-1}[J] = \frac{N(x\omega_1 - y\omega_2)}{N(J)}, \tag{3.5}$$

*donde $\omega_1, \omega_2$ es una base bien ordenada de $J$.*

**Observación 3.19.** De la proposición anterior se deduce que $C(d)$ es un conjunto finito, ya que $\mathrm{Pic}^+(\mathcal{O}_d)$ es un cociente del grupo de clase de $\mathcal{O}_d$ que es un grupo finito. Alternativamente, se puede probar directamente que $C(d)$ es un conjunto finito, ver [Per12, §1.4].

*Demostración.* Por el lema 3.14, $I_q$ es un ideal fraccional propio del orden $[1, a\tau]$. Ahora, si $f$ es el conductor de $\mathcal{O}$, entonces $d = f^2 d_K$ y

$$\begin{aligned} a\tau &= \frac{-b+\sqrt{d}}{2} = \frac{-b+f\sqrt{d_K}}{2} \\ &= \frac{-b+fd_K}{2} + f\left(\frac{d_K+\sqrt{d_K}}{2}\right) = \frac{-b+fd_K}{2} + f\alpha, \end{aligned}$$

donde $\alpha = \frac{d_K+\sqrt{d_K}}{2}$. Como $d = b^2 - 4ac$, tenemos que $b \equiv b^2 \equiv d \equiv fd_K \pmod 2$, de modo que $\frac{-b+fd_K}{2} \in \mathbb{Z}$. Se sigue que $[1, a\tau] = [1, f\alpha] = \mathcal{O}$, y luego $I_q$ es un ideal fraccional propio de $\mathcal{O}$. Ahora veamos que esta asignación induce un mapa bien definido $C(d) \to \mathrm{Pic}^+(\mathcal{O})$. Sean $q_1, q_2 \in R_{\mathrm{disc}}(d)$ tales que $q_i = a_i x^2 + b_i xy + c_i y^2$ y sea

$$\tau_i = \frac{-b_i + a_i\sqrt{d}}{2a_i},$$

para $i = 1, 2$. Afirmamos que vale lo siguiente:

$$q_1 \text{ es propiamente equivalente a } q_2 \iff [I_{q_1}] = [I_{q_2}] \text{ en } \mathrm{Pic}^+(\mathcal{O}_d).$$

En efecto, si $q_1$ y $q_2$ son propiamente equivalentes, existe $g \in \mathrm{SL}_2(\mathbb{Z})$ tal que $q_2 = q_1 \circ g$. En ese caso tenemos que $\tau_2 = g^{-1} \cdot \tau_1$ por la proposición 3.17.(*b*). Supongamos que $g^{-1} = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right)$. Entonces

$$[1, \tau_2] = \left[1, \frac{p\tau_1 + q}{r\tau_1 + s}\right] = (r\tau_1 + s)^{-1}[p\tau_1 + q, r\tau_1 + s] = (r\tau_1 + s)^{-1}[1, \tau_1],$$

así que $[1, \tau_1] = (r\tau_1 + s)[1, \tau_2]$. Se sigue que

$$I_{q_1} = \eta_1 a_1 [1, \tau_1] = \frac{\eta_1 a_1}{\eta_2 a_2}(r\tau_1 + s) I_{q_2}.$$

Como $g = \left(\begin{smallmatrix} s & -q \\ -r & p \end{smallmatrix}\right)$, por el punto 3.17.(*a*) tenemos que $a_2 = a_1 N(r\tau_1 + s)$. Entonces

$$N\left(\frac{\eta_1 a_1}{\eta_2 a_2}\right) N(r\tau_1 + s) = \frac{N(\eta_1)a_1^2}{N(\eta_2)a_2^2} \cdot \frac{a_2}{a_1} = \frac{N(\eta_1)a_1}{N(\eta_2)a_2} > 0.$$

Luego $[I_{q_1}] = [I_{q_2}]$ en $\mathrm{Pic}^+(\mathcal{O}_d)$.

Recíprocamente, supongamos que $I_{q_1}$ e $I_{q_2}$ están en la misma clase de $\mathrm{Pic}^+(\mathcal{O}_d)$. Entonces existe $\mu \in K$ con $N(\mu) > 0$ tal que $\eta_1 a_1 [1, \tau_1] = \mu \eta_2 a_2 [1, \tau_2]$. Sea

$$\lambda = \frac{\eta_2 a_2}{\eta_1 a_1}\mu.$$

Luego $[1, \tau_1] = \lambda[1, \tau_2]$ así que existe $g = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z})$ tal que

$$\begin{cases} \lambda\tau_2 = p\tau_1 + q, & (3.6) \\ \lambda = r\tau_1 + s. & (3.7) \end{cases}$$

Si dividimos la primera ecuación por la segunda obtenemos $\tau_2 = \frac{p\tau_1+q}{r\tau_1+s}$. Además tenemos que

$$\begin{pmatrix} p & q \\ r & s \end{pmatrix}\begin{pmatrix} \tau_1 & \tau_1' \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \lambda\tau_2 & \lambda'\tau_2' \\ \lambda & \lambda' \end{pmatrix} = \begin{pmatrix} \tau_2 & \tau_2' \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \lambda & 0 \\ 0 & \lambda' \end{pmatrix}.$$

y tomando determinantes resulta que

$$\det(g)(\tau_1 - \tau_1') = (\tau_2 - \tau_2')N(\lambda),$$

Como $\tau_i - \tau_i' = \sqrt{d}/a_i$, es

$$\det(g) = \frac{a_1}{a_2} \cdot \frac{N(\eta_2)a_2^2}{N(\eta_1)a_1^2} N(\mu) = \frac{N(\eta_2)a_2}{N(\eta_1)a_1} N(\mu) > 0.$$

Por lo tanto $g \in \mathrm{SL}_2(\mathbb{Z})$. Por el punto 3.17.(*b*), las formas $q_2$ y $q_1 \circ g^{-1}$ tienen las mismas raíces, puesto que $\tau_2 = g.\tau_1$. Se sigue que $q_2 = \pm q_1 \circ g^{-1}$. Ahora, el coeficiente de $x^2$ en $q_1 \circ g^{-1}$ es igual a $a_1 N(r\tau_1 + s) = a_1 N(\lambda)$ por el punto 3.17.(*a*). Por definición de $\lambda$,

$$N(\lambda) = \frac{N(\eta_2)a_2^2}{N(\eta_1)a_1^2} N(\mu),$$

así que $N(\lambda)$ tiene igual signo que $a_2/a_1$, y luego $a_1 N(\lambda)$ tiene igual signo que $a_2$. Por lo tanto $q_2 = q_1 \circ g^{-1}$.

De esta equivalencia se deduce que el mapa $C(d) \to \mathrm{Pic}^+(\mathcal{O})$ está bien definido y es inyectivo. Queda probar que es sobreyectivo. Dado un ideal fraccional inversible $I$ de $\mathcal{O}$, elegimos $\omega_1 \in I \cap \mathbb{N}$ no nulo y luego elegimos $\omega_2 \in I$ tal que $I = [\omega_1, \omega_2]$. Claramente $N(\omega_1) > 0$, así que $I$ representa la misma clase de $\mathrm{Pic}^+(\mathcal{O})$ que $J = [1, \tau]$, donde $\tau = \omega_2/\omega_1$. Reemplazando a $\tau$ por $-\tau$ podemos suponer que $(\tau - \tau') > 0$. Sea $ax^2 + bx + c \in \mathbb{Z}[x]$ el polinomio primitivo de $\tau$, y sea $q(x, y) = ax^2 + bxy + cy^2$. Por el lema 3.14 $J$ es un ideal fraccional propio de $[1, a\tau]$, así que $[1, a\tau] = \mathcal{O}$. En particular

$$\mathrm{disc}[1, a\tau] = \begin{vmatrix} 1 & 1 \\ a\tau & a\tau' \end{vmatrix}^2 = d,$$

$$a^2(\tau - \tau')^2 = d.$$

Por otro lado, $\tau = \frac{-b+\sqrt{\Delta}}{2a}$, donde $\Delta = b^2 - 4ac$. Entonces $\tau - \tau' = \sqrt{\Delta}/a$, y $a^2(\tau - \tau')^2 = \Delta$. Por lo tanto $\Delta = d$, es decir, $q \in R_{\mathrm{disc}}(d)$. Claramente el ideal asociado a $q$ está en la misma clase de $\mathrm{Pic}^+(\mathcal{O})$ que $J$.

Finalmente, veamos que el inverso de $\psi$ está dado por la fórmula (3.5). Definimos una aplicación $\phi : \mathcal{I}(\mathcal{O}) \to \mathbb{Q}[x, y]_2/\mathrm{SL}_2(\mathbb{Z})$, donde $\mathbb{Q}[x, y]_2$ es el conjunto de polinomios homogéneos de grado 2 sobre $\mathbb{Q}$:

$$\phi(J) = \frac{N(x\omega_1 - y\omega_2)}{N(J)},$$

donde $\omega_1, \omega_2$ es una base bien ordenada de $J$. Si elegimos otra base con la misma propiedad, esto genera una forma cuadrática que es propiamente equivalente a la original, luego $\phi$ está bien definido.

En primer lugar afirmamos que $\phi(I_q) = q$ para todo $q \in R_{\mathrm{disc}}(d)$. Por el lema 3.9, tenemos que

$$N(I_q) = \left| \frac{\Delta(\eta a, \eta a\tau)}{f^2 d_K} \right|^{1/2} = |aN(\eta)| = aN(\eta).$$

Notemos además que la base $\omega_1 = \eta a$, $\omega_2 = \omega a\tau$ es bien ordenada:

$$(\eta a)'\eta a\tau - \eta a(\eta a\tau)' = N(\eta)a^2(\tau - \tau') = N(\eta)a\sqrt{d} > 0.$$

Luego

$$\begin{aligned} \phi(I_q)(x, y) &= \frac{1}{N(I_q)} N(\eta a x - \eta a\tau y) \\ &= \frac{N(\eta a)}{N(I_q)} N(x - \tau y) \\ &= \frac{a^2 N(\eta)}{aN(\eta)} N(x - \tau y) = aN(x - \tau y) = q(x, y). \end{aligned}$$

Es fácil ver que si $J_1, J_2 \in \mathcal{I}(\mathcal{O})$ son congruentes módulo $\mathcal{P}^+$ entonces $\phi(J_1) = \phi(J_2)$, luego $\phi(J)$ induce una aplicación $\phi_1 : \mathrm{Pic}^+(\mathcal{O}) \to \mathbb{Q}[x, y]_2/\mathrm{SL}_2(\mathbb{Z})$. Como $\psi$ es sobreyectivo, la imagen de $\phi$ siempre es un polinomio primitivo. Como $\phi_1 \circ \psi$ es la identidad, concluimos que $\phi_1 = \psi^{-1}$. □

Sea $q$ una forma cuadrática binaria. Dado un entero $n$, llamamos *representación* de $n$ por $q$ a un par $(x_0, y_0) \in \mathbb{Z}^2$ tal que $q(x_0, y_0) = n$. La correspondencia entre formas y clases de ideales nos permite relacionar las representaciones de un entero $n > 0$ con ideales de norma $n$. Sin embargo,

en el caso de discriminante $d > 0$, que es el que nos interesa, el conjunto de representaciones de un entero $n$ por $q$ es infinito siempre que no es vacío. Esto se debe a que el subgrupo $\mathcal{O}_{d,1} \subseteq \mathcal{O}_d^\times$ de unidades de norma 1 es un grupo infinito. De hecho $\mathcal{O}_d^\times$ es un grupo finitamente generado de rango 1 por el teorema de las unidades de Dirichlet y $\mathcal{O}_{d,1}$ es un subgrupo de índice $\leq 2$. Si definimos $\varepsilon_0$ y $\varepsilon$ como los menores elementos en $\mathcal{O}_d^\times \cap (1, +\infty)$ y $\mathcal{O}_{d,1} \cap (1, +\infty)$ respectivamente, es fácil ver que $\mathcal{O}_d^\times = \{\pm 1\}\varepsilon_0^{\mathbb{Z}}$ y $\mathcal{O}_{d,1} = \{\pm 1\}\varepsilon^{\mathbb{Z}}$. El elemento $\varepsilon_0$ se suele llamar la *unidad fundamental* de $K = \mathbb{Q}(\sqrt{d})$. En particular el regulador de $\mathcal{O}_d$ es igual a $\log \varepsilon_0$. Para ver cómo esto afecta la cantidad de representaciones, notemos que por la proposición 3.18 todo $q \in R_{\text{disc}}(d)$ es propiamente equivalente a una forma $q_J(x, y) = N(J)^{-1}N(x\omega_1 - y\omega_2)$ donde $J \subseteq \mathcal{O}_d$ es un ideal inversible y $\{\omega_1, \omega_2\}$ es una base bien ordenada de $J$. Supongamos que $(x_0, y_0)$ es una representación de $n$ por $q = q_J$ y sea $\beta = x_0\omega_1 - y_0\omega_2 \in J$. Entonces $N(\beta) = nN(J)$. Luego $N(\pm\varepsilon^m\beta) = nN(J)$ para todo $m \in \mathbb{Z}$ y esto induce infinitas soluciones distintas de $q(x, y) = n$.

Sin embargo, observemos que para todo $\gamma \in \mathcal{O}_d^\times \cap \mathbb{R}_{>0}$ existe un único entero $m$ tal que

$$1 \leq \frac{(\gamma\varepsilon^m)'}{\gamma\varepsilon^m} < \varepsilon^2.$$

Decimos que una representación $(x, y)$ de $n$ por $q$ es *primaria* si

$$x - \tau y > 0, \qquad \text{y} \qquad 1 \leq \frac{x - \tau' y}{x - \tau y} < \varepsilon^2, \tag{3.8}$$

donde $\tau$ es la raíz principal de $q$. Notemos que si $q = q_J$ entonces $\tau = \omega_2/\omega_1$. En efecto, como

$$q(x, 1) = \frac{1}{N(J)}(x\omega_1 - \omega_2)(x\omega_1' - \omega_2')$$

vemos que $\tau = \frac{\omega_2}{\omega_1}$ o $\tau = (\frac{\omega_2}{\omega_1})'$. Por otro lado, el coeficiente $a$ de $x^2$ en $q$ es igual a $N(J)^{-1}N(\omega_1)$, así que $a$ y $N(\omega_1)$ tienen igual signo. Como $\omega_1, \omega_2$ es una base bien ordenada, tenemos que

$$\frac{a}{N(\omega_1)}(\omega_1'\omega_2 - \omega_1\omega_2') = a\left(\frac{\omega_2}{\omega_1} - \frac{\omega_2'}{\omega_1'}\right) > 0.$$

Usando esto es fácil ver que $\tau \neq (\frac{\omega_2}{\omega_1})'$, así que $\tau = \frac{\omega_2}{\omega_1}$. Luego las condiciones de (3.8) se traducen en términos de $\gamma = \beta/\omega_1 = x - \tau y$ como $\gamma > 0$ y

$$1 \leq \gamma'/\gamma < \varepsilon^2.$$

Ahora fijemos un conjunto de representantes $q_1, \ldots, q_h \in R_{\text{disc}}(d)$ de $C(d)$. Sea $R_d(n)$ la cantidad de representaciones primitivas de $n$ por alguna forma $q_i$. Afirmamos que $R_d(n)$ es igual al número de ideales inversibles $I \subseteq \mathcal{O}_d$ de norma $n$, donde $\mathcal{O}_d$ es el orden cuadrático de discriminante $d$. En efecto, supongamos que $(x, y)$ es una representación primitiva de $n$ por una forma cuadrática $q = q_i$. Podemos suponer que $q = q_J$ para algún ideal inversible $J \subseteq \mathcal{O}_d$, es decir,

$$q(x, y) = \frac{N(x\omega_1 - y\omega_2)}{N(J)}.$$

Entonces existe $\beta \in J$ tal que $N(\beta) = nN(J)$. Como $\beta \in J$, el ideal fraccionario $I = \beta J^{-1}$ es integral y tiene norma $N(I) = n$. Notemos que necesariamente $N(\beta) > 0$ así que la clase de $J$ en $\text{Pic}^+(\mathcal{O}_d)$ está determinada por $I$ y en particular $I$ no se puede obtener por este proceso a partir de $q_j$ para $j \neq i$. Si $(x', y')$ es otra representación primitiva de $n$ por $q_i$ que genera $I$ y $\beta' \in J$ es el elemento correspondiente, entonces $(\beta) = (\beta')$. Luego $\beta$ y $\beta'$ difieren en una unidad de norma positiva ya que $N(\beta), N(\beta') > 0$. El hecho de que $(x, y)$, $(x', y')$ son primitivos implica que $x = x'$ e $y = y'$.

**Teorema 3.20.** *Sea $d > 0$, $d \equiv 0, 1 \pmod 4$ y sea $n \in \mathbb{N}$ tal que $\text{mcd}(n, d) = 1$. Entonces*

$$R_d(n) = \sum_{m|n} \left(\frac{d}{m}\right), \tag{3.9}$$

*donde $\left(\frac{d}{m}\right)$ es el símbolo de Kronecker. En particular $R_d(n) \leq \tau(n)$ para todo $n \in \mathbb{N}$, donde $\tau(n)$ es la cantidad de divisores de $n$.*

*Demostración.* Ver [Lan86, Teo. 204]. □

## 3.3. El espacio de retículos

Sea $X = \mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R}) \cong \mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$ el espacio considerado en el capítulo 2. Veamos que podemos identificar los elementos con retículos en $\mathbb{R}^2$. Sea $\mathscr{L}_2^{(1)}(\mathbb{R})$ el conjunto de retículos $L$ de $\mathbb{R}^2$ de covolumen 1, es decir tales que $|\det(b_1\ \ b_2)| = 1$ para cualquier base $b_1, b_2$ de $L$ como $\mathbb{Z}$-módulo. En general, si $L \subseteq \mathbb{R}^2$ es un retículo con base $b_1, b_2$, definimos su *covolumen* como $\mathrm{vol}(L) = |\det(b_1\ \ b_2)|$. Notar que $\mathscr{L}_2^{(1)}(\mathbb{R})$ está en biyección con el conjunto $[\mathscr{L}_2(\mathbb{R})]$ de retículos de $\mathbb{R}^2$ módulo homotecia vía $[L] \mapsto \mathrm{vol}(L)^{-1/2}L$. El grupo $\mathrm{PSL}_2(\mathbb{R})$ actúa a derecha por multiplicación en $\mathscr{L}_2^{(1)}(\mathbb{R})$ y la acción es transitiva. El estabilizador de $\mathbb{Z}^2$ es igual a $\mathrm{PSL}_2(\mathbb{Z})$, entonces hay una biyección entre $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$ y $\mathscr{L}_2^{(1)}(\mathbb{R})$ dada por $\overline{g} \mapsto \mathbb{Z}^2.g$. Similarmente, el grupo $\mathrm{PGL}_2(\mathbb{R})$ actúa a derecha en $[\mathscr{L}_2(\mathbb{R})]$ y esto induce una biyección $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R}) \xrightarrow{\sim} [\mathscr{L}_2(\mathbb{R})]$, $\overline{g} \mapsto [\mathbb{Z}^2.g]$. Claramente el siguiente diagrama conmuta:

$$\begin{array}{ccc} X & \longrightarrow & \mathscr{L}_2^{(1)}(\mathbb{R}). \\ & \searrow \quad \nearrow & \\ & [\mathscr{L}_2(\mathbb{R})] & \end{array}$$

Dado un retículo $L = \mathbb{Z}^2.g \subseteq \mathbb{R}^2$, definimos su *altura* como

$$\mathrm{ht}(L) = \left(\frac{\mathrm{mín}_{x\in L\setminus\{0\}} \|x\|}{\mathrm{vol}(L)^{1/2}}\right)^{-1} = \left(\frac{\mathrm{mín}_{x\in\mathbb{Z}^2\setminus\{0\}} \|xg\|}{|\det(g)|^{1/2}}\right)^{-1}, \tag{3.10}$$

donde $\|\cdot\|$ denota la norma usual en $\mathbb{R}^2$. Notemos que $\mathrm{ht}(L)$ solo depende de la clase de homotecia de $L$. Recordemos del capítulo 2 que el conjunto

$$S = \{(z, v) \in T^1\mathbb{H} : |\mathrm{Re}(z)| \leq \tfrac{1}{2}, |z| \geq 1\}$$

es un dominio fundamental para la acción de $\mathrm{PSL}_2(\mathbb{Z})$ en $T^1\mathbb{H} \cong \mathrm{PSL}_2(\mathbb{R})$. Sorprendentemente, la altura tiene una interpretación sencilla dentro de $S$:

**Lema 3.21.** *Si $(z, v) \in S$ representa un retículo $L \in \mathscr{L}_2^{(1)}(\mathbb{R})$, entonces $\mathrm{ht}(L)^2 = \mathrm{Im}(z)$.*

*Demostración.* Sea $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$ tal que $(z, v) = g.(i, i)$. Entonces

$$z = \frac{ai + b}{ci + d} = \frac{ac + bd}{c^2 + d^2} + \frac{i}{c^2 + d^2}.$$

En particular $\mathrm{Im}(z) = (c^2 + d^2)^{-1}$ y las condiciones $|\mathrm{Re}(z)| \leq \frac{1}{2}$, $|z| \geq 1$ son equivalentes a $|ac + bd| \leq \frac{1}{2}(c^2 + d^2)$, y $a^2 + b^2 \geq c^2 + d^2$, respectivamente. El retículo correspondiente a $(z, v)$ es $L = \mathbb{Z}^2.g = \mathbb{Z}(a, b) \oplus \mathbb{Z}(c, d)$. Si definimos $v_1 = (a, b)$, $v_2 = (c, d)$, tenemos que $\|v_1\| \geq \|v_2\|$ y $|\langle v_1, v_2\rangle| \leq \frac{1}{2}\|v_2\|^2$. Sean $A = \|v_1\|^2$, $B = \|v_2\|^2$. Supongamos que $x = mv_1 + nv_2 \in L \setminus \{0\}$ minimiza la norma. Tenemos

$$\begin{aligned} \|x\|^2 &= Am^2 + 2mn\langle v_1, v_2\rangle + Bn^2 \\ &\geq Am^2 - Bmn + Bn^2 = B\left(n - \frac{m}{2}\right)^2 + \left(A - \frac{B}{4}\right)m^2. \end{aligned}$$

Notemos que $|m| \leq 1$ ya que si $|m| \geq 2$ entonces $\|x\|^2 \geq 4(A - \frac{B}{4}) \geq 3B > \|v_2\|^2$, absurdo. Si $m = 0$, entonces $\|x\|^2 = Bn^2 \geq B$, con igualdad si $(m, n) = (0, 1)$. Si $m = \pm 1$, entonces

$$\|x\|^2 = B\left(n \pm \frac{1}{2}\right)^2 + A - \frac{B}{4}.$$

Como $n$ es entero, $(n \pm \frac{1}{2})^2 \geq \frac{1}{4}$ y luego $\|x\|^2 \geq A \geq B$. Por lo tanto la norma mínima es $\|v_2\| = (c^2 + d^2)^{1/2}$ y $\mathrm{ht}(L) = (c^2 + d^2)^{-1/2}$. □

De este lema se deduce que el conjunto de puntos $x \in X$ con $\mathrm{ht}(x) \leq H$ es compacto para todo $H > 0$, ya que el conjunto de puntos $(z, v) \in S$ con $\mathrm{Im}(z) \leq H^2$ es compacto. Dado un número real $H > 0$, definimos $X_{\geq H} = \{x \in X : \mathrm{ht}(x) \geq H\}$. Los subconjuntos $X_{>H}, X_{\leq H}, X_{<H}$ se definen de manera análoga. Recordemos que habíamos definido en el capítulo anterior el subgrupo

$$A = \left\{ \begin{pmatrix} a & 0 \\ 0 & \pm a^{-1} \end{pmatrix} : a \in \mathbb{R}^\times \right\} \subseteq \mathrm{GL}_2(\mathbb{R}),$$

y que dada $q \in R_{\mathrm{disc}}(d)$, el punto $x_q \in X$ es la clase de algún elemento $g \in \mathrm{GL}_2(\mathbb{R})$ tal que

$$g \cdot q_0(x, y) = \frac{1}{\det(g)} q_0((x, y)g) = d^{-1/2} q(x, y),$$

donde $q_0(x, y) = xy$.

**Proposición 3.22.** *Sea $\mathcal{O}_d$ el orden de discriminante $d > 0$. Supongamos que un ideal propio $J \subseteq \mathcal{O}_d$ se corresponde con la clase de la forma $q = ax^2+bxy+cy^2$ en $C(d)$. Entonces $x_q A \cap X_{\geq H}$ es no vacío si y solo si $J^{-1}$ es equivalente en $\mathrm{Pic}(\mathcal{O}_d)$ a un ideal integral $I \subseteq \mathcal{O}_d$ con $N(I) \leq \frac{1}{2} H^{-2} d^{1/2}$. Más aún, existe un conjunto de ideales inversibles $I \subseteq \mathcal{O}_d$ con $N(I) \leq \frac{1}{2} H^{-2} \sqrt{d}$ que está en correspondencia dos a uno con el conjunto de componentes conexas de $\mathcal{G}_d \cap X_{\geq H}$.*

*Demostración.* Como $\mathbb{R}^\times I$ actúa trivialmente en el espacio de formas cuadráticas podemos suponer $\det(g) = \pm 1$. Entonces $x_q$ se corresponde con el retículo $L = \mathbb{Z}^2 g$ de covolumen 1. Ahora, tenemos que $gA \cap X_{\geq H} \neq \varnothing$ si y solo si existe $\alpha \in A$ tal que $h = x_q \alpha$ tiene altura $\geq H$. Luego la condición $gA \cap X_{\geq H} \neq \varnothing$ es equivalente a

$$\sup_{\alpha \in A} \mathrm{ht}(x_q \alpha) \geq H. \tag{3.11}$$

La altura de $x_q \alpha$ se calcula como

$$\mathrm{ht}(x_q \alpha) = \left( \inf_{x \in \mathbb{Z}^2 \setminus \{0\}} \|xg\alpha\| \right)^{-1} = \left( \inf_{(u,v) \in L \setminus \{0\}} \|(u, v)\alpha\| \right)^{-1},$$

así que (3.11) es equivalente a

$$\inf_{\alpha \in A} \left( \inf_{(u,v) \in L \setminus \{0\}} \|(u, v)\alpha\| \right) \leq H^{-1}. \tag{3.12}$$

Ahora, todo elemento $\alpha \in A$ es de la forma $\left(\begin{smallmatrix} \lambda & 0 \\ 0 & \pm 1/\lambda \end{smallmatrix}\right)$ para algún $\lambda \in \mathbb{R}^\times$ y notemos que

$$\inf_{\lambda \neq 0} \left\| (u, v) \begin{pmatrix} \lambda & 0 \\ 0 & \pm 1/\lambda \end{pmatrix} \right\| = \inf_{\lambda \neq 0} \sqrt{\lambda^2 u^2 + v^2 \lambda^{-2}} = \sqrt{2|uv|}$$

por la desigualdad AM-GM. Entonces (3.12) es equivalente a

$$\inf_{(u,v) \in L \setminus \{0\}} |uv| \leq \tfrac{1}{2} H^{-2}. \tag{3.13}$$

Por otro lado, todo $(u, v) \in L$ se escribe como $(u, v) = (m, n)g$ con $(m, n) \in \mathbb{Z}^2$ y luego

$$|uv| = |q_0((m, n)g)| = \left| \frac{q(m, n)}{\sqrt{d}} \right| = d^{-1/2} \frac{|N(\omega_1 m - \omega_2 n)|}{N(J)}$$

donde $\omega_1, \omega_2$ es una base bien ordenada de $J$. Esta última igualdad se deduce de (3.5). Por lo tanto tenemos que $x_q A \cap X_{\geq H} \neq \varnothing$ si y solo si existe $\beta \in J$ no nulo tal que

$$|N(\beta)| \leq \frac{1}{2} H^{-2} d^{1/2} N(J).$$

Esto es equivalente a que $N(\beta J^{-1}) \leq \frac{1}{2} H^{-2} d^{1/2}$. Notemos además que $I = \beta J^{-1}$ es un ideal integral, ya $\beta J^{-1} \subseteq J J^{-1} = \mathcal{O}$.

Queda probar la última afirmación. Como $\mathscr{G}_d$ es la unión disjunta de cerrados $x_qA$, $[q] \in C(d)$, basta considerar las componentes conexas de $x_qA \cap X_{\geq H}$ para cada $[q]$. Fijemos $[q] \in C(d)$ y sea $J \subseteq \mathcal{O}_d$ su ideal asociado. Como vimos antes, todo elemento $x_q\alpha \in x_qA \cap X_{\geq H}$ tiene asociado al menos un ideal $I$ equivalente a $J^{-1}$ en $\mathrm{Pic}(\mathcal{O}_d)$. Este ideal proviene de elegir un punto $(u,v) \in \mathbb{Z}^2g$ no nulo con $\|(u,v)\alpha\| \leq H^{-1}$, que a su vez determina un elemento $\beta \in J$ con $|N(\beta)| \leq \frac{1}{2}H^{-2}\sqrt{d}$. Consideremos el conjunto $\mathscr{A}$ de ideales $I$ generados de esta manera por puntos $(u,v) \in \mathbb{Z}^2g$ primitivos, es decir, que no son múltiplos enteros de ningún otro punto en $\mathbb{Z}^2g$. A cada elemento $x \in x_qA \cap X_{\geq H}$ le podemos asociar un ideal $I \in \mathscr{A}$ ya que si $(u,v) \in \mathbb{Z}^2g$ cumple $\|(u,v)\alpha\| \leq H^{-1}$ y es múltiplo de otro punto en el retículo, podemos reemplazar $(u,v)$ por el punto más chico. Para cada $I \in \mathscr{A}$ sea $C_I \subseteq x_qA \cap X_{\geq H}$ el conjunto de puntos que tienen asociado a $I$. Afirmamos que estos conjuntos son disjuntos dos a dos. En efecto, supongamos que $x \in C_{I_1} \cap C_{I_2}$ para dos ideales $I_1 \neq I_2$ y sean $(u_1,v_1),(u_2,v_2) \in \mathbb{Z}^2g$ los puntos asociados. Notemos que $(u_1,v_1),(u_2,v_2)$ no son iguales ni opuestos entre sí ya que en tal caso producirían el mismo ideal. Entonces la condición de primitividad implica que $(u_1,v_1)$ y $(u_2,v_2)$ son linealmente independientes. Pero entonces el retículo generado por $(u_1,v_1)\alpha$ y $(u_2,v_2)\alpha$ tiene covolumen $\leq H^{-2} < 1$, lo cual es absurdo ya que $\mathbb{Z}^2g\alpha$ tiene covolumen 1. Además cada $C_I$ es cerrado ya que es la preimagen de $[0,H^{-1}]$ bajo el mapa $x_q\alpha \mapsto \|(u,v)\alpha\|$, que es continuo. Como vimos en la sección anterior, el conjunto de ideales inversibles de norma menor que $T$ es finito para todo $T > 0$ y en particular $\mathscr{A}$ es finito. Luego $(x_qA \cap X_{\geq H}) \setminus C_I$ es cerrado. Como $C_I$ es abierto y cerrado en $x_qA \cap X_{\geq H}$, es unión de componentes conexas de $x_qA \cap X_{\geq H}$. Para concluir la demostración basta ver que $C_I$ tiene dos componentes conexas para todo $I \in \mathscr{A}$.

El grupo $A$ es isomorfo a $\mathbb{R}_{>0} \times \{\pm 1\}$ y todo $\alpha \in A$ se escribe como $\left(\begin{smallmatrix} \lambda & 0 \\ 0 & s/\lambda \end{smallmatrix}\right)$ con $\lambda > 0$ y $s \in \{\pm 1\}$. Las componentes conexas de $A$ están definidas por las ecuaciones $s = 1$, $s = -1$ respectivamente. En estas coordenadas la función $\|(u,v)\alpha\|^2 = u^2\lambda^2 + v^2\lambda^{-2}$ es convexa en cada una de las componentes conexas de $A$, y luego el conjunto de puntos $\alpha$ en cada componente conexa con $\|(u,v)\alpha\|^2 \leq H^{-2}$ es un intervalo cerrado. Finalmente, $C_I$ es la imagen de estos dos intervalos dentro de $x_qA$. □

**Observación 3.23.** Aplicando el resultado anterior a $H = d^{1/4}$ vemos que $\mathscr{G}_d \cap X_{d^{1/4}} = \varnothing$, donde $\mathscr{G}_d$ es la unión de las órbitas $x_qA$, con $q \in R_{\mathrm{disc}}(d)$.

## 3.4. Números $p$-ádicos

Fijemos un número primo $p$. Dado un entero $n \neq 0$, sea $v_p(n)$ el exponente de $p$ en la factorización en primos de $n$. Es decir, $v_p(n)$ es el mayor entero $k \geq 0$ tal que $p^k \mid n$. Es claro que $v_p(mn) = v_p(m)+v_p(n)$ para todo par de enteros $m, n \neq 0$. En particular podemos extender la función $v_p$ a $\mathbb{Q}^\times$ definiendo $v_p(\frac{m}{n}) = v_p(m)-v_p(n)$ para $m, n \in \mathbb{Z}$ no nulos. Además $v_p$ tiene la siguiente propiedad: $v_p(x+y) \geq \mín(v_p(x), v_p(y))$ para todo par de racionales $x, y \in \mathbb{Q}^\times$ con $x+y \neq 0$. Esto es claro si $x, y$ son enteros, ya que si $k = \mín(v_p(x), v_p(y))$ entonces $p^k$ divide a $x, y$ en cuyo caso $p^k \mid x+y$. En el caso general podemos multiplicar a $x$ e $y$ por un entero $m$ para hacerlos ambos enteros. Entonces $v_p(x+y)+v_p(m) = v_p(xm+ym) \geq \mín(v_p(xm), v_p(ym)) = \mín(v_p(x), v_p(y))+v_p(m)$ y la desigualdad se sigue. En otras palabras, $v_p$ es una *valuación discreta*. Podemos definir un valor absoluto $|\cdot|_p : \mathbb{Q} \to \mathbb{R}_{\geq 0}$ como $|x|_p = p^{-v_p(x)}$ para $x \neq 0$ y $|0|_p = 0$. Es fácil verificar que $|\cdot|_p$ cumple las siguientes propiedades:

(1) $|x|_p \geq 0$ con igualdad si y solo si $x = 0$.

(2) $|xy|_p = |x|_p|y|_p$.

(3) $|x+y|_p \leq \máx(|x|_p, |y|_p)$.

En particular la función $d(x,y) = |x-y|_p$ define una métrica en $\mathbb{Q}$. El cuerpo $\mathbb{Q}_p$ se define como la completación del espacio métrico $(\mathbb{Q}, d)$. Su construcción es similar a la de los números reales. Consideremos el conjunto $\mathscr{C}$ de sucesiones $(a_n)_{n\in\mathbb{N}}$ en $\mathbb{Q}$ que son de Cauchy con respecto a $|\cdot|_p$. Es fácil ver que si $(a_n)_n, (b_n)_n \in \mathscr{C}$ entonces $(a_n+b_n)_n, (a_nb_n)_n \in \mathscr{C}$. Luego $\mathscr{C}$ es un subanillo de $\mathbb{Q}^{\mathbb{N}}$. Sea $\mathscr{N} \subseteq \mathscr{C}$ el conjunto de sucesiones $(a_n)_n$ tales que $|a_n|_p \to 0$. Claramente $\mathscr{N}$ es cerrado con respecto la suma y de hecho es un ideal de $\mathscr{C}$: en efecto, si $(a_n)_n \in \mathscr{N}$ y $(b_n)_n \in \mathscr{C}$, entonces

existe $M > 0$ tal que $|b_n|_p \leq M$ para todo $n$ y luego $|a_n b_n|_p \leq M|a_n|_p \to 0$. Ahora veamos que $\mathcal{N}$ es un ideal maximal. Sea $a = (a_n)_n \in \mathcal{C} - \mathcal{N}$. Como $|a_n|_p \not\to 0$, existe $\varepsilon > 0$ tal que $|a_n|_p \geq \varepsilon$ para infinitos $n$. Como $(a_n)_n$ es de Cauchy, esto implica que existe $n_0 \in \mathbb{N}$ tal que $|a_n|_p \geq \varepsilon/2$ para todo $n \geq n_0$. En particular solo hay finitos $n \in \mathbb{N}$ tales que $a_n = 0$. Consideremos la siguiente sucesión:

$$b_n = \begin{cases} 1/a_n, & \text{si } a_n \neq 0, \\ 0, & \text{si } a_n = 0. \end{cases}$$

Afirmamos que $b = (b_n)_n$ es una sucesión de Cauchy. En efecto, si $m, n \neq n_0$ entonces

$$|b_m - b_n| = \frac{|a_m - a_n|}{|a_m a_n|} \leq \frac{4}{\varepsilon^2}|a_m - a_n|.$$

El término $|a_m - a_n|$ se vuelve arbitrariamente chico si uno toma $m$ y $n$ suficientemente grandes, luego lo mismo vale para $|b_m - b_n|$. Entonces $b \in \mathcal{C}$. Como $ab - 1$ solo tiene finitas entradas no nulas, $ab - 1 \in \mathcal{N}$. Luego $1 \in \mathcal{N} + (a)$, así que $\mathcal{N} + (a) = (1)$. Como esto vale para cualquier $a \in \mathcal{C} - \mathcal{N}$, concluimos que $\mathcal{N}$ es maximal. Luego $\mathbb{Q}_p = \mathcal{C}/\mathcal{N}$ es un cuerpo. Podemos extender el valor absoluto $|\cdot|_p$ a $\mathbb{Q}_p$ de la siguiente manera: dado un elemento $a = (a_n)_n \in \mathcal{C} - \mathcal{N}$, existe $\varepsilon > 0$ y $n_0 \in \mathbb{N}$ tales que $|a_n|_p \geq \varepsilon$ para todo $n \geq n_0$. Como $(a_n)_n$ es de Cauchy, existe $n_1 \geq n_0$ tal que $|a_m - a_n|_p < \varepsilon/2$ siempre que $m, n \geq n_1$. Supongamos que $|a_m|_p > |a_n|_p$, entonces

$$|a_m|_p = |a_n + (a_m - a_n)|_p \leq \text{máx}(|a_n|_p, |a_m - a_n|_p) = |a_n|_p$$

lo cual es absurdo. Análogamente, no puede pasar $|a_m|_p < |a_n|_p$. Por lo tanto $|a_m|_p = |a_n|_p$ y la sucesión $(|a_n|_p)_n$ es eventualmente constante. Entonces $|a|_p$ se define como $|a_n|_p$ para $n > 0$ suficientemente grande. Notemos que hay una inmersión $\mathbb{Q} \to \mathbb{Q}_p$, $a \mapsto \overline{(a, a, a, \dots)}$. No es difícil ver que $\mathbb{Q}_p$ es completo y que $\mathbb{Q}$ es denso en $\mathbb{Q}_p$. El anillo de enteros $p$-ádicos $\mathbb{Z}_p$ es el conjunto de números $p$-adicos $x \in \mathbb{Q}_p$ tales que $|x|_p \leq 1$. Este subconjunto es un anillo por la desigualdad ultramétrica, es decir, la propiedad (3) de $|\cdot|_p$.

Podemos definir una norma $\|\cdot\|$ en $\mathbb{Q}_p^d$:

$$\|(a_1, \dots, a_d)\| = \max_{1 \leq i \leq d} |a_i|_p.$$

Es claro que $\|\lambda v\| = |\lambda|_p \|v\|$ para todo $\lambda \in \mathbb{Q}_p$ y que $\|v + w\| \leq \text{máx}(\|v\|, \|w\|)$ para todos $v, w \in \mathbb{Q}_p^d$.

**Teorema 3.24** (Lema de Hensel multivariable)**.** *Sea $F = (f_1, \dots, f_d) \in \mathbb{Z}_p[X_1, \dots, X_d]^d$ y supongamos que $v = (v_1, \dots, v_d) \in \mathbb{Z}_p^d$ verifica*

$$\|F(v)\| < |J_F(v)|_p^2,$$

*donde $J_F$ es el Jacobiano de $F$. Entonces existe un único $\widehat{v} \in \mathbb{Z}_p^d$ tal que $F(\widehat{v}) = 0$ y $\|\widehat{v} - v\| < |J_F(v)|_p$.*

*Demostración.* Ver [Con]. □

**Corolario 3.25** (Lema de Hensel)**.** *Sea $f \in \mathbb{Z}_p[x]$ y supongamos que $a \in \mathbb{Z}_p$ satisface*

$$|f(a)|_p < |f'(a)|_p^2.$$

*Entonces existe un único $\widehat{a} \in \mathbb{Z}_p$ tal que $f(\widehat{a}) = 0$ y $|\widehat{a} - a|_p < |f'(a)|_p$.*

Llamamos $\mathbb{Z}_p$-*retículo* o simplemente *retículo* de $\mathbb{Q}_p^n$ a un $\mathbb{Z}_p$-módulo libre $\Lambda \subseteq \mathbb{Q}_p^n$ de rango $n$. Observemos que los $\mathbb{Z}_p$-retículos no son retículos en el sentido del capítulo 2 ya que no son discretos. El *determinante* de un retículo $\Lambda$ se define como $p^m$ donde $m = v_p(\det(b_1 \,|\, b_2 \,|\, \dots \,|\, b_n))$, y donde $b_1, b_2, \dots, b_n$ es cualquier base de $\Lambda$ como $\mathbb{Z}_p$-módulo.

## 3.5. Representaciones de formas cuadráticas

Esta sección está dedicada a demostrar el teorema 3.26, que usaremos en el siguiente capítulo para probar el *lema básico de Linnik*.

Supongamos que $q$ y $Q$ son dos formas cuadráticas enteras en 2 y 3 variables respectivamente. Llamamos $\mathrm{Emb}(q,Q)$ al conjunto de morfismos $j:\mathbb{Z}^2\to\mathbb{Z}^3$ tales que $Q\circ j=q$. El grupo $\mathrm{SO}_Q(\mathbb{Z})$ actúa a izquierda en $\mathrm{Emb}(q,Q)$ por composición ya que si $j\in\mathrm{Emb}(q,Q)$ y $g\in\mathrm{SO}_Q(\mathbb{Z})$ entonces $g\circ j(\mathbb{Z}^2)\subseteq\mathbb{Z}^3$ y $Q\circ g\circ j=Q\circ j=q$, lo que muestra que $g\circ j\in\mathrm{Emb}(q,Q)$. Como veremos más adelante, en vez de contar la cantidad de embeddings de $q$ en $Q$, resulta más sencillo contar la cantidad de embeddings módulo la acción de $K=\mathrm{SO}_Q(\mathbb{Z})$, es decir, la cardinalidad del conjunto $R(q,Q)=K\backslash\mathrm{Emb}(q,Q)$. Llamamos $N(q,Q)=|R(q,Q)|$.

**Teorema 3.26.** *Sea $Q$ una forma cuadrática entera no degenerada en $\mathbb{Z}^3$ y sea*

$$q(x,y)=a_1x^2+a_2xy+a_3y^2$$

*una forma cuadrática no degenerada en $\mathbb{Z}^2$. Sea $f^2$ el mayor cuadrado que divide a $\mathrm{mcd}(a_1,a_2,a_3)$. Entonces $N(q,Q)\ll_{Q,\varepsilon} f\,\mathrm{máx}(|a_1|,|a_2|,|a_3|)^{\varepsilon}$ para todo $\varepsilon>0$.*

Claramente podemos suponer que $N(q,Q)>0$, en caso contrario el teorema 3.26 vale trivialmente. Fijamos un embedding $j_0\in\mathrm{Emb}(q,Q)$ y denotamos por $L_0=j_0(\mathbb{Z}^2)$ a su imagen. Observemos que cualquier otro embedding $(\mathbb{Z}^2,q)\to(\mathbb{Z}^3,Q)$ se puede escribir como $g\circ j_0$ para algún $g\in\mathrm{SO}_Q(\mathbb{Q})$. En efecto, si $j:(\mathbb{Z}^2,q)\to(\mathbb{Z}^3,Q)$ es otro embedding, tanto $j_0$ como $j$ se extienden por extensión de escalares a morfismos $j_0,j:(\mathbb{Q}^2,q)\to(\mathbb{Q}^3,Q)$. Entonces por el teorema de extensión de Witt 3.3, existe $g\in\mathrm{O}_Q(\mathbb{Q})$ tal que $g\circ j_0=j$. Ahora, si $S=j(\mathbb{Q}^2)$ es la imagen de $j$, entonces $\mathbb{Q}^3=S\oplus S^{\perp}$ por la proposición 3.2.(*a*) y componiendo a $g$ con una reflexión que deja fijo $S$ podemos suponer que $g\in\mathrm{SO}_Q(\mathbb{Q})$. La matriz $g$ cumple que $g(L_0)=g\circ j_0(\mathbb{Z}^2)=j(\mathbb{Z}^2)\subseteq\mathbb{Z}^3$. Luego hay una aplicación sobreyectiva del conjunto

$$X=\{\overline{g}\in\mathrm{SO}_Q(\mathbb{Z})\backslash\mathrm{SO}_Q(\mathbb{Q}):\ g(L_0)\subseteq\mathbb{Z}^3\}$$

hacia $R(q,Q)$, dada por $\overline{g}\mapsto[g\circ j_0]$. Afirmamos que esta aplicación es inyectiva. Para ver esto, supongamos que $[g_1\circ j_0]=[g_2\circ j_0]$ para dos elementos $\overline{g}_1,\overline{g}_2\in X$, donde la barra denota tomar clase módulo $K$. Entonces $g_1j_0=hg_2j_0$ para algún $h\in K$ y luego la matriz $g_3=g_1^{-1}hg_2$ fija a $j_0$. En particular $g_3$ deja fijo puntualmente al subespacio vectorial $S_0\subseteq\mathbb{Q}^3$ generado por $L_0$. Como $Q$ es no degenerada en $S_0$, hay una descomposición $\mathbb{Q}^3=S_0\oplus S_0^{\perp}$. Ahora, si $y\in S_0^{\perp}$ entonces

$$\langle x,g_3y\rangle=\langle g_3x,g_3y\rangle=\langle x,y\rangle=0$$

para todo $x\in S_0$, lo que implica que $g_3y\in S_0^{\perp}$ y por lo tanto $g_3(S_0^{\perp})\subseteq S_0^{\perp}$. Como $S_0^{\perp}$ es unidimensional, $g_3$ debe actuar en este subespacio como multiplicación por una constante $\lambda\in\mathbb{Q}$. El hecho que $\det(g_3)=1$ implica que $\lambda$ debe ser 1. Luego $g_3=1$, así que $g_1=hg_2$ y entonces $\overline{g_1}=\overline{g_2}$. En conlusión $N(q,Q)=|X|$. También usaremos la notación $N(L_0)$ para denotar a $N(q,Q)$.

Es bien sabido que para estimar la cantidad soluciones de ecuaciones diofánticas suele ser útil estudiar el problema en cuerpos $p$-ádicos, y en este caso sobre $\mathbb{Q}_p$, con $p$ primo. Sea $K_p=\mathrm{SO}_Q(\mathbb{Z}_p)$ y sea $L_p=L\otimes_{\mathbb{Z}}\mathbb{Z}_p$. Como $\mathbb{Z}_p$ es playo, el morfismo canónico $L\otimes_{\mathbb{Z}}\mathbb{Z}_p\to\mathbb{Q}^3\otimes_{\mathbb{Z}}\mathbb{Z}_p\cong\mathbb{Q}_p^3$ es inyectivo y luego podemos identificar a $L_p$ con su imagen en $\mathbb{Q}_p^3$, que es un $\mathbb{Z}_p$-retículo. Además definimos

$$X_p=\{g\in\mathrm{SO}_Q(\mathbb{Z}_p)\backslash\mathrm{SO}_Q(\mathbb{Q}_p):\ g(L_p)\subseteq\mathbb{Z}_p^3\}$$

y $N_p(L_0)=|X_p|$. Hay una biyección entre $X_p$ y el cociente del conjunto $\mathrm{Emb}_p(q,Q)$ de embeddings $(\mathbb{Z}_p^2,q)\to(\mathbb{Z}_p^3,Q)$ por la acción de $K_p$. El razonamiento para justificar este hecho es análogo al caso anterior. A su vez, usando la biyección $K_pg\mapsto g^{-1}K_p$ entre coclases a izquierda y derecha, vemos que

$$N_p(L_0)=\left|\{g\in\mathrm{SO}_Q(\mathbb{Q}_p)/K_p:\ L_p\subseteq g.\mathbb{Z}_p^3\}\right|.$$

Por otro lado, $\mathrm{SO}_Q(\mathbb{Q}_p)$ actúa a izquierda en los $\mathbb{Z}_p$-retículos $\Lambda\subseteq\mathbb{Q}_p^3$. Como el estabilizador de $\mathbb{Z}_p^3$ en $\mathrm{SO}_Q(\mathbb{Q}_p)$ es $K_p$, el mapa $\mathrm{SO}_Q(\mathbb{Q}_p)/K_p\to\mathrm{SO}_Q(\mathbb{Q}_p).\mathbb{Z}_p^3$ dado por $\overline{g}\mapsto g\cdot\mathbb{Z}_p^3$ es una biyección. Luego

$$N_p(L_0)=\left|\{\Lambda\in\mathrm{SO}_Q(\mathbb{Q}_p).\mathbb{Z}_p^3:\ L_p\subseteq\Lambda\}\right|. \tag{3.14}$$

**Lema 3.27.** $N(L_0) \leq \prod_p N_p(L_0)$.

Más adelante veremos que $N_p(L_0) = 1$ para casi todo $p$, de modo que el producto que aparece en este lema es finito.

*Demostración.* Para cada primo $p$ se tiene un mapa $K\backslash \mathrm{SO}_Q(\mathbb{Q}) \to K_p\backslash \mathrm{SO}_Q(\mathbb{Q}_p)$ inducido por la inclusión $\mathrm{SO}_Q(\mathbb{Q}) \to \mathrm{SO}_Q(\mathbb{Q}_p)$. Y observemos que el mapa

$$\phi : K\backslash \mathrm{SO}_Q(\mathbb{Q}) \to \prod_p K_p\backslash \mathrm{SO}_Q(\mathbb{Q}_p), \qquad \phi(Kg) = (K_p g)_p$$

es inyectivo. En efecto, supongamos que $\phi(Kg_1) = \phi(Kg_2)$ para dos elementos $g_1$, $g_2$. Entonces $K_p g_1 = K_p g_2$ para todo primo $p$, es decir, $g_1 g_2^{-1} \in \mathrm{SO}_Q(\mathbb{Z}_p)$ para todo $p$. Esto quiere decir que las entradas de $g_1 g_2^{-1}$ no tienen denominador divisible por $p$ para ningún primo $p$, de modo que sus entradas son enteros. Se sigue que $g_1 g_2^{-1} \in K$, o equivalentemente, $Kg_1 = Kg_2$. Restringiendo este mapa a $X$ obtenemos una aplicación inyectiva $X \to \prod_p X_p$. En particular

$$N(L_0) \leq \left| \prod_p X_p \right| = \prod_p N_p(L_0). \qquad \square$$

El siguiente lema reduce el problema a estudiar formas isotrópicas sobre $\mathbb{Q}_p$:

**Lema 3.28.** *Sea $(V, Q)$ un espacio cuadrático anisotrópico sobre $\mathbb{Q}_p$ con $p \neq 2$. Entonces existe un único retículo maximal $\Lambda^+$ tal que $Q(\Lambda^+) \subseteq \mathbb{Z}_p$ y está caracterizado por*

$$\Lambda^+ = \{v \in V : Q(v) \in \mathbb{Z}_p\}. \tag{3.15}$$

*En particular $N_p(L_0) = 1$.*

*Demostración.* Basta probar que el conjunto $\Lambda^+$ definido en (3.15) es un retículo, ya que en tal caso es claro que es maximal con respecto a la propiedad $Q(\Lambda^+) \subseteq \mathbb{Z}_p$. Claramente $\Lambda^+$ es cerrado bajo multiplicación por elementos de $\mathbb{Z}_p$, así que debemos ver que es cerrado bajo la suma. Sean $x, y \in \Lambda^+$ no nulos. Podemos suponer sin pérdida de generalidad que $v_p(Q(x)) \leq v_p(Q(y))$. Supongamos que $z = x + y$ no pertenece a $\Lambda^+$. Como $Q(z) = Q(x) + Q(y) + \langle x, y\rangle \notin \mathbb{Z}_p$, se sigue que $\langle x, y\rangle \notin \mathbb{Z}_p$. Sea $n$ el menor entero tal que

$$\frac{\langle x, p^n y\rangle}{\langle x, x\rangle} \in \mathbb{Z}_p.$$

Notemos que $n > 0$ ya que $\langle x, y\rangle \notin \mathbb{Z}_p$. Consideremos el siguiente polinomio:

$$f(t) = \frac{\langle tx + p^n y, tx + p^n y\rangle}{\langle x, x\rangle} = t^2 + 2t\frac{\langle x, p^n y\rangle}{\langle x, x\rangle} + \frac{\langle p^n y, p^n y\rangle}{\langle x, x\rangle}.$$

Los coeficientes de este polinomio son enteros $p$-ádicos. Por la elección de $n$, y puesto que 2 es una unidad en $\mathbb{Z}_p$, el coeficiente lineal de $f$ es una unidad en $\mathbb{Z}_p$. Por otro lado, el coeficiente constante es divisible por $p^{2n}$ ya que $v_p(\langle x, x\rangle) \leq v_p(\langle y, y\rangle)$. Entonces el elemento $\alpha_1 = -2\frac{\langle x, p^n y\rangle}{\langle x, x\rangle} \in \mathbb{Z}_p$ es una raíz de $f$ módulo $p$ y $f'(\alpha_1) \equiv \alpha_1 \not\equiv 0 \pmod p$. Por el lema de Hensel, $f$ tiene una raíz $\alpha \in \mathbb{Z}_p$. Pero esto implica que

$$\langle \alpha x + p^n y, \alpha x + p^n y\rangle = 0,$$

contradiciendo que $Q$ es anisotrópica. Por lo tanto $z \in \Lambda^+$. $\square$

La clave de la demostración de 3.26 reside en los siguientes dos lemas. Antes de enunciarlos hagamos la siguiente observación. Sea $Q$ una forma $p$-isotrópica, es decir, isotrópica en $\mathbb{Q}_p$, y sea $Q_0(x, y, z) = xy + z^2$. Entonces $(\mathbb{Q}_p^3, Q)$ contiene un plano hiperbólico, es decir, existen $e_1, e_2 \in \mathbb{Q}_p^3$ isotrópicos tales que $\langle e_1, e_2\rangle = 1$. Llamando $H = \mathbb{Q}_p e_1 + \mathbb{Q}_p e_2$, tenemos una descomposición $\mathbb{Q}_p^3 = H \oplus H^\perp$. Sea $e_3$ un generador de $H^\perp$. Como $H$ es un subespacio regular, $Q$ es no degenerada en $H^\perp$, así que $Q(e_3) = \lambda \neq 0$. Se sigue que

$$Q(xe_1 + ye_2 + ze_3) = xy + \lambda z^2.$$

Reemplazando a $e_1$ por $\lambda e_1$ obtenemos $Q(xe_1 + ye_2 + ze_3) = \lambda Q_0(x, y, z)$. Multiplicando estos vectores por una constante podemos suponer que $e_1, e_2, e_3 \in \mathbb{Z}_p^3$. Entonces hay un morfismo de retículos cuadráticos

$$\psi : (\mathbb{Z}_p^3, \lambda^{-1}Q) \to (\mathbb{Z}_p^3, Q_0), \qquad \psi(x, y, z) = xe_1 + ye_2 + ze_3.$$

Como $\mathrm{Emb}_p(q, Q) = \mathrm{Emb}_p(\lambda^{-1}q, \lambda^{-1}Q)$, hay un mapa $\psi_* : \mathrm{Emb}_p(q, Q) \to \mathrm{Emb}_p(\lambda^{-1}q, Q_0)$, dado por $j \mapsto \psi \circ j$, que es inyectivo. Para relacionar $N_p(q, Q)$ con $N_p(\lambda^{-1}q, Q_0)$ debemos tomar en cuenta las acciones de $\mathrm{SO}_Q(\mathbb{Z}_p)$ y $\mathrm{SO}_{Q_0}(\mathbb{Z}_p)$. Notemos que si $T \in K_p$ entonces $\psi T \psi^{-1} \in \mathrm{SO}_{Q_0}(\mathbb{Q}_p)$ ya que

$$Q_0(\psi T \psi^{-1}(v)) = Q(T\psi^{-1}(v)) = Q_0\psi^{-1}(v) = Q(v).$$

Sin embargo $\psi T \psi^{-1}$ puede no tener coeficientes enteros. Sea $K_{p,0} = \mathrm{SO}_{Q_0}(\mathbb{Z}_p)$ y consideremos los subgrupos

$$G_p = K_p \cap \psi^{-1} K_{p,0} \psi \subseteq K_p, \qquad G_{p,0} = \psi K_p \psi^{-1} \cap K_{p,0} \subseteq K_{p,0}.$$

Es claro que hay un isomorfismo $G_p \to G_{p,0}$, $x \mapsto \psi x \psi^{-1}$. Además el mapa $\psi_*$ manda $G_p$-órbitas en $G_{p,0}$-órbitas, ya que si $j_1, j_2 \in \mathrm{Emb}_p(q, Q)$ son tales que $j_2 = Tj_1$ para algún $T \in G_p$, entonces $\psi j_2 = \psi T j_1 = (\psi T \psi^{-1})\psi j_2$. Luego hay una aplicación bien definida

$$G_p \backslash \mathrm{Emb}_p(q, Q) \to G_{p,0} \backslash \mathrm{Emb}_p(\lambda^{-1}q, Q_0)$$

que es inyectiva.

Ahora afirmamos que $(K_p : G_p)$ es finito. Mirando los endomorfismos de $\mathbb{Q}_p^3$ como matrices, vemos que existe $N > 0$ tal que $p^N \psi^{-1} \in M_3(\mathbb{Z}_p)$. Sea $H = \mathrm{Ker}\left(K_p \to \mathrm{SO}_Q(\mathbb{Z}_p/p^N\mathbb{Z}_p)\right)$. Entonces $H$ tiene índice finito ya que $\mathrm{SO}_Q(\mathbb{Z}_p/p^N\mathbb{Z}_p)$ es finito. Ahora, todo elemento de $H$ se escribe como $I + p^N a$ para algún $a \in M_3(\mathbb{Z}_p)$. Entonces $\psi(I + p^N a)\psi^{-1} = I + p^N \psi a \psi^{-1}$ pertenece a $M_3(\mathbb{Z}_p)$. Entonces $\psi(I + p^N a)\psi^{-1} \in K_{p,0}$ y luego $H \subseteq G_p$, lo que demuestra la afirmación. Similarmente uno ve que $(K_{p,0} : G_{p,0})$ es finito. Como $\psi$ solo depende de $Q$ y $p$, tenemos $(K_{p,0} : G_{p,0}) = O_{Q,p}(1)$. Luego

$$N_p(q, Q) \leq |G_p \backslash \mathrm{Emb}_p(q, Q)| \leq |G_{p,0} \backslash \mathrm{Emb}_p(\lambda^{-1}q, Q_0)| \ll_{Q,p} N_p(\lambda^{-1}q, Q_0). \tag{3.16}$$

Finalmente notemos que $\mathrm{disc}(Q_0) = 2$ así que $Q_0$ es no degenerada módulo $p$ para $p \neq 2$. Por lo tanto basta acotar $N_p(L_0)$ cuando $p \nmid \mathrm{disc}(Q)$ o $p = 2$. En el primer caso $Q$ es no degenerada en $\mathbb{F}_p^3$.

Sea $p \neq 2$ un primo y sea $L$ un retículo de rango 2 en un espacio cuadrático $(V, q)$ sobre $\mathbb{Q}_p$. Entonces por [Cas78, Teo. 8.3.1] existe una base $e_1, e_2$ de $L$ tal que

$$q(xe_1 + ye_2) = u_1 p^a x^2 + u_2 p^b y^2$$

donde $u_1, u_2 \in \mathbb{Z}_p^\times$ y $a \leq b$ son enteros. Los enteros $a, b$ son los únicos con esta propiedad y los llamaremos las *invariantes* del par $(L, q)$. En el caso $p = 2$ no siempre es posible llevar a $q$ a una forma diagonal, sin embargo, por [Cas78, Teo. 8.4.1] existe una base $e_1, e_2$ de $L$ en la que $q$ toma una de las siguientes formas:

$$q(xe_1 + ye_2) = u_1 2^a x^2 + u_2 2^b y^2,$$

donde $u_1, u_2 \in \mathbb{Z}_2^\times$ y $a \leq b$, o bien

$$q(xe_1 + ye_2) = 2^a xy, \qquad a \in \mathbb{Z}, \tag{3.17}$$

$$q(xe_1 + ye_2) = 2^a(x^2 + xy + y^2), \qquad a \in \mathbb{Z}. \tag{3.18}$$

En el primer caso diremos que $q$ tiene invariantes $(a, b)$ *de tipo diagonal*.

Si $(L, q)$ es un retículo cuadrático sobre $\mathbb{Z}$, definimos sus *p-invariantes* como las invariantes de $(L \otimes_{\mathbb{Z}} \mathbb{Z}_p, \widetilde{q})$ donde $\widetilde{q}$ es la única extensión posible de $q$ a una forma cuadrática sobre $\mathbb{Q}_p$.

**Lema 3.29.** *Sea $p > 2$ primo y sea $Q$ una forma cuadrática isotrópica en $\mathbb{Q}_p^3$ tal que $p \nmid \mathrm{disc}(Q)$. Si $L \subseteq \mathbb{Q}_p^3$ es un $\mathbb{Z}_p$-submódulo de rango 2 tal que $Q|_L$ tiene invariantes $(a, b)$ entonces*

$$N(L) \ll (b + 1)^2 p^{\lfloor a/2 \rfloor}. \tag{3.19}$$

*Además, si $(a, b) = (0, 0)$ entonces $N(L) = 1$.*

**Lema 3.30.** *Sea $Q(x,y,z) = xy+z^2$ y sea $L \subseteq \mathbb{Q}_2^3$ un $\mathbb{Z}_2$-submódulo de rango 2 tal que $Q(L) \subseteq \mathbb{Z}_2$. Si $Q|_L$ tiene invariantes $(a,b)$ de tipo diagonal, entonces*

$$N(L) \ll (b+1)^2 2^{\lfloor a/2 \rfloor}. \tag{3.20}$$

*Si $q$ es de la forma (3.17) o (3.18) entonces*

$$N(L) \ll (a+1)^2 2^{\lfloor a/2 \rfloor}. \tag{3.21}$$

Vamos a postergar la demostración de estos lemas para la siguiente sección. Ahora veremos cómo estos lemas implican el teorema 3.26. Para esto necesitamos la siguiente cota elemental:

**Lema 3.31.** *Si $\tau(n)$ denota la cantidad de divisores positivos de un entero $n \in \mathbb{N}$, entonces*

$$\tau(n) \ll_\varepsilon n^\varepsilon, \tag{3.22}$$

*para todo $\varepsilon > 0$.*

*Demostración.* Fijemos $\varepsilon > 0$. Si $n = p_1^{a_1} \dots p_r^{a_r}$ es la factorización en primos de $n$, entonces

$$\frac{\tau(n)}{n^\varepsilon} = \prod_{i=1}^{r} \frac{a_i+1}{p_i^{\varepsilon a_i}}. \tag{3.23}$$

Si $p_i \geq \exp(1/\varepsilon)$, entonces $p_i^{\varepsilon a_i} \geq e^{a_i} \geq 1 + a_i$, así que la contribución del $i$-ésimo factor en (3.23) es a lo sumo 1. Para los primos $p < \exp(1/\varepsilon)$, la expresión $(k+1)/p^{\varepsilon k}$ tiende a 0 cuando $k \to \infty$, así que $(a_i+1)/p^{\varepsilon a_i} = O_\varepsilon(1)$. Como solo hay finitos de estos primos, deducimos que

$$\prod_{p < \exp(1/\varepsilon)} \frac{v_p(n)+1}{p^{\varepsilon v_p(n)}} = O_\varepsilon(1)$$

y luego $\tau(n) n^{-\varepsilon} = O_\varepsilon(1)$. □

*Demostración de 3.26 asumiendo 3.29 y 3.30.* Por el lema 3.28, si $Q$ es $p$-anisotrópica para un primo $p$ distinto de 2 entonces $N_p(L_0) = 1$, así que

$$N(L_0) = N_2(L_0) \prod_{\substack{p \neq 2 \\ Q \text{ } p\text{-isotrópica}}} N_p(L_0). \tag{3.24}$$

Para $p$ primo, sean $(a_p, b_p)$ las invariantes de $Q|_{L_p}$. Como $Q \circ j_0 = q$, estas invariantes coinciden con las respectivas $p$-invariantes de $q$. Supongamos primero que $p \neq 2$. Esto quiere decir que $q$ es $\mathrm{GL}_2(\mathbb{Z}_p)$-equivalente a una forma cuadrática $up^{a_p}x^2 + vp^{b_p}y^2$ donde $u, v \in \mathbb{Q}_p^\times$. Ahora notemos que tanto $d = \mathrm{disc}(q)$ como $\min_j v_p(a_j)$ son invariantes bajo la acción de $\mathrm{GL}_2(\mathbb{Z}_p)$. Entonces $v_p(\mathrm{disc}(q)) = a_p + b_p$ y $\min_j v_p(a_j) = a_p$. En particular, si $v_p(\mathrm{disc}(q)) = 0$ entonces $a_p = b_p = 0$, en cuyo caso $N_p(L_0) = 1$ por el lema 3.29. Además $\mathrm{disc}(q) > b_p$ y $v_p(f) = \lfloor a_p/2 \rfloor$. Si $p \nmid d$, por el lema 3.29 tenemos

$$N_p(L_0) \ll (b_p+1)^2 p^{\lfloor a_p/2 \rfloor} \leq (v_p(d)+1)^2 p^{v_p(f)}$$

Si $p \mid d$, vimos que $N_p(L_0) \ll_{Q,p} N_p(\lambda q, Q_0)$ donde $\lambda = \lambda(Q,p) \in \mathbb{Q}_p$. Además, las invariantes de $\lambda q$ son $(a + O_Q(1), b + O_Q(1))$, así que $N_p(L_0) \ll_Q (b_p+1)^2 p^{\lfloor a_p/2 \rfloor}$ también en este caso.

Análogamente, en el caso $p = 2$ también tenemos $N_2(L_0) \ll_Q N_2(\lambda q, Q_0)$. En este caso $v_2(d) = a_2 + b_2 + 2 \geq b_2$ si $Q|_L$ toma la forma diagonal y $v_2(d) = 2a_2$ si toma las formas (3.17) o (3.18). Además, $v_p(f) = \lfloor a_2/2 \rfloor$. En ambos casos tenemos

$$N_2(L_0) \ll (v_2(d)+1)^2 2^{v_2(f)},$$

por el lema 3.30. Juntando estas cotas, obtenemos

$$\begin{aligned} N(L_0) &\ll_Q \prod_{p \mid 2d} (v_p(d)+1)^2 \prod_{p \mid 2d} p^{v_p(f)} \\ &= \tau(d)^2 f \ll_\varepsilon |d|^{\varepsilon/2} f \ll f \left( \max_{1 \leq j \leq 3} |a_j| \right)^\varepsilon, \end{aligned}$$

ya que $|d| = |a_2^2 - 4a_1a_3| \ll (\max_j |a_j|)^2$. □

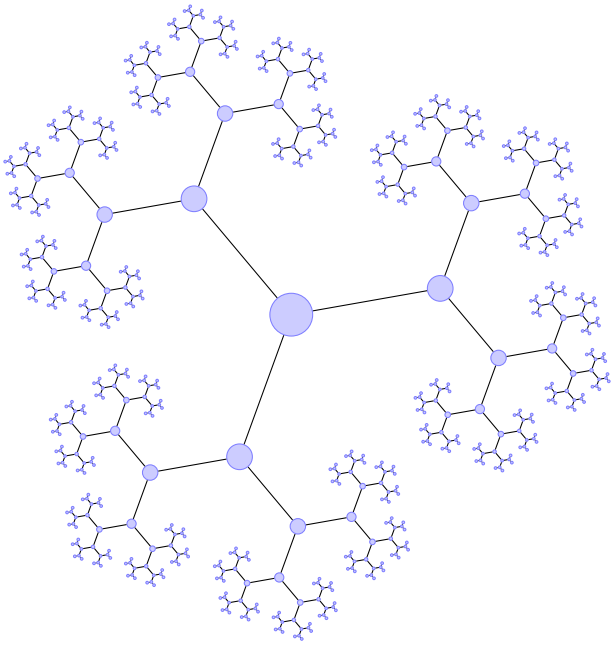

Figura 3.1: El árbol de Bruhat-Tits de $Q_0 = xy + z^2$ en $\mathbb{Q}_2$.

## 3.6. El árbol de Bruhat-Tits

En esta sección asumimos que $Q$ es una forma cuadrática $p$-isotrópica en tres variables con $p \nmid \operatorname{disc}(Q)$, o tal que $Q = xy + z^2$. Para demostrar los lemas 3.29 y 3.30 usaremos que le podemos dar a $\mathrm{SO}_Q(\mathbb{Q}_p).\mathbb{Z}_p^3$ una estructura de grafo $(p+1)$-regular, conocido como el *árbol de Bruhat-Tits*[1] de $\mathrm{SO}_Q(\mathbb{Q}_p)$.

En general, si $G$ es un grupo algebraico reductivo sobre $\mathbb{Q}_p$, entonces existe un complejo simplicial $\Delta = \Delta_G$ llamado el *edificio de Bruhat-Tits* que es contráctil y de dimensión finita. Además $G$ actúa en $\Delta$ por morfismos simpliciales. En el caso de $G = \mathrm{SO}_Q(\mathbb{Q}_p)$, $\Delta$ tiene dimensión 1, así que es un grafo. La construcción explícita de $\Delta$ para grupos ortogonales está detallada en [Gar95, §19.8]. En nuestro caso el conjunto $\mathcal{T}_Q$ de vértices consiste de los retículos $\Lambda \subseteq \mathbb{Q}_p^3$ tales que $Q(\Lambda) \subseteq \mathbb{Z}_p$ y que son maximales con esta propiedad. Se puede ver (sin recurrir a la teoría de edificios, cf. [GI63, Teo. 4.11]) que $\mathrm{SO}_Q(\mathbb{Q}_p)$ actúa transitivamente en $\mathcal{T}_Q$ así que $\mathcal{T}_Q = \mathrm{SO}_Q(\mathbb{Q}_p)\Lambda_0$ para cualquier $\Lambda_0 \in \mathcal{T}_Q$. Es fácil ver que $\mathbb{Z}_p^3 \in \mathcal{T}_Q$, así que $\mathcal{T}_Q = \mathrm{SO}_Q(\mathbb{Q}_p).\mathbb{Z}_p^3$.

Queda describir cuáles vértices son adyacentes. Dos retículos $\Lambda, \Gamma \in \mathcal{T}_Q$ se dicen adyacentes si $(\Lambda : \Lambda \cap \Gamma) = (\Gamma : \Lambda \cap \Gamma) = p$. Para probar los resultados de esta sección es necesario entender con mayor detalle la estructura de adyacencia. Supongamos que $\Lambda, \Gamma \in \mathcal{T}_Q$ son adyacentes. Tomemos $\gamma \in \Gamma - \Lambda$, $\delta \in \Lambda - \Gamma$. Como $(\Gamma : \Lambda \cap \Gamma) = p$, tenemos que $v = p\gamma \in \Lambda$. Además $Q(v) = p^2 Q(\gamma)$ es divisible por $p^2$ y $\langle v, z\rangle \equiv 0 \pmod{p}$ para todo $z \in \Lambda \cap \Gamma$. En otras palabras, $\Lambda \cap \Gamma$ está contenido en el conjunto

$$S_v = \{z \in \Lambda : \langle v, z\rangle \equiv 0 \pmod{p}\}. \tag{3.25}$$

Afirmamos que $\Lambda \cap \Gamma = S_v$. En efecto, sea $z \in S_v$. Puesto que $\Lambda = (\Lambda \cap \Gamma) + \mathbb{Z}_p\delta$, podemos escribir $z = z_0 + \lambda\delta$ con $z_0 \in \Lambda \cap \Gamma$ y $\lambda \in \mathbb{Z}_p$. Notemos que $\langle \gamma, \delta\rangle \notin \mathbb{Z}_p$ ya que en caso contrario $Q(\Lambda+\Gamma) \subseteq \mathbb{Z}_p$, lo que contradice la maximalidad de $\Lambda$ y $\Gamma$. Luego $p \nmid \langle v, \delta\rangle$. Entonces la congruencia

$$0 \equiv \langle v, z\rangle \equiv \lambda\langle v, \delta\rangle \pmod{p}$$

implica que $\lambda \equiv 0 \pmod{p}$, y luego $\lambda\delta \in \Lambda \cap \Gamma$. Se sigue que $z = z_0 + \lambda\delta \in \Lambda \cap \Gamma$, como queríamos ver. En conclusión, tenemos que

$$\Gamma = \mathbb{Z}_p\gamma + \Lambda \cap \Gamma = \tfrac{1}{p}\mathbb{Z}_p v + S_v. \tag{3.26}$$

[1] Se puede ver que este grafo efectivamente es un árbol, pero no necesitaremos este resultado.

Dado un retículo $\Lambda \in \mathcal{T}_Q$ y un elemento $v \in \Lambda$ primitivo (i.e. tal que $v \notin p\Lambda$) tal que $p^2 \mid Q(v)$, denotamos $\Lambda(\overline{v})$ al lado derecho de (3.26). Es posible extender la definición de $\Lambda(\overline{v})$ al caso de un elemento $v \in \Lambda$ primitivo tal que $p \mid Q(v)$. Para esto uno elige $z_0 \in \Lambda$ de modo que $p^2 \mid Q(v+pz_0)$ y define $\Lambda(\overline{v}) = \Lambda(\overline{v+pz_0})$. Veamos que tal $z_0$ existe. Como $p \nmid \operatorname{disc}(Q)$, la forma $Q$ es no degenerada en $\Lambda/p\Lambda$. Esto es claro si $\Lambda = \mathbb{Z}_p^3$. Ahora, si $\Lambda = g.\mathbb{Z}_p^3$ con $g \in \mathrm{SO}_Q(\mathbb{Q}_p)$ entonces $g$ induce un isomorfismo $\mathbb{Z}_p^3/p\mathbb{Z}_p^3 \to \Lambda/p\Lambda$ que conmuta con $Q$. Luego existe $w_0 \in \Lambda/p\Lambda$ tal que $p \nmid \langle v, w_0\rangle$. Puesto que

$$Q(v + \lambda p w_0) = Q(v) + \lambda^2 p^2 Q(w_0) + \lambda p \langle v, w_0\rangle,$$

podemos hallar $\lambda \in \mathbb{Z}_p$ tal que $p^2 \mid Q(v + p\lambda w_0)$, en cuyo caso podemos tomar $z_0 = \lambda w_0$. Notemos que $\Lambda(\overline{v})$ no depende de la elección de $z_0 \in \Lambda$, ya que si uno toma otro $z_1 \in \Lambda$ con $p^2 \mid Q(v + pz_1)$, es fácil ver que $z_0 - z_1 \in S_v$. Además $\Lambda(\overline{v})$ solo depende de la recta $\mathbb{F}_p\overline{v} \subseteq \Lambda/p\Lambda$, es decir, si uno toma cualquier otro $v_1 \in \Lambda$ primitivo tal que $\overline{v}_1 \in \mathbb{F}_p\overline{v}$, entonces $\Lambda(\overline{v}_1) = \Lambda(\overline{v})$.

Hemos visto que todo retículo adyacente a $\Lambda$ es de la forma $\Lambda(\overline{v})$. Recíprocamente, veamos que $\Lambda(\overline{v}) \in \mathcal{T}_Q$ es adyacente a $\Lambda$ para todo $v \in \Lambda$ con $p \mid Q(v)$. En primer lugar $\Lambda(\overline{v})$ es integral, es decir, $Q(\Lambda(\overline{v})) \subseteq \mathbb{Z}_p$. Para ver esto podemos suponer que $p^2 \mid Q(v)$. En ese caso, $\Lambda(\overline{v}) = \frac{1}{p}\mathbb{Z}_p v + S_v$. Si $\lambda \in \mathbb{Z}_p$ y $z \in S_v$, entonces

$$Q(\lambda v/p + z) = \frac{\lambda^2}{p^2}Q(v) + Q(z) + \lambda\frac{\langle v, z\rangle}{p} \in \mathbb{Z}_p$$

ya que los tres sumandos pertenecen a $\mathbb{Z}_p$. Se sigue que $\Lambda(\overline{v})$ está contenida en un retículo $\Gamma$ que es maximal con la propiedad de ser integral. Como veremos más adelante -ver 3.36.(*a*)-, tenemos que $\Lambda \cap \Lambda(\overline{v}) = S_v$. Luego $(\Lambda : \Lambda \cap \Lambda(\overline{v})) = p = (\Lambda(\overline{v}) : \Lambda \cap \Lambda(\overline{v}))$. Entonces $\Lambda$ y $\Lambda(\overline{v})$ deben tener igual determinante. Por otro lado, como los retículos $\Lambda, \Gamma \in \mathcal{T}_Q$ están en la misma $\mathrm{SO}_Q(\mathbb{Q}_p)$-órbita, también tienen el mismo determinante. Por lo tanto $\Lambda(\overline{v}) = \Gamma \in \mathcal{T}_Q$. Además $(\Lambda : \Lambda \cap \Lambda(\overline{v})) = (\Lambda : S_v) = p$ ya que $S_v$ es el núcleo del morfismo de grupos

$$\Lambda \to \mathbb{F}_p, \qquad w \mapsto \overline{\langle v, w\rangle}.$$

Dados dos retículos $\Lambda, \Gamma \subseteq \mathbb{Q}_p^n$ de igual determinante, denotamos por $I(\Lambda, \Gamma)$ al índice común $(\Lambda : \Lambda \cap \Gamma) = (\Gamma : \Lambda \cap \Gamma)$. Definimos la *distancia* $d(\Lambda, \Gamma)$ entre dos retículos $\Lambda, \Gamma \in \mathcal{T}_Q$ como la menor longitud $d$ de un camino $\Lambda_0 = \Lambda, \Lambda_1, \ldots, \Lambda_d = \Gamma$ que los conecta, es decir, tal que $\Lambda_i$ y $\Lambda_{i+1}$ son adyacentes para todo $i$.

**Lema 3.32.** *Sea $(V, Q)$ un espacio cuadrático sobre $\mathbb{Q}_p$. Sean $\Lambda, \Gamma \subseteq V$ dos $\mathbb{Z}_p$-retículos integrales con respecto a $Q$ de igual determinante y supongamos que $I(\Lambda, \Gamma) = p^n$. Entonces existe un $\mathbb{Z}_p$-retículo integral $\Delta \subseteq V$ tal que $I(\Lambda, \Delta) = p$, $I(\Gamma, \Delta) = p^{n-1}$ y $\Lambda \cap \Gamma \subseteq \Delta$.*

*Demostración.* Sea $a \in \Gamma - \Lambda$ tal que $pa \in \Lambda$. Como $Q$ es integral en $\Lambda$, tenemos

$$p\langle a, b\rangle = \langle pa, b\rangle \in \mathbb{Z}_p, \qquad \text{para todo } b \in \Lambda$$

y por la integralidad de $\Gamma$,

$$\langle a, c\rangle \in \mathbb{Z}_p, \qquad \text{para todo } c \in \Gamma.$$

Consideremos el retículo

$$\Lambda^{(1)} = \{c \in \Lambda : \langle a, c\rangle \in \mathbb{Z}_p\}.$$

Entonces $\Lambda \cap \Gamma \subseteq \Lambda^{(1)}$. Además $(\Lambda : \Lambda^{(1)}) \mid p$ ya que $\Lambda^{(1)}$ es el núcleo del morfismo de grupos

$$\psi : \Lambda \to \mathbb{Z}/p\mathbb{Z}, \qquad \psi(x) = \overline{\langle pa, x\rangle},$$

donde la barra denota tomar clase módulo $p$. Luego podemos elegir un sub-retículo $\Lambda^{(2)} \subseteq \Lambda^{(1)}$ tal que $(\Lambda : \Lambda^{(2)}) = p$ y $\Lambda \cap \Gamma \subseteq \Lambda^{(2)}$. Ahora tomamos

$$\Delta = \Lambda^{(2)} + \mathbb{Z}_p a.$$

El $\mathbb{Z}_p$-módulo $\Delta$ es libre ya que es finitamente generado y sin torsión. Además es un retículo por contener a $\Lambda^{(2)}$. Afirmamos que $\Delta$ tiene las propiedades requeridas. En primer lugar, $\Delta$ es integral,

ya que si $x = c + \lambda a \in \Delta$, donde $c \in \Lambda^{(2)}$ y $\lambda \in \mathbb{Z}_p$, entonces $\langle a, c\rangle \in \mathbb{Z}_p$ puesto que $c \in \Lambda^{(1)}$, lo que implica que

$$Q(x) = Q(c) + \lambda^2 Q(a) + \lambda\langle a, c\rangle \in \mathbb{Z}_p.$$

Ahora, notemos que $pa \in \Lambda \cap \Gamma \subseteq \Lambda^{(2)}$ pero $a \notin \Lambda^{(2)}$ ya que $a$ no pertenece a $\Lambda$. Se sigue que $(\Delta : \Lambda^{(2)}) = p$. Y esto a su vez implica que $\Lambda^{(2)} = \Delta \cap \Lambda$. En efecto, es claro que $\Lambda^{(2)} \subseteq \Delta \cap \Lambda$. Si $\Lambda^{(2)} \subsetneq \Delta \cap \Lambda \subseteq \Delta$, esto fuerza la igualdad $\Delta \cap \Lambda = \Delta$, en cuyo caso $\Delta \subseteq \Lambda$. Pero esto se contradice con que $a \notin \Lambda$. Por lo tanto $\Lambda^{(2)} = \Delta \cap \Lambda$ es un retículo que tiene índice $p$ tanto en $\Lambda$ como en $\Delta$, de modo que $I(\Delta, \Lambda) = p$.

Queda ver que $I(\Gamma, \Delta) = p^{n-1}$. Para esto, basta probar que $\Delta \cap \Gamma = \Lambda \cap \Gamma + \mathbb{Z}_p a$, ya que en tal caso

$$(\Gamma : \Delta \cap \Gamma) = \frac{(\Gamma : \Lambda \cap \Gamma)}{(\Delta \cap \Gamma : \Lambda \cap \Gamma)} = \frac{p^n}{p} = p^{n-1}.$$

Es claro que $\Lambda \cap \Gamma + \mathbb{Z}_p a \subseteq \Delta \cap \Gamma$. Recíprocamente, si $x \in \Delta \cap \Gamma$, lo podemos escribir como $x = c + \lambda a$ con $c \in \Lambda^{(2)}$ y $\lambda \in \mathbb{Z}_p$. Entonces $c = x - \lambda a \in \Gamma$. Se sigue que $c \in \Lambda \cap \Gamma$ y luego $x \in \Lambda \cap \Gamma + \mathbb{Z}_p a$. □

**Corolario 3.33.** *$\mathcal{T}_Q$ es un grafo conexo.*

**Proposición 3.34.** *Sea $Q$ una forma cuadrática isotrópica en $\mathbb{Q}_p^3$ y supongamos que $p \nmid \operatorname{disc}(Q)$. Entonces $\mathcal{T}_Q$ es un grafo $(p+1)$-regular.*

*Demostración.* Fijemos $\Lambda \in \mathcal{T}_Q$ y sea $V = \Lambda/p\Lambda$, que es un espacio vectorial sobre $\mathbb{F}_p$. Consideremos el anillo de coordenadas $\mathbb{F}_p[V] = \operatorname{Sym}(V^*)$ de $V$, que es igual al álgebra simétrica de $V^*$. Recordemos que el álgebra simétrica de un espacio vectorial $W$ se define como $\operatorname{Sym}(W) = T(W)/I$, donde $T(W) = \bigoplus_{k=0}^{\infty} W^{\otimes k}$ es el álgebra tensorial e $I$ es el ideal generado por $\alpha \otimes \beta - \beta \otimes \alpha$, con $\alpha, \beta$ variando en $W$. Como $I$ es un ideal homogéneo, $\operatorname{Sym}(V)$ es un álgebra graduada conmutativa sobre $\mathbb{F}_p$. De hecho, $\mathbb{F}_p[V]$ es isomorfa a $\mathbb{F}_p[x, y, z]$ como álgebra graduada. Los elementos homogéneos de grado 1 son exactamente los elementos no nulos de $V^*$. Finalmente notemos que todo elemento de $\mathbb{F}_p[V]$ determina una función $V \to \mathbb{F}_p$. En efecto, por la propiedad universal del álgebra simétrica, la inclusión $V^* \hookrightarrow \mathbb{F}_p^V$ induce un morfismo de $\mathbb{F}_p$-álgebras $\mathbb{F}_p[V] \to \mathbb{F}_p^V$.

Podemos ver a $Q$ como un elemento homogéneo de grado 2. Afirmamos que $Q$ es irreducible en $\mathbb{F}_p[V]$. Si esto no pasa entonces existen elementos homogéneos $\ell_1, \ell_2 \in \mathbb{F}_p[V]$ de grado 1 (es decir, $\ell_1, \ell_2 \in V^*$) tales que $Q = \ell_1 \ell_2$. En tal caso,

$$\langle v, w\rangle = Q(v + w) - Q(v) - Q(w) = \ell_1(v)\ell_2(w) + \ell_2(v)\ell_1(w),$$

para $v, w \in V$. Se sigue que si elegimos $z \in \operatorname{Ker}(\ell_1) \cap \operatorname{Ker}(\ell_2)$ no nulo, entonces $\langle v, z\rangle = 0$ para todo $v \in V$. Pero esto contradice que $Q$ es no degenerada en $V$. Por lo tanto $Q$ es irreducible.

Vimos que el número de vértices adyacentes a $\Lambda$ en $\mathcal{T}_Q$ es igual al cardinal del conjunto $Z(Q)$ de ceros de $Q$ en $\mathbb{P}(V) \cong \mathbb{P}^2(\mathbb{F}_p)$. Como $Q$ es polinomio irreducible de grado 2, $Z(Q)$ es una cónica. Además $Z(Q)$ tiene puntos en $\mathbb{F}_p^3$ ya que $Q$ es isotrópica. Luego $Z(Q)$ es isomorfo a $\mathbb{P}^1(\mathbb{F}_p)$. En particular $|Z(Q)| = p + 1$. □

**Corolario 3.35.** *Para todo $\Gamma \in \mathcal{T}_Q$ y para todo $r \geq 1$ hay $O(p^r)$ retículos a distancia menor o igual que $r$ de $\Gamma$.*

*Demostración.* Basta ver que hay a lo sumo $(p+1)p^{r-1}$ a distancia igual a $r$ de $\Gamma$, para todo $r \geq 1$. Para $r = 1$ esto se deduce de la proposición anterior. Sea $r > 1$ y supongamos que hay $\leq (p+1)p^{r-2}$ retículos a distancia $r-1$ de $\Gamma$. Si $\Lambda \in \mathcal{T}_Q$ está a distancia $r$ de $\Gamma$, entonces debe existir un retículo $\Lambda'$ adyacente a $\Lambda$ que esté a distancia $r-1$ de $\Gamma$. Ahora, cada $\Lambda' \in \mathcal{T}_Q$ con $d(\Gamma, \Lambda') = r - 1$ tiene a lo sumo $p$ vecinos a distancia $r$ de $\Gamma$, ya que $\Lambda'$ debe tener al menos un vecino a distancia $r-2$ de $\Gamma$. Luego hay a lo sumo $(p+1)p^{r-1}$ retículos en $\mathcal{T}_Q$ a distancia $r$ de $\Gamma$. □

**Lema 3.36.** *Sea $\Lambda \subseteq \mathbb{Q}_p^3$, $\Lambda \in \mathcal{T}_Q$ y sean $\overline{v}_1, \overline{v}_2, \overline{v}_3 \in \Lambda/p\Lambda$ elementos isotrópicos que generan rectas distintas. Entonces*

(a) $\Lambda \cap \Lambda(\overline{v}_1) = S_{v_1} = \{z \in \Lambda : \langle v_1, z\rangle \equiv 0 \pmod{p}\}$.

(*b*) *Si* $w \in \Lambda$ *es tal que* $(\mathbb{F}_p\overline{v}_1 + \mathbb{F}_p\overline{v}_2)^\perp = \mathbb{F}_p\overline{w}$ *entonces*

$$\Lambda(\overline{v}_1) \cap \Lambda(\overline{v}_2) = \mathbb{Z}_p w + p\Lambda.$$

(*c*) *Si* $p \nmid \operatorname{disc}(Q)$ *entonces*

$$\Lambda(\overline{v_1}) \cap \Lambda(\overline{v_2}) \cap \Lambda(\overline{v_3}) = p\Lambda.$$

*Demostración.* Como vimos antes, podemos suponer que el elemento $v_i \in \Lambda$ que representa a $\overline{v}_i$ es tal que $p^2 \mid Q(v_i)$ para todo $i$.

(*a*) Es claro que $S_{v_1} \subseteq \Lambda \cap \Lambda(\overline{v}_1)$. Recíprocamente, supongamos que $x \in \Lambda \cap \Lambda(\overline{v}_1)$. Entonces $x$ se escribe como $x = \frac{\lambda}{p}v_1 + z$ para algún $\lambda \in \mathbb{Z}_p$ y $z \in S_{v_1}$. En tal caso $x - z = \frac{\lambda}{p}v_1 \in \Lambda$, lo que implica que $p \mid \lambda$ ya que $v_1$ es primitivo. Como $v_1 \in S_{v_1}$ concluimos que $x \in S_{v_1}$.

(*b*) La inclusión $\mathbb{Z}_p w + p\Lambda \subseteq \Lambda(\overline{v}_1) \cap \Lambda(\overline{v}_2)$ es clara; para ver la otra, sea $x \in \Lambda(\overline{v}_1) \cap \Lambda(\overline{v}_2)$. Por (3.26) existen $\lambda_1, \lambda_2 \in \mathbb{Z}_p$, $z_1 \in S_{v_1}$ y $z_2 \in S_{v_2}$ tales que

$$x = \frac{\lambda_1}{p}v_1 + z_1 = \frac{\lambda_2}{p}v_2 + z_2.$$

Multiplicando esta ecuación por $p$ obtenemos $\lambda_1 v_1 = \lambda_2 v_2 + p(z_2 - z_1)$ y luego $\lambda_1\overline{v}_1 = \lambda_2\overline{v}_2$. Por hipótesis $\overline{v}_1$ y $\overline{v}_2$ son linealmente independientes, así que $\lambda_1 \equiv \lambda_2 \equiv 0 \pmod p$. En particular $x \in \Lambda$. Además, tenemos que $\langle x, v_1\rangle \equiv 0 \equiv \langle x, v_2\rangle \pmod p$, lo que implica que $\overline{x} \in \mathbb{F}_p\overline{w}$. Luego $x \in \mathbb{Z}_p w + p\Lambda$.

(*c*) Dado $x \in \Lambda(\overline{v}_1) \cap \Lambda(\overline{v}_1) \cap \Lambda(\overline{v}_3)$, tenemos que $x \in \Lambda$ y $\langle x, v_i\rangle \equiv 0 \pmod p$ para todo $i$ por el punto anterior. Luego basta ver que $\overline{v}_1, \overline{v}_2$ y $\overline{v}_3$ son linealmente independientes ya que en ese caso $(\mathbb{F}_p\overline{v}_1 + \mathbb{F}_p\overline{v}_2 + \mathbb{F}_p\overline{v}_3)^\perp = 0$ y entonces $x \in p\Lambda$. Como $Q$ es no degenerada en $\Lambda/p\Lambda$, no puede haber un subespacio totalmente isotrópico de dimensión 2. Se sigue que $\langle\overline{v}_1, \overline{v_2}\rangle \neq \overline{0}$ en $\mathbb{F}_p$ y luego es fácil ver que $Q(\overline{z}) \neq \overline{0}$ para todo $z \in \mathbb{F}_p\overline{v}_1 + \mathbb{F}_p\overline{v}_2$ que no es proporcional a $\overline{v}_1$ ni $\overline{v}_2$. En particular $\overline{v}_3 \notin \mathbb{F}_p\overline{v}_1 + \mathbb{F}_p\overline{v}_2$, así que los tres vectores son linealmente independientes. □

*Demostración de* 3.29. Sea $R(L) = \{\Lambda \in \mathcal{T}_Q : L \subseteq \Lambda\}$, de modo que $N(L) = |R(L)|$. Notemos que $R(L)$ forma un subgrafo conexo de $\mathcal{T}_Q$. Esto se debe a que por el lema 3.32, para cada par de elementos $\Lambda, \Gamma \in R(L)$ es posible hallar un camino en el grafo $\mathcal{T}_Q$ que conecta a $\Lambda$ y $\Gamma$ y en el que todos los vértices contienen a $\Lambda \cap \Gamma$. En particular todos los vértices contienen a $L$ y luego el camino está en $R(L)$.

Podemos suponer que $N(L) > 0$ y elegimos un elemento $\Lambda \in R(L)$. Como $Q|_L$ tiene invariantes $(a, b)$, existe una base $e_1, e_2$ de $L$ como $\mathbb{Z}_p$-módulo tal que

$$Q(e_1) = u_1 p^a, \qquad Q(e_2) = u_2 p^b, \qquad \langle e_1, e_2\rangle = 0, \tag{3.27}$$

donde $u_1, u_2 \in \mathbb{Z}_p^\times$. Consideramos varios casos:

**Caso 1:** $a = b = 0$. En este caso veamos que $R(L) = \{\Lambda\}$. Si esto no pasa, como $R(L)$ es conexo $\Lambda$ debe tener un vecino de la forma $\Lambda(\overline{v})$ para algún vector isotrópico $\overline{v}$. Entonces $L \subseteq \Lambda \cap \Lambda(\overline{v}) = S_v$. En particular $\overline{e}_1, \overline{e}_2 \in \mathbb{F}_p\overline{v}^\perp$. Como $\overline{e}_1, \overline{e}_2$ son linealmente independientes en $\Lambda/p\Lambda$, generan el subespacio $\mathbb{F}_p\overline{v}^\perp$. En particular $\mathbb{F}_p\overline{v}^\perp$ es un subespacio regular por (3.27). Pero por la proposición 3.2 esto contradice el hecho que $\mathbb{F}_p\overline{v}$ no es regular. Por lo tanto $\Lambda$ es el único elemento de $R(L)$.

**Caso 2:** $a = 0$, $b \geq 1$. Supongamos que $N(L) > 1$. Notemos que $\overline{e}_2 \in \mathbb{F}_p\overline{e}_1^\perp$ es un vector isotrópico y luego $\mathbb{F}_p\overline{e}_1^\perp$ es un plano hiperbólico. Entonces, por el lema de Hensel multivariable, es fácil ver que existen $w_1, w_2 \in \Lambda$ tales que

$$\langle e_1, w_1\rangle = \langle e_1, w_2\rangle = 0,$$

$$Q(w_1) = Q(w_2) = 0, \qquad \langle w_1, w_2\rangle = u_3 \in \mathbb{Z}_p^\times$$

y tales que $\{e_1, w_1, w_2\}$ es una base de $\Lambda$ como $\mathbb{Z}_p$-módulo.

En particular, $\overline{w}_1$ y $\overline{w}_2$ son vectores isotrópicos y determinan retículos $\Lambda(\overline{w}_1), \Lambda(\overline{w}_2)$ que contienen a $e_1$. Es fácil ver que $S_{w_1} = \mathbb{Z}_p w_1 + \mathbb{Z}_p e_1 + \mathbb{Z}_p p w_2$, luego

$$\Lambda(\overline{w}_1) = \mathbb{Z}_p \frac{w_1}{p} \oplus \mathbb{Z}_p e_1 \oplus \mathbb{Z}_p p w_2.$$

Análogamente, $\Lambda(\overline{w}_2) = \mathbb{Z}_p p w_1 \oplus \mathbb{Z}_p e_1 \oplus \mathbb{Z}_p \frac{w_2}{p}$. Notemos que no puede haber ninguna otro retículo $\Lambda(\overline{v})$ adyacente a $\Lambda$ que contenga a $e_1$ ya que por el lema 3.36.(*c*),

$$\Lambda(\overline{w}_1) \cap \Lambda(\overline{w}_2) \cap \Lambda(\overline{v}) = p\Lambda$$

y $e_1 \notin p\Lambda$ puesto que $p \nmid Q(e_1)$. El mismo argumento vale para $\Lambda(\overline{w}_1)$ y $\Lambda(\overline{w}_2)$, luego por inducción vemos que todas los retículos en $R(L)$ deben ser de la forma

$$\Lambda_n := p^{-n}\mathbb{Z}_p w_1 \oplus \mathbb{Z}_p e_1 \oplus p^n \mathbb{Z}_p w_2,$$

para algún $n \in \mathbb{Z}$. Además $\Lambda_m$ y $\Lambda_n$ son adyacentes si y solo si $|m-n| = 1$. Veamos que si $|n| > b/2$ entonces $e_2 \notin \Lambda_{2n}$. Como $R(L)$ es conexo esto implica que $e_2 \notin \Lambda_m$ para los enteros $m$ con $|m| \geq b+2$. De esto se deduce que $N(L) \leq 2b+3$. Por simetría podemos suponer que $n > 0$. Supongamos que $e_2 \in \Lambda_{2n}$ para algún $n > b/2$. Entonces

$$e_2 \in \Lambda \cap \Lambda_{2n} = \mathbb{Z}_p w_1 \oplus \mathbb{Z}_p e_1 \oplus p^{2n}\mathbb{Z}_p w_2 = \mathbb{Z}_p e_1 + p^n \Lambda_n$$

así que podemos escribir $e_2 = \lambda e_1 + p^n z$ para algún $\lambda \in \mathbb{Z}_p$ y $z \in \Lambda_n$. Como

$$0 = \langle e_1, e_2 \rangle \equiv \lambda \langle e_1, e_1 \rangle \equiv 2\lambda u_1 \pmod{p^n},$$

tenemos que $p^n \mid \lambda$ y luego

$$u_2 p^b = Q(e_2) = \lambda^2 Q(e_1) + p^{2n} Q(z) + \lambda p^n \langle e_1, z \rangle \equiv 0 \pmod{p^{2n}}.$$

Esto es una contradicción ya que $2n > b$.

**Caso 3:** $a = 1$. Supongamos que $\Lambda$ tiene algún vecino $\Lambda(\overline{v})$ en $R(L)$. Al igual que en el caso 1 vemos que $\overline{e}_1 \in \mathbb{F}_p \overline{v}^{\perp}$. Como $v_p(Q(e_1)) = 1$, el elemento $\overline{e}_1 \in \Lambda/p\Lambda$ es no nulo e isotrópico. Entonces $\overline{e}_1$ y $\overline{v}$ no pueden ser linealmente independientes ya que en ese caso el subespacio $\mathbb{F}_p \overline{v} + \mathbb{F}_p \overline{e}_1$ sería totalmente isotrópico de dimensión 2. Luego $\overline{v}$ es un múltiplo de $\overline{e}_1$. En particular solo puede haber un vecino de $\Lambda$. Usando el mismo argumento para $\Lambda(\overline{v})$, vemos que $\Lambda(\overline{v})$ solo tiene un vecino, que debe ser $\Lambda$. En conclusión $N(L) \leq 2$.

**Caso 4:** $a \geq 2$. Afirmamos que existe un $\mathbb{Z}_p$-retículo $\Gamma \in \mathcal{T}_Q$ con $d(\Gamma, \Lambda) \leq \lfloor a/2 \rfloor$ que contiene un submódulo de la forma

$$L_r^{(1)} = \mathbb{Z}_p\, p^{-r} e_1 \oplus \mathbb{Z}_p\, p^{-\lfloor b/2 \rfloor} e_2, \qquad 0 \leq r \leq \lfloor a/2 \rfloor \tag{3.28}$$

o de la forma

$$L_s^{(2)} = \mathbb{Z}_p\, p^{-\lfloor a/2 \rfloor} e_1 \oplus \mathbb{Z}_p\, p^{-s} e_2, \qquad 0 \leq s \leq \lfloor b/2 \rfloor. \tag{3.29}$$

Para esto procedemos por inducción en $a+b$, siendo los casos $a+b \in \{0,1\}$ triviales. Suponiendo $a+b \geq 2$, podemos tomar $k, \ell \geq 0$ tales que los vectores $e_1' = p^{-k} e_1$ y $e_2' = p^{-\ell} e_2$ son primitivos. El submódulo $L' = \mathbb{Z}_p e_1' \oplus \mathbb{Z}_p e_2'$ tiene invariantes $a-2k$ y $b-2\ell$, donde el orden depende de cuál de estos números es menor. Si $p \nmid Q(e_1')$ o $p \nmid Q(e_2')$ entonces $L'$ debe ser de la forma $L_x^{(i)}$ y la afirmación vale en este caso. Si no, tanto $\overline{e_1'}$ como $\overline{e_2'}$ son vectores isotrópicos. Luego deben ser linealmente dependientes, en caso caso contrario $Q$ se anularía en el plano $\mathbb{F}_p \overline{e_1'} + \mathbb{F}_p \overline{e_2'}$. Se sigue que $\Lambda' = \Lambda\left(\overline{e_1'}\right)$ contiene a $e_1'/p$ y $e_2'/p$, es decir, contiene a $\frac{1}{p}L'$. Ahora, $\frac{1}{p}L'$ tiene invariantes $a-2k-2$ y $b-2\ell-2$ así que por hipótesis inductiva existe un retículo $\Gamma \in \mathcal{T}_Q$ con $d(\Gamma, \Lambda') \leq \lfloor a/2 \rfloor - 1$ que contiene un retículo $L_x^{(i)}$. Como $\Lambda'$ es adyacente a $\Lambda$, tenemos $d(\Gamma, \Lambda) \leq \lfloor a/2 \rfloor$ lo que demuestra la afirmación. En conclusión

$$R(L) \subseteq \bigcup_{r=0}^{\lfloor a/2 \rfloor} \bigcup_{\Gamma \in R(L_r^{(1)})} \overline{B}(\Gamma, \lfloor a/2 \rfloor) \cup \bigcup_{s=0}^{\lfloor b/2 \rfloor} \bigcup_{\Gamma \in R(L_s^{(2)})} \overline{B}(\Gamma, \lfloor a/2 \rfloor), \tag{3.30}$$

donde $\overline{B}(\Gamma, \lfloor a/2 \rfloor)$ denota la bola cerrada en $\mathcal{T}_Q$ de centro $\Gamma$ y radio $\lfloor a/2 \rfloor$, con respecto a la distancia $d$ definida antes. Todo $L_x^{(i)}$ tiene invariantes $(0, b')$ o $(1, b')$ para algún $b' \leq b$, así que $|R(L_x^{(i)})| \ll b+1$ por los casos anteriores. Además $|\overline{B}(\Gamma, \lfloor a/2 \rfloor)| \ll p^{\lfloor a/2 \rfloor}$ para cualquier retículo $\Gamma$, así que

$$|N(L)| \ll (\lfloor a/2 \rfloor + \lfloor b/2 \rfloor)\,(b+1)p^{\lfloor a/2 \rfloor} \leq (b+1)^2 p^{\lfloor a/2 \rfloor}. \qquad \square$$

*Demostración de* 3.30. Sean $f_1 = (1,0,0)$, $f_2 = (0,1,0)$ y $f_3 = (1,-1,1)$. Entonces $f_1, f_2, f_3 \in \mathbb{Z}_2^3$ son vectores isotrópicos primitivos y luego determinan los vecinos de $\mathbb{Z}_2^3$. Además el elemento $\overline{k} = (0,0,1) \in \mathbb{Z}_2^3/2\mathbb{Z}_2^3$ cumple que $\langle \overline{k}, \overline{v} \rangle = 0$ en $\mathbb{F}_2$, para todo $v \in \mathbb{Z}_2^3$ y es el único con esta propiedad. Lo llamaremos el *elemento especial* de $\mathbb{Z}_2^3/2\mathbb{Z}_2^3$. Para cualquier retículo $\Lambda \in \mathcal{T}_Q$, el espacio cuadrático $(\Lambda/2\Lambda, Q)$ es isomorfo a $(\mathbb{Z}_2^3/2\mathbb{Z}_2^3, Q)$, y en particular contiene exactamente tres vectores isotrópicos y un elemento especial, que también denotamos $\overline{k}$.

Supongamos primero que $Q$ tiene invariantes $(a,b)$ de tipo diagonal, es decir, existe una base $\{e_1, e_2\}$ de $L$ tal que

$$Q(e_1) = u_1 2^a, \qquad Q(e_2) = u_2 2^b, \qquad \langle e_1, e_2 \rangle = 0.$$

**Caso 1:** $a = b = 0$. Afirmamos que en este caso $\Lambda$ tiene a lo sumo un vecino en $R(L)$. Supongamos que $\Lambda$ tiene dos vecinos $\Lambda(\overline{v})$ y $\Lambda(\overline{w})$. Entonces por el lema 3.36.(*b*) tenemos

$$L \subseteq \Lambda(\overline{v}) \cap \Lambda(\overline{w}) = \mathbb{Z}_2 k + 2\Lambda$$

ya que $(\mathbb{F}_p\overline{v} + \mathbb{F}_p\overline{w})^\perp = \mathbb{F}_p k$. En particular $\overline{e}_1, \overline{e}_2 \in \mathbb{F}_p\overline{k}$. Como $e_1$ y $e_2$ son primitivos, se sigue que $e_1 \equiv k \equiv e_2 \pmod{2\Lambda}$. Entonces podemos escribir $e_2 = e_1 + 2z$ para algún $z \in \Lambda$. Pero entonces

$$\langle e_1, e_2 \rangle = \langle e_1, e_1 + 2z \rangle = 2u_1 + 2\langle e_1, z \rangle$$

y notemos que $\langle e_1, z \rangle \equiv \langle k, z \rangle \equiv 0 \pmod 2$, mientras que $u_1$ es impar. Luego $\langle e_1, e_2 \rangle \neq 0$, absurdo. En conclusión, $\Lambda$ tiene a lo sumo un vecino. Como esto vale para cualquier elemento de $R(L)$ concluimos que $N(L) \leq 2$.

**Caso 2:** $a = 0$, $b \geq 1$. En este caso veremos que existe un camino $\gamma$ de longitud finita en $R(L)$ tal que todos los elementos de $R(L)$ están a distancia 1 de algún elemento de $\gamma$. Llamamos a un elemento $\Gamma \in R(L)$ *central* si $\overline{e}_1 = \overline{k}$ en $\Gamma$, donde $\overline{k}$ denota el elemento especial, y sea $C$ el conjunto de vértices centrales.

En primer lugar afirmamos que si $\Gamma$ no es central entonces $\Gamma$ tiene a lo sumo un vecino en $R(L)$. En efecto, si $\Gamma$ tiene dos vecinos $\Gamma(\overline{v})$ y $\Gamma(\overline{w})$, argumentando como en el caso 1 obtenemos que $\overline{e}_1$ debe ser igual a $\overline{k}$, en cuyo caso $\Gamma$ es central. Esto implica que $C$ es conexo, puesto que dados $\Gamma_1, \Gamma_2 \in C$, existe un camino en $R(L)$ que los conecta. Los elementos intermedios de tal camino tienen al menos dos vértices adyacentes en $R(L)$ y luego deben ser centrales, lo que demuestra que el camino está contenido en $C$.

Finalmente veamos que cada elemento de $C$ tiene a lo sumo dos vecinos en $C$. Como todo vértice en $\mathcal{T}_Q$ tiene exactamente 3 vértices adyacentes, basta ver que todo retículo central posee un vecino que no es central. Sea $\Gamma \in C$. Podemos suponer sin pérdida de generalidad que $\Gamma = \mathbb{Z}_2^3$ y que todos los vecinos de $\Gamma$ pertenecen a $R(L)$. Supongamos que $e_1 = (\alpha, \beta, \gamma)$ y consideremos los vectores $f_1, f_2, f_3$ definidos al principio de la demostración. En este caso tenemos $\overline{e}_1 = \overline{k} = (0,0,1)$ así que $\alpha \equiv \beta \equiv 0 \pmod 2$ y $\gamma \equiv 1 \pmod 2$. Si $4 \nmid \alpha$, entonces $\langle e_1, \frac{1}{2} f_2 \rangle = \alpha/2 \equiv 1 \pmod 2$. Esto quiere decir que en el retículo $\Gamma(\overline{f}_2)$ el elemento $e_1$ no es especial. Luego $\Gamma(\overline{f}_2) \notin C$. Análogamente, si $4 \nmid \beta$ entonces $\langle e_1, \frac{1}{2} f_1 \rangle \equiv 1 \pmod 2$ lo que implica que $e_1$ no es especial en $\Gamma(\overline{f}_1)$. Finalmente, si $\alpha \equiv \beta \equiv 0 \pmod 4$ entonces

$$\langle e_1, \tfrac{1}{2} f_3 \rangle = \frac{\beta - \alpha}{2} + \gamma \equiv 1 \pmod 2$$

y luego $e_1$ no es especial en $\Gamma(\overline{f}_3)$. Esto demuestra la afirmación. De estas propiedades se deduce que $C$ es un segmento o un camino cerrado. Argumentando como en el caso 2 del lema 3.29 deducimos que $|C| = O(b+1)$ y luego $N(L) = O(b+1)$.

**Caso 3:** $a = 1$. Al igual que en el caso 3 del lema 3.29, vamos a probar que todo $\Lambda \in R(L)$ tiene a lo sumo un vértice adyacente. Sin embargo ahora no podemos usar que $Q$ es no degenerada.

Como $a = 1$, el vector $e_1 \in \Lambda$ es primitivo y $Q(e_1) \equiv 0 \pmod 2$. Supongamos que $\Lambda(\overline{v}) \in R(L)$ es un vecino de $\Lambda$. Entonces $\overline{e}_1 \in \mathbb{F}_2\overline{v}^\perp = \mathbb{F}_2\overline{v} + \mathbb{F}_2\overline{k}$. Además $\overline{e}_1 \neq 0$ ya que $e_1$ es primitivo. Si $\overline{e}_1 \neq \overline{v}$, entonces $\overline{e}_1 = \overline{k}$ o $\overline{e}_1 = \overline{v} + \overline{k}$. En el primer caso obtenemos $Q(e_1) \equiv Q(k) \equiv 1 \pmod 2$ y en el segundo caso

$$Q(e_1) \equiv Q(v + k) = Q(v) + Q(k) + \langle v, k\rangle \equiv 1 \pmod 2.$$

En ambos casos llegamos a una contradicción, luego debe ser $\overline{e}_1 = \overline{v}$. En particular, solo hay una posibilidad para $\overline{v}$. Por lo tanto $N(L) \leq 2$ en este caso.

**Caso 4:** $a \geq 2$. Este caso se demuestra de la misma manera que en el lema 3.29.

Ahora pasemos al caso no diagonal, es decir, supongamos que $Q|_L$ es de la forma (3.17) o (3.18) para algún $a \geq 0$. Equivalentemente, existe una base $e_1, e_2$ de $L$ tal que $\langle e_1, e_2\rangle = 2^a$ y

$$Q(e_1) = Q(e_2) = \varepsilon_0 2^a,$$

donde $\varepsilon_0 \in \{0, 1\}$.

**Caso 5:** $a = 0$. En este caso $\overline{e}_1$ y $\overline{e}_2$ deben ser linealmente independientes, ya que en caso contrario $\langle e_1, e_2\rangle \equiv 0 \pmod 2$. Supongamos primero que $\varepsilon_0 = 1$. Como $Q(e_1 + e_2) \equiv 1 \pmod 2$, vemos que no hay vectores isotrópicos en $\mathbb{F}_2\overline{e}_1 + \mathbb{F}_2\overline{e}_2$. Se sigue que no hay ningún vector isotrópico $\overline{v}$ que sea ortogonal a ambos, ya que entonces $\mathbb{F}_2\overline{v}^\perp$ contendría a $\overline{v}, \overline{e}_1, \overline{e}_2$ que son linealmente independientes. En particular $\Lambda$ no tiene ningún vecino y $N(L) = 1$. Supongamos ahora que $\varepsilon_0 = 0$. Entonces $\overline{e}_1$ y $\overline{e}_2$ son vectores isotrópicos linealmente independientes. Si $\Lambda(\overline{v})$ es un vecino de $\Lambda$, entonces $\overline{e}_1 \in \mathbb{F}_2\overline{v}^\perp = \mathbb{F}_2\overline{v} + \mathbb{F}_2\overline{k}$ y al igual que en el caso 3 obtenemos $\overline{v} = \overline{e}_1$. Análogamente, $\overline{v} = \overline{e}_2$ y esto contradice la independencia lineal de $\overline{e}_1, \overline{e}_2$. En conclusión $\Lambda$ no puede tener vecinos y $N(L) = 1$.

**Caso 6:** $a = 1$. Notemos que uno de los vectores $e_1, e_2$ debe ser primitivo, ya que en caso contrario $4 \,|\, \langle e_1, e_2\rangle$. Entonces podemos suponer que $e_1$ es primitivo. En tal caso $\overline{e}_1$ es un vector isotrópico no nulo. Al igual que en casos anteriores, vemos que $e_1 \in \Lambda(\overline{v})$ solo si $\overline{v} = \overline{e}_1$. Se sigue que todo vértice en $R(L)$ tiene a lo sumo un vecino y luego $N(L) \leq 2$.

**Caso 7:** En el caso $a \geq 2$, la demostración es similar al caso final de 3.29. Al igual que en ese caso, hay que probar por inducción que

$$R(L) \subseteq \bigcup_{L' \in \mathfrak{S}} \bigcup_{\Gamma \in R(L')} \overline{B}(\Gamma, \lfloor a/2 \rfloor),$$

donde $\mathfrak{S}$ es un conjunto con $O(a + 1)$ elementos. Más precisamente, si definimos

$$L_{r,s} = \mathbb{Z}_2 2^{-r} e_1 \oplus \mathbb{Z}_2 2^{-s} e_2$$

entonces

$$\mathfrak{S} = \{L_{r,s} : r, s \geq 0, r + s = a\}$$

en el caso $\varepsilon_0 = 0$ y

$$\mathfrak{S} = \{L_{r,s} : r, s \geq 0, \text{máx}(r, s) = \lfloor a/2 \rfloor\}$$

en el caso $\varepsilon_0 = 1$. En el caso $\varepsilon_0 = 0$, $Q$ es de la forma (3.17) en todos los $L_{r,s} \in \mathfrak{S}$, con $a \leq 1$. En el caso $\varepsilon_0 = 1$ se puede ver que o bien $Q$ tiene invariantes $(a', b')$ de tipo diagonal con $a' \leq 1$ y $b' \leq a$ en todos los elementos de $\mathfrak{S}$, (hay que elegir una base distinta), o bien es de la forma (3.18) con $a \leq 1$. Luego en todos los casos hay $O(a + 1)$ retículos en $R(L')$, para todo $L' \in \mathfrak{S}$. Luego $N(L) \ll (a + 1)^2 2^{\lfloor a/2 \rfloor}$. □

# Capítulo 4

# Demostración del teorema de Duke

En este capítulo retornamos al estudio de la curva modular $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$ para completar la demostración del teorema de Duke. Los dos ingredientes más importantes de la demostración son el lema básico de Linnik, que demostraremos en la primera sección usando el teorema 3.26, y el teorema 4.6.

## 4.1. Resultados preliminares

En esta sección demostramos el lema básico de Linnik, así como un par de resultados auxiliares que necesitaremos más adelante. Sea $G = \mathrm{PSL}_2(\mathbb{R})$. Fijamos una métrica riemanniana $\langle\cdot,\cdot\rangle$ invariante a izquierda en $G$, que induce una distancia $d = d_G : G \times G \to \mathbb{R}$ invariante a izquierda. Como vimos en la introducción del capítulo 2, para demostrar el teorema de Duke basta probar el teorema 2.1 que trata sobre la convergencia débil* de la sucesión de medidas $(\mu_d)_d$ definidas en esa misma sección a la medida de Haar $\mu_X$.

Definimos una norma en $\mathrm{PSL}_2(\mathbb{R})$: si $g \in \mathrm{PSL}_2(\mathbb{R})$ está representado por una matriz $a$, entonces

$$\|g\| = \mathrm{tr}(a^t a)^{1/2}.$$

Esta norma está bien definida ya que $(-a)^t(-a) = a^t a$ para cualquier matriz $a$.

**Lema 4.1.** *Sea $g$ un elemento del conjunto*

$$S = \{h \in \mathrm{PSL}_2(\mathbb{R}) : h(i) \in F\},$$

*donde $F \subseteq \mathbb{H}$ es el dominio fundamental de la acción de $G$. Supongamos además que $\mathrm{ht}(\Gamma g) \leq H$ para algún $H > 1$. Entonces $\|g\| \ll H$.*

*Demostración.* Supongamos que $g$ está representado por una matriz $\left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right)$. Como $g \in S$, tenemos

$$\mathrm{ht}(\Gamma g) = \left(\mathrm{Im}\,\frac{pi+q}{ri+s}\right)^{1/2} = (r^2+s^2)^{-1/2} \in \left[\tfrac{1}{2}, H\right],$$

de modo que $H^{-2} \leq r^2 + s^2 \leq 4$. Por otro lado,

$$|\mathrm{Re}\, g(i)| = \left|\frac{qs+pr}{r^2+s^2}\right| \leq \frac{1}{2},$$

así que

$$|qs+pr| \leq \tfrac{1}{2}(r^2+s^2) \leq 2.$$

Además $ps - qr = \pm 1$ por ser el determinante de la matriz, y luego

$$(p^2+q^2)(r^2+s^2) = (ps-qr)^2 + (qs+pr)^2 \leq 1 + 4 = 5,$$

lo que implica que

$$p^2 + q^2 \leq 5(r^2 + s^2)^{-1} \leq 5H^2.$$

Juntando estas cotas obtenemos que

$$\|g\|^2 = p^2 + q^2 + r^2 + s^2 \leq 5H^2 + 4 \ll H^2. \qquad \square$$

**Lema 4.2.** *Para todo compacto $K \subseteq \mathrm{PSL}_2(\mathbb{R})$ existen constantes $c_1, c_2 > 0$ que dependen solamente de $K$, tales que*

$$c_1\|g - 1\| \leq d(g, 1) \leq c_2\|g - 1\|,$$

*para todo $g \in K$.*

*Demostración.* Para todo $r > 0$ el conjunto $K \setminus B_r(1)$ es compacto, así que la función $F(g) = \frac{d(g,1)}{\|g-1\|}$ atiene un máximo y un mínimo en este espacio. Entonces para probar el lema basta analizar cómo se comporta $d(g, 1)$ en un entorno de la identidad. Como la proyección $\pi : \mathrm{SL}_2(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R})$ es un revestimiento, existe un entorno $U$ de $I \in \mathrm{SL}_2(\mathbb{R})$ tal que $\pi|_U$ es un homeomorfismo en su imagen. Además podemos darle a $\mathrm{SL}_2(\mathbb{R})$ una métrica riemanniana de modo que $\pi$ sea una isometría local. Luego podemos suponer que estamos trabajando en un entorno de $I$ en $\mathrm{SL}_2(\mathbb{R})$.

Para todo $g \in \mathrm{SL}_2(\mathbb{R})$ hay una inmersión canónica $T_g\mathrm{SL}_2(\mathbb{R}) \to T_gM_2(\mathbb{R}) \cong M_2(\mathbb{R})$, así que podemos mirar a los vectores tangentes como matrices. Sea $\|\cdot\|_{M_2(\mathbb{R})}$ la norma $\ell^2$ usual en $M_2(\mathbb{R})$. Sea $\exp_I : \mathscr{D} \subseteq T_I\mathrm{SL}_2(\mathbb{R}) \to \mathrm{SL}_2(\mathbb{R})$ el mapa exponencial en el sentido de geometría riemanniana. Este mapa es un difeomorfismo en un entorno del 0 ya que $d_0 \exp_I : T_I\mathrm{SL}_2(\mathbb{R}) \to T_I\mathrm{SL}_2(\mathbb{R})$ es la identidad. En particular $\exp_I$ es bi-Lipschitz en un entorno del 0, es decir, existe un entorno $U \subseteq T_I\mathrm{SL}_2(\mathbb{R})$ del 0 y constantes $c_{11}, c_{21} > 0$ tales que

$$c_{11}\|v - w\|_{M_2(\mathbb{R})} \leq \|\exp_I(v) - \exp_I(w)\|_{M_2(\mathbb{R})} \leq c_{21}\|v - w\|_{M_2(\mathbb{R})} \tag{4.1}$$

para todo par de puntos $v, w \in U$. A su vez existen $c_{12}, c_{22} > 0$ tales que

$$c_{12}\|v - w\| \leq \|v - w\|_{M_2(\mathbb{R})} \leq c_{22}\|v - w\|,$$

para $v, w \in T_I\mathrm{SL}_2(\mathbb{R})$. Ahora, si $g = \exp_I(v)$ para algún $v \in U$, entonces $\|v\|$ es la longitud de la geodésica entre 1 y $g$, y luego coincide con $d(g, 1)$ si $\|v\|$ es chico. Tomando $w = 0$ en (4.1) obtenemos que

$$c_{11}c_{12}\|v\| \leq \|g - I\|_{M_2(\mathbb{R})} \leq c_{21}c_{22}\|v\|,$$

lo cual demuestra el lema ya que $d(g, 1) = \|v\|$. $\square$

En la demostración del lema 4.4 usaremos una versión de la fórmula del número de clase para órdenes en cuerpos cuadráticos:

**Teorema 4.3** (Fórmula del número de clase)**.** *Sea $d \geq 2$, $d \equiv 0, 1 \pmod 4$ un entero. Entonces*

$$h(d) \log \varepsilon_d = \sqrt{d}L(1, \chi), \tag{4.2}$$

*donde $\varepsilon_d$ es la unidad fundamental de $\mathcal{O}_d$, $h(d) = |\mathrm{Pic}(\mathcal{O}_d)|$ es el número de clase, $\chi(n) = \left(\frac{d}{n}\right)$ es el símbolo de Kronecker y*

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}.$$

Que $L(1, \chi)$ converge se deduce del criterio de convergencia de Dirichlet, ya que la sucesión $(\chi(n))_{n\in\mathbb{N}}$ tiene sumas parciales acotadas y $\frac{1}{n}$ tiende a 0 de forma monótona.

*Demostración.* Ver [Lan86, Teo. 209]. Alternativamente, se puede deducir de la fórmula del número de clase usual. $\square$

El lema de Siegel [Dav82, cap. 21] nos dice que $L\left(1, \left(\frac{d}{\cdot}\right)\right) \gg_\varepsilon d^{-\varepsilon}$ para todo $\varepsilon > 0$, aunque la cota no es explícita, es decir, no se conoce ninguna constante $C(\varepsilon) > 0$ calculable explícitamente que haga valer la desigualdad. Por la fórmula del número de clase para órdenes, el lema de Siegel implica que $h(d) \log \varepsilon_d = \sqrt{d}L(1, \chi) \gg_\varepsilon d^{1/2-\varepsilon}$.

**Lema 4.4.** *Para todo $\varepsilon > 0$ y $H \geq 1$, se tiene*

$$\mu_d(X_{\geq H}) \ll_\varepsilon d^\varepsilon H^{-2}. \tag{4.3}$$

*Demostración.* Recordemos que la medida $\mu_d$ se define como $\mu_d = \frac{1}{\rho(\mathcal{G}_d)}\rho_d$, donde $\rho_d$ es una suma de medidas $A$-invariantes soportadas en las geodésicas $x_q A$ con $[q] \in C(d)$. Por la observación 3.23 tenemos que $\mathcal{G}_d \subseteq X_{\leq d^{1/4}}$. Entonces las componentes conexas de $\mathcal{G}_d \cap X_{\geq H}$ son segmentos geodésicos con altura entre $H$ y $d^{1/4}$. Por el corolario 2.13 se sigue que la longitud de estas curvas, y luego su medida con respecto a $\rho_d$, está acotada por $\ll \log(d)$. Por la proposición 3.22,

$$\rho_d(\mathcal{G}_d \cap X_{\geq H}) \ll \log(d) N_{\leq H}(d),$$

donde $N_{\leq H}(d)$ denota la cantidad de ideales inversibles $I \subseteq \mathcal{O}_d$ con $N(I) \leq \frac{1}{2}H^{-2}\sqrt{d}$. Recordemos que por 3.20, $R_d(n) \ll_\varepsilon n^\varepsilon$, donde $R_d(n)$ es la cantidad de ideales inversibles de norma $n$. Sumando sobre $1 \leq n \leq \frac{1}{2}H^{-2}\sqrt{d}$ obtenemos

$$N_{\leq H}(d) \ll_\varepsilon (H^{-2}\sqrt{d})^{1+\varepsilon} \ll_\varepsilon H^{-2} d^{(1+\varepsilon)/2}.$$

Luego

$$\rho_d(\mathcal{G}_d \cap X_{\geq H}) \ll_\varepsilon \log(d) d^{(1+\varepsilon)/2} H^{-2} \ll_\varepsilon d^{1/2+\varepsilon} H^{-2}.$$

Por otro lado, $\rho_d(\mathcal{G}_d) = h(d)\log \varepsilon_d \gg_\varepsilon d^{1/2-\varepsilon}$ por el lema de Siegel así que

$$\mu_d(X_{\geq H}) = \frac{\rho_d(\mathcal{G}_d \cap X_{\geq H})}{\rho_d(\mathcal{G}_d)} \ll_\varepsilon d^{2\varepsilon} H^{-2}. \qquad \square$$

**Lema 4.5** (Lema Básico de Linnik)**.** *Sea $d \in \mathbb{N}$ y sea $H \geq 1$. Entonces*

$$\mu_d \times \mu_d\{(x,y) \in X_{\leq H}^2 : d(x,y) < \delta\} \ll_\varepsilon H^4 \delta^3 d^\varepsilon \tag{4.4}$$

*siempre que $d^{-1/4} \leq \delta \leq \frac{1}{3}H^{-2}$ y $\varepsilon > 0$.*

*Demostración.* Hay una relación entre pares de puntos $(x_1, x_2) \in (\mathcal{G}_d \cap X_{\leq H})^2$ a distancia menor que $\delta$ y representaciones de la forma $q(x,y) = dx^2 + \ell xy + dy^2$.

Sea $D_\delta = \{(x,y) \in (\mathcal{G}_d \cap X_{\leq H})^2 : d(x,y) < \delta\}$ el conjunto que aparece en (4.4) y sea $S' \subseteq \mathbb{H}$ un entorno abierto del dominio fundamental $S$. Si $S'$ es suficientemente chico, entonces dados $x_1, x_2 \in \mathcal{G}_d \cap X_{\leq H}$, existe un único par $(g_1, g_2) \in S \times S'$ tal que $\Gamma g_1 = x_1$, $\Gamma g_2 = x_2$ y $d(g_1, g_2) < \delta$. Luego la imagen del conjunto

$$C_\delta = \{(g_1, g_2) \in S \times S' : \Gamma g_1, \Gamma g_2 \in \mathcal{G}_d \cap X_{\leq H}, d(g_1, g_2) < \delta\}$$

en el cociente $X = \Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$ contiene a $D_\delta$, y tras sacar un subconjunto $Z \subseteq C_\delta$ de medida 0, $C_\delta$ se aplica inyectivamente en $X$. Se sigue que

$$\mu_d \times \mu_d(D_\delta) \leq \widetilde{\mu}_d \times \widetilde{\mu}_d(C_\delta),$$

donde $\widetilde{\mu}_d$ es la medida en $\mathrm{PSL}_2(\mathbb{R})$ asociada a $\mu_d$ bajo la correspondencia de la proposición 1.20. Esta medida es simplemente una suma de las medidas de Haar soportadas en las $A$-órbitas $g_q A$, con $q \in R_{\mathrm{disc}}(d)$.

Ahora, podemos dividir el conjunto $C_\delta$ según las formas cuadráticas $q_1, q_2$ asociadas a cada punto $(g_1, g_2)$. Sean $\Gamma(q_1^{(j)}, q_2^{(j)})$, $1 \leq j \leq k$ las $\Gamma$-órbitas de pares de formas cuadráticas asociadas a puntos en $C_\delta$. Para cada $1 \leq j \leq k$ elegimos un par $(g_1^{(j)}, g_2^{(j)}) \in C_\delta$ asociado a $(q_1^{(j)}, q_2^{(j)})$. A cualquier par $(q_1, q_2) \in R_{\mathrm{disc}}(d)^2$ le podemos asociar la forma cuadrática binaria

$$\widehat{q}_\ell(u,v) = \mathrm{disc}(uq_1 + vq_2) = du^2 + \ell uv + dv^2,$$

donde $\ell$ es un entero. Entonces el par $(q_1, q_2)$ da lugar a una representación de $\widehat{q}_\ell$ por la forma cuadrática ternaria disc. Ahora, como $g_1$ y $g_2$ están cerca entre sí, $q_1$ y $q_2$ deben estar cerca y

luego $\ell$ no puede ser muy grande. Efectivamente, sea $M_0 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ la matriz asociada a $q_0$ y sea $h = g_2 g_1^{-1}$. Entonces $hM_0h^t$ es la matriz asociada a $h.q_0$. Como

$$hM_0h^t - M_0 = hM_0(h^t - 1) + (h-1)M_0,$$

tenemos que

$$\|hM_0h^t - M_0\| \leq \|hM_0\| \cdot \|h^t - 1\| + \|h-1\| \cdot \|M_0\| \ll \|h-1\| \ll d(h,1),$$

donde usamos el lema 4.2 en la cota final. Luego $\|h \cdot q_0 - q_0\| \ll \delta$, donde $\|\cdot\|$ denota la norma $\ell^2$ usual en $\mathbb{R}^3$. Además tenemos que $\|g_1\| \ll H$, por el lema 4.1. Luego

$$\|q_2 - q_1\| \ll \sqrt{d}\|g_1\|^2 \|h.q_0 - q_0\| \ll \sqrt{d}H^2\delta. \tag{4.5}$$

Pero entonces

$$|2d - \ell| = |\widehat{q}(-1,1)| = \operatorname{disc}(q_2 - q_1) \ll \|q_2 - q_1\|^2 \ll dH^4\delta^2. \tag{4.6}$$

Notar que no hemos usado todavía las condiciones $d^{-1/4} \leq \delta \leq \frac{1}{3}H^{-2}$.

Ahora, la forma $\widehat{q}_\ell$ no cambia si multiplicamos a $q_1$ y $q_2$ por un mismo $\gamma \in \Gamma$. Se sigue que podemos asociarle a cada $\Gamma$-órbita $\Gamma(q_1^{(j)}, q_2^{(j)})$ una forma $\widehat{q}_\ell$. Sea

$$N_{\ell,d} = |\Gamma \backslash \operatorname{Emb}(\widehat{q}_\ell, \operatorname{disc})|.$$

Notemos que la acción de $\Gamma$ en $R_{\operatorname{disc}}(d)$ se extiende a una acción en $\mathbb{Z}^3$ que es lineal, es decir, el mapa $(a,b,c) \mapsto g \cdot (a,b,c)$ es $\mathbb{Z}$-lineal para todo $g \in \Gamma$. Luego la acción induce un morfismo de grupos $\phi : \Gamma \to \mathrm{GL}_3(\mathbb{Z})$. Además $\phi(\Gamma) \subseteq \mathrm{SO}_{\operatorname{disc}}(\mathbb{Z})$: efectivamente, si $M_q$ es la matriz asociada a una forma cuadrática representada por $(a,b,c)$, entonces la matriz asociada a $g \cdot (a,b,c)$ es igual a $gM_qg^t$ y luego

$$\operatorname{disc}(g \cdot q) = -\det(gM_qg^t) = -\det(M_q) = \operatorname{disc}(q),$$

de modo que la acción de $g$ preserva el discriminante. Se puede ver que $\phi(\Gamma)$ tiene índice finito en $\mathrm{SO}_{\operatorname{disc}}(\mathbb{Z})$. Entonces, por el teorema 3.26 tenemos que $N_{\ell,d} \ll_\varepsilon f\,\mathrm{máx}(d,\ell)^\varepsilon$. Sea $c_1$ la constante implícita en la desigualdad (4.6) y sea $L = c_1 dH^4\delta^2$. Luego el número $k$ de $\Gamma$-órbitas de pares está acotado por

$$k \leq \sum_{0<|\ell-2d|\leq L} N_{\ell,d} \leq \sum_{f^2|d} \sum_{\substack{f^2|\ell \\ 0<|\ell-2d|\leq L}} N_{\ell,d} \ll_\varepsilon \sum_{f^2|d} \sum_{\substack{f^2|\ell \\ 0<|\ell-2d|\leq L}} fd^\varepsilon$$

$$\ll \sum_{f^2|d} fd^\varepsilon L/f^2 \ll d^{1+\varepsilon}H^4\delta^2 \sum_{f^2|d} \frac{1}{f} \ll_\varepsilon d^{1+2\varepsilon}H^4\delta^2.$$

Ahora veamos que

$$d(g_1^{(j)}a_t, g_2^{(j)}A) \gg d^{-1}$$

para todo $t$ y para todo $1 \leq j \leq k$ tal que $q_1 \neq q_2$. Supongamos que $d(g_1^{(j)}a_t, g_2^{(j)}a_s) \geq d^{-1}$. Podemos elegir $\gamma \in \Gamma$ tal que $\gamma g_1^{(j)}a_t \in S$ y entonces $\gamma g_2^{(j)}a_s \in S'$. Por la observación 3.23, tenemos que $\mathscr{G}_d \subseteq X_{\leq d^{1/4}}$. Tomando $H' = d^{1/4}$, $\delta' = d^{-1}$, obtenemos por (4.5) que $\|q_2 - q_1\| \ll d^{-1/4}$ y luego es menor que 1 para $d > 0$ suficientemente grande. Pero esto se contradice con que los coeficientes de $q_1$ y $q_2$ son enteros.

Se puede ver que todo par $(x_1, x_2)$ asociado a cada par $(g_1^{(j)}, g_2^{(j)})$ en el que las formas cuadráticas asociadas son distintas es de la forma $x_1 = \Gamma g_1^{(j)}\alpha$ con $\alpha$ contenido en un intervalo $I_j \subseteq A$ de longitud $\ll_\varepsilon d^\varepsilon$. Luego la medida del conjunto de pares asociados a $q_1^{(j)} \neq q_2^{(j)}$ es

$$\ll \sum_{j=1}^{k} |I_j|\delta \ll_\varepsilon d^\varepsilon \delta k \ll_\varepsilon d^{1+2\varepsilon}\delta^3 H^4.$$

Finalmente, la $\rho_d \times \rho_d$-masa de los pares asociados a la misma forma es $\ll_\varepsilon d^{1/2+\varepsilon}\delta$. Como $\rho_d(\mathscr{G}_d) \gg_\varepsilon d^{1/2-\varepsilon}$ se sigue que la $\mu_d \times \mu_d$-medida está acotada por $\ll_\varepsilon \delta d^{-1/2+3\varepsilon} \leq \delta^3 d^{3\varepsilon}$, puesto que $\delta \geq d^{-1/4}$. □

**Teorema 4.6.** *Supongamos que $(\mu_i)_i$ es una sucesión de medidas $A$-invariantes en $X$ y supongamos que existe una sucesión $\delta_i \to 0$ de reales positivos tales que las alturas $H_i = \delta_i^{-\varepsilon}$ satisfacen*

(1) $\mu_i(X_{\geq H_i}) \to 0$ *cuando* $i \to \infty$.

(2) $\mu_i \times \mu_i\{(x,y) \in (X_{\leq H_i})^2 : d(x,y) < \delta_i\} \ll_\varepsilon \delta_i^{3-5\varepsilon}$.

*Entonces* $\mu_i \xrightarrow{w*} \mu_X$.

Los lemas 4.4 y 4.5 garantizan que la sucesión $(\mu_i)_i$ cumple las condiciones del teorema anterior tomando $\delta_d = d^{-1/3}$ por ejemplo, así que del teorema 4.6 se deduce el teorema 2.1. La siguiente sección está dedicada a probar el teorema anterior.

## 4.2. Controlando el escape de masa

En esta sección se demuestra el teorema 4.6. La demostración está basada en el siguiente resultado:

**Teorema 4.7.** *Sea $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$ y sea $T : X \to X$ el flujo geodésico a tiempo 1. Entonces $h_\nu(T) \leq 1$ para cualquier medida $\nu \in \mathscr{M}_T^1(X)$, con igualdad si y solo si $\nu = \mu_X$ es la medida de Haar.*

*Demostración.* Ver [EL10, §7]. □

El teorema 4.6 se puede ver como una versión finitaria del teorema 4.7. La idea de la demostración consiste en probar que la entropía de las medidas $\mu_i$ tiende a 1. Este es esencialmente el contenido del lema 4.14. Además necesitaremos ver que la sucesión $(\mu_i)_i$ es tight para evitar el escape de masa; esto se demuestra en el lema 4.12.

Sea

$$B_{N,\eta} = \bigcap_{n=-N}^{N} a^{-n} B_\eta^G a^n,$$

donde $a = \begin{pmatrix} e^{1/2} & 0 \\ 0 & e^{-1/2} \end{pmatrix}$ y $B_\eta^G$ es la bola de radio $\eta$ en $G$. Una *$N$-bola de Bowen* de radio $\eta > 0$ es un conjunto de la forma $xB_{N,\eta}$ con $x \in X$. Si $\eta$ es más chico que el radio de inyectividad en $\{T^n x : n \in [-N,N] \cap \mathbb{Z}\}$ entonces $xB_{N,\eta}$ es igual al conjunto de puntos $y \in X$ tales que $d(T^n x, T^n y) < \eta$ para todo $n \in [-N,N]$. Esta última condición es la definición usual de las bolas de Bowen en espacios topológicos compactos. Las bolas de Bowen están íntimamente relacionadas con la entropía topológica, que es un invariante similar a la entropía medible definida en el capítulo 1. En el caso de un sistema $(X,T)$ donde $X$ es compacto y $T : X \to X$ es un homeomorfismo, estas dos invariantes están relacionadas por el *principio variacional*, que dice que la entropía topológica es el supremo de las entropías medibles $h_\mu(T)$ con $\mu \in \mathscr{M}_T^1(X)$. Luego es razonable que incluso en el caso no compacto, las bolas de Bowen resulten útiles para estudiar la entropía medible.

Dado $x \in X$ y un conjunto $B \subseteq X$ boreliano, definimos el conjunto de recurrencia

$$\mathrm{Rec}(x,B) = \{n \in \mathbb{Z} : T^n x \in B\}. \tag{4.7}$$

**Teorema 4.8.** *Sean $M \geq 1$ y $N \in \mathbb{N}$. Entonces para todo $V \subseteq [0,N-1] \cap \mathbb{Z}$, el conjunto*

$$Z(V) = \{x \in X_{<M} \cap T^{-N} X_{<M} : \mathrm{Rec}(x, X_{\geq M}) \cap [0,N-1] = V\} \tag{4.8}$$

*puede ser cubierto por $\ll_M e^{N - \frac{1}{2}|V|}$ $N$-bolas de Bowen. Más aún, hay $\ll_M \exp\left(\frac{\log\log M}{\log M} N\right)$ subconjuntos $V \subseteq [0,N-1] \cap \mathbb{Z}$ tales que $Z(V) \neq \varnothing$.*

*Demostración.* Este es esencialmente el enunciado del teorema 4.3 en [Ein+14], salvo por el intervalo en el que está contenido $V$ y un cambio acorde en la cota de la cantidad de subconjuntos. La demostración es la misma. □

El siguiente teorema no forma parte de la demostración del teorema de Duke, pero tiene interés propio ya que muestra que una sucesión de medidas con alta entropía no puede presentar escape de masa. Este resultado es usado en [EKP15].

**Teorema 4.9.** *Sea $T$ el flujo geodésico a tiempo 1. Entonces existe $M_0 > 0$ tal que*

$$h_\mu(T) \leq 1 + \frac{\log\log M}{\log M} - \frac{\mu(X_{\geq M})}{2} \tag{4.9}$$

*para toda $\mu \in \mathscr{M}_T^1(X)$ y todo $M \geq M_0$. En particular, si $(\mu_i)_{i\in\mathbb{N}} \subseteq \mathscr{M}_T^1(X)$ es una sucesión con $h_{\mu_i}(T) \geq c$ entonces todo límite débil$^*$ $\mu$ verifica $\mu(X) \geq 2c - 1$.*

Para probar el teorema necesitaremos el siguiente resultado, cuya demostración omitiremos. El lector interesado puede consultar [Ein+14, Lema B.2].

**Lema 4.10.** *Sea $\mu$ una medida $A$-invariante en $X$. Fijemos $\eta > 0$, $\varepsilon \in (0,1)$. Para $N \geq 1$, sea $BC_\eta(N,\varepsilon)$ el número mínimo de $(N,\eta)$-bolas de Bowen necesarias para cubrir cualquier subconjunto de $X$ de medida mayor a $1-\varepsilon$. Entonces*

$$h_\mu(T) \leq \lim_{\varepsilon\to 0} \liminf_{N\to\infty} \frac{\log BC_\eta(N,\varepsilon)}{2N}. \tag{4.10}$$

*Demostración de 4.9.* Si $\mu = \int_Y \mu_y \, d\tau(y)$ es la descomposición ergódica de $\mu$, entonces

$$\mu(X_{\geq M}) = \int_Y \mu_y(X_{\geq M}) \, d\tau(y)$$

y

$$h_\mu(T) = \int_Y h_{\mu_y}(T) \, d\tau(y).$$

Luego podemos suponer que $\mu$ es ergódica. Vamos a aplicar 4.10 para acotar $h_\mu(T)$. Para esto debemos ver que se puede cubrir la mayor parte de $X$ por pocas $N$-bolas de Bowen. Si $M > 3$ entonces toda $T$-órbita interseca a $X_{<M}$. Dado $x_0 \in \operatorname{sop}(\mu)$, existe $n \in \mathbb{Z}$ tal que $T^n(x_0) \in X_{<M}$. Entonces $T^{-n}X_{<M}$ es un entorno abierto de $x_0$, así que $\mu(T^{-n}X_{<M}) > 0$ y luego $\mu(X_{<M}) > 0$ por $T$-invariancia. Esto implica que $\mu\left(\bigcup_{k=0}^{\infty} T^{-k}X_{<M}\right) = 1$ por ergodicidad, así que para cada $\varepsilon > 0$ existe $R > 0$ tal que el conjunto

$$Y = \bigcup_{k=0}^{R-1} T^{-k}X_{<M}$$

tiene $\mu(Y) > 1 - \varepsilon$. Por el teorema ergódico puntual,

$$\frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{1}_{X_{\geq M}}(T^n x) \to \mu(X_{\geq M}) \quad \text{c.t.p.}$$

cuando $N \to \infty$. Como $\mu$ es una medida de probabilidad, la convergencia c.t.p. implica convergencia en medida, así que para todo $N \in \mathbb{N}$ suficientemente grande existe un conjunto $X_1 \subseteq X$ que depende de $N$ con $\mu(X_1) > 1 - \varepsilon$ tal que

$$\frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{1}_{X_{\geq M}}(T^n x) > \kappa = \mu(X_{\geq M}) - \varepsilon$$

para todo $x \in X_1$. Claramente el conjunto $Z = X_1 \cap T^N Y \cap T^{-N} Y$ tiene medida mayor que $1 - 3\varepsilon$. Por definición de $Y$ podemos descomponer a $Z$ en $R^2$ conjuntos de la forma

$$Z_{r,s} = X_1 \cap T^{N-r} X_{<M} \cap T^{-N-s} X_{<M}$$

con $r, s \in [0, R-1]$. A su vez podemos dividir a cada $Z_{r,s}$ en conjuntos $Z_{r,s} \cap Z(V)$ con $V \subseteq [-N, N]$, donde $Z(V)$ es el conjunto definido en el teorema 4.8. Por la definición de $X_1$, solo hace falta

considerar aquellos conjuntos $V \subseteq [-N, N]$ tales que $|V| \geq \kappa(2N+1)$, ya para los otros $Z(V)$ va a tener intersección vacía con $X_1$. Por el teorema 4.8 hay a lo sumo $\ll_M \exp\left(\frac{2\log\log M}{\log M} N\right)$ tales $V$ con $Z(V) \neq \varnothing$ y para cada uno de ellos, $Z(V)$ puede ser cubierto por

$$\ll_M e^{2N - \frac{1}{2}|V|} \ll e^{2N(1-\kappa/2)}$$

$N$-bolas de Bowen. Por lo tanto $Z$ puede ser cubierto por $\ll_{M,R} \exp\left(\frac{2\log\log M}{\log M} N + 2N(1-\frac{\kappa}{2})\right)$ $N$-bolas de Bowen. En particular $BC_\eta(N,\varepsilon) \leq \frac{2\log\log M}{\log M} N + 2N(1-\frac{\kappa}{2}) + O_{M,\varepsilon}(1)$. Por el lema 4.10 deducimos que

$$h_\mu(T) \leq \frac{\log\log M}{\log M} + 1 - \frac{\kappa}{2} = 1 + \frac{\log\log M}{\log M} - \frac{\mu(X_{\geq M}) - \varepsilon}{2}.$$

Como $\varepsilon > 0$ es arbitrario, el teorema se sigue. □

El siguiente lema será usado en los lemas 4.12 y 4.14.

**Lema 4.11.** *Supongamos que $S_1, \ldots, S_\ell$ son $N$-bolas de Bowen de radio $\eta > 0$ contenidas en $X_{\leq M}$. Si $\eta$ es suficientemente chico con respecto a $M$, entonces existen elementos $a_1, \ldots, a_J \in A$ con $J \ll e^N$ tales que*

$$\bigcup_{i=1}^{\ell} S_i \times S_i \subseteq \bigcup_{j=1}^{J} \{(x, ya_j) : d(x,y) < e^{-N}\}.$$

*Demostración.* Consideremos una bola de Bowen $S_i = x_i B_{N,\eta}$ y sean $x, y \in S_i$. Asumiendo que $\eta$ es más chico que el radio de inyectividad en $\bigcup_{n=-N}^{N} X_{\leq M} a^n$, donde $a = \left(\begin{smallmatrix} e^{1/2} & 0 \\ 0 & e^{-1/2} \end{smallmatrix}\right)$, tenemos que $S_i$ es igual al conjunto de puntos $w \in X$ con $d(T^n w, T^n x_i) < \eta$ para todo $n \in [-N, N]$. Se sigue que $d(T^n x, T^n y) < 2\eta$ para todo $n \in [-N, N]$. Ahora, si definimos $h = x^{-1}y$, tenemos que

$$d(xa^n, ya^n) = d(a^n, ha^n) = d(1, a^{-n} h a^n) < 2\eta. \tag{4.11}$$

Supongamos que $h$ está representado por la matriz $\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$. Como

$$\begin{pmatrix} e^{-n/2} & 0 \\ 0 & e^{n/2} \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} e^{n/2} & 0 \\ 0 & e^{-n/2} \end{pmatrix} = \begin{pmatrix} \alpha & e^{-n}\beta \\ e^{n}\gamma & \delta \end{pmatrix},$$

la condición (4.11) implica que $\max(|\beta|, |\gamma|) \ll e^{-N}\eta$ y $\max(|\alpha - 1|, |\delta - 1|) \ll \eta$ por el lema 4.2. Sea $c > 0$ la constante implícita en las últimas dos desigualdades. Podemos elegir $\alpha_1 < \alpha_2 < \ldots < \alpha_J$ en $[1 - c\eta, 1 + c\eta]$ con $J \ll e^N$, tales que $|\log \alpha_{j+1} - \log \alpha_j| < 2c\eta e^{-N}$ para todo $j$. Como $\alpha \in [1 - c\eta, 1 + c\eta]$, existe $1 \leq i \leq J$ tal que $|\log \alpha - \log \alpha_i| < c\eta e^{-N}$. Entonces

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \alpha_i^{-1} & 0 \\ 0 & \alpha_i \end{pmatrix} = \begin{pmatrix} \alpha/\alpha_i & \beta\alpha_i \\ \gamma/\alpha_i & \delta\alpha_i \end{pmatrix}.$$

Sea $a_i = \left(\begin{smallmatrix} \alpha_i & 0 \\ 0 & \alpha_i^{-1} \end{smallmatrix}\right)$. Veamos que $\|h a_i^{-1} - 1\| \ll \eta e^{-N}$ para $\eta$ suficientemente chico. En primer lugar,

$$\left|\frac{\alpha}{\alpha_i} - 1\right| = \left|e^{\log(\alpha/\alpha_i)} - 1\right| \ll \left|\log\left(\frac{\alpha}{\alpha_i}\right)\right| \ll \eta e^{-N}.$$

Además, como $\frac{1}{2} < 1 - c\eta < 1 + c\eta < 2$ para $\eta$ chico, tenemos que $|\beta\alpha_i|$ y $|\gamma/\alpha_i|$ son $\ll \eta e^{-N}$. Finalmente, es

$$\begin{aligned} |\alpha_i\delta - 1| &= \left|\alpha\delta\frac{\alpha_i}{\alpha} - 1\right| = \left|(1+\beta\gamma)\frac{\alpha_i}{\alpha} - 1\right| \\ &\leq \left|\frac{\alpha_i}{\alpha} - 1\right| + \left|\beta\gamma\frac{\alpha_i}{\alpha}\right| \ll \eta e^{-N} + \eta^2 e^{-2N} \ll \eta e^{-N}. \end{aligned}$$

Esto demuestra la afirmación. Como $d(1, ha_i^{-1}) \ll \|ha_i^{-1} - 1\|$, se ve que $d(1, ha_i^{-1}) < e^{-N}$ siempre que $\eta$ es suficientemente chico con respecto a las constantes. Luego $d(x, ya_i^{-1}) < e^{-N}$, así que $(x,y) \in \{(u, va_i) \in X_{\leq M}^2 : d(u,v) < e^{-N}\}$. Como esto vale para cualquier $N$-bola de Bowen, el resultado se sigue. □

**Lema 4.12.** *Sea $(\mu_i)_i$ una sucesión de medidas como en el teorema 4.6. Entonces la sucesión $(\mu_i)_i$ es tight y todo punto límite $\mu \in \mathscr{M}^1(X)$ de la sucesión cumple*

$$\mu(X_{\leq M}) \geq 1 - \frac{2\log\log M}{\log M}, \tag{4.12}$$

*para todo $M > 0$ suficientemente grande.*

*Demostración de 4.12.* Sea $M > 0$ suficientemente grande. Basta ver que $\mu(X_{<M}) \geq 1 - \kappa$ para todo $\kappa > \frac{2\log\log M}{\log M}$. Fijemos un tal $\kappa$, sea $\varepsilon = \varepsilon(\kappa)$ una constante a determinar y sean $N_i = \lceil -\log \delta_i \rceil$, $H_i = \delta_i^{-\varepsilon}$. Como $H_i \to \infty$, podemos suponer que $H_i > M$. Usando la fórmula (2.10) y la descripción de las geodésicas dada en la proposición 2.11 es fácil ver que la trayectoria geodésica de cualquier punto $x \in X_{\leq H_i}$ cae en $X_{<M}$ en a lo sumo $2\log H_i - 2\log M + O(1) \leq 2\varepsilon N_i$ pasos, ya sea en el pasado o en el futuro. Entonces

$$X_{\leq H_i} \subseteq \bigcup_{-2\varepsilon N_i \leq n \leq 2\varepsilon N_i} T^n X_{<M}.$$

Definiendo $N_i' = N_i + \lfloor 2\varepsilon N_i \rfloor$, se sigue que

$$T^{N_i'} X_{\leq H_i} \cap T^{-N_i'} X_{\leq H_i} \subseteq \bigcup_{-2\varepsilon N_i \leq m, n \leq 2\varepsilon N_i} T^{N_i'+m} X_{<M} \cap T^{-N_i'+n} X_{<M}.$$

Ahora consideremos el conjunto

$$X_\kappa = \left\{ x \in T^{N_i'} X_{\leq H_i} \cap T^{-N_i'} X_{\leq H_i} : \tfrac{1}{2N_i'+1} \sum_{n=-N_i'}^{N_i'} \mathbb{1}_{X_{\geq M}}(T^n x) \geq \kappa \right\},$$

que consiste de puntos cuya $T$-órbita pasa una proporción positiva del tiempo en $X_{\geq M}$. Notemos que

$$\mu_i(X_{\geq M}) = \int \mathbb{1}_{X_{\geq M}}\, d\mu_i = \int \frac{1}{2N_i'+1} \sum_{n=-N_i'}^{N_i'} \mathbb{1}_{X_{\geq M}} \circ T^n \, d\mu_i \tag{4.13}$$

$$\leq \kappa + \mu_i(X_\kappa) + \mu_i\left( (T^{N_i'} X_{\leq H_i} \cap T^{-N_i'} X_{\leq H_i})^c \right) \leq \kappa + \mu_i(X_\kappa) + 2\mu_i(X_{\geq H_i}). \tag{4.14}$$

Como $\mu_i(X_{\geq H_i}) \to 0$ por hipótesis, se ve que para acotar $\mu(X_{\geq M})$ basta acotar $\mu_i(X_\kappa)$. Podemos escribir a $X_\kappa$ como unión de conjuntos

$$Z' = X_\kappa \cap T^{N_i'+r} X_{<M} \cap T^{-N_i'+s} X_{<M}$$

con $r, s \in [-2\varepsilon N_i, 2\varepsilon N_i]$. Reemplazando a $Z'$ por el shift $Z = T^{-h} Z'$ donde $h = N_i' + r$ tenemos que $Z \subseteq X_{<M} \cap T^{-N} X_{<M}$ donde $N = 2N_i' + r - s \in [2N_i, 2N_i + 4\varepsilon N_i]$. Las $T$-órbitas de los puntos en $Z'$ pasan una proporción positiva del tiempo en $X_{\geq M}$, así que lo mismo vale para $Z$. En efecto, dado $y \in Z$ tenemos que $T^{-h} y \in Z' \subseteq X_\kappa$ y luego

$$\kappa(2N_i'+1) < \sum_{n=-N_i'}^{N_i'} \mathbb{1}_{X_{\geq M}}(T^{-(n+h)} y) = \sum_{n=r}^{2N_i'+r} \mathbb{1}_{X_{\geq M}}(T^{-n} y) = \sum_{n=r}^{N+s} \mathbb{1}_{X_{\geq M}}(T^{-n} y).$$

Entonces

$$\sum_{n=0}^{N} \mathbb{1}_{X_{\geq M}}(T^{-n} y) \geq \sum_{n=r}^{N} \mathbb{1}_{X_{\geq M}}(T^{-n} y) \geq \kappa(2N_i'+1) - s \geq (\kappa - O(\varepsilon))(N+1).$$

Esto implica que

$$Z \subseteq \left\{ y \in T^N X_{<M} \cap T^{-N} X_{<M} : \tfrac{1}{N+1} \sum_{n=0}^{N} \mathbb{1}_{X_{\geq M}}(T^n y) > \kappa - O(\varepsilon) \right\}.$$

En particular $\mathrm{Rec}(x, X_{\geq M}) \cap [0, N]$ tiene por lo menos $(\kappa - O(\varepsilon))N$ elementos, para todo $x \in Z$. En otras palabras, $Z$ está contenido en la unión de los conjuntos $Z(V)$ del teorema 4.8 con $|V| \geq (\kappa - O(\varepsilon))N$. Se sigue que $Z$ puede ser cubierto por

$$\ell \ll_M \exp\left(\frac{\log\log M}{\log M} N + \left(1 - \frac{\kappa}{2} - O(\varepsilon)\right) N\right) \leq \exp\left(\frac{2\log\log M}{\log M} N_i + (2 - \kappa - O(\varepsilon))N_i\right)$$

$N$-bolas de Bowen. Como $N \geq N_i$, también podemos cubrir a $Z$ con $\ell$ $N_i$-bolas de Bowen, digamos $S_1, \ldots, S_\ell$. Por el lema 4.11 se deduce que

$$\bigcup_{j=1}^{\ell} S_j \times S_j \subseteq \bigcup_{j=1}^{J} \{(x, ya_j) : d(x, y) < \delta_i\}$$

para una elección adecuada de elementos $a_1, \ldots, a_J \in A$, con $J \ll e^N$, puesto que $e^{-N_i} \leq \delta_i$. La identidad sigue valiendo si reemplazamos a los $S_i$ por $\widehat{S}_i = S_i \setminus \left(\bigcup_{j<i} S_i\right)$, que son disjuntos dos a dos. Por la hipótesis (2) del teorema 4.6, resulta

$$\sum_{j=1}^{\ell} \mu_i(\widehat{S}_j)^2 \ll_\varepsilon \delta_i^{3-5\varepsilon} J \ll e^{-3N_i + 5\varepsilon N_i} e^{N_i} = e^{-2N_i + 5\varepsilon N_i}.$$

Luego, por Cauchy-Schwarz, es

$$\mu_i(Z) \leq \sum_{j=1}^{\ell} \mu(S_j) \leq \sqrt{\ell}\left(\sum_{j=1}^{\ell} \mu(\widehat{S}_j)^2\right)^{1/2} \ll_\varepsilon \exp\left(\frac{\log\log M}{\log M} N_i - \frac{\kappa}{2} N_i + O(\varepsilon) N_i\right).$$

Sumando sobre $r, s \in [-2\varepsilon N_i, 2\varepsilon N_i]$, obtenemos

$$\mu_i(X_\kappa) \ll_{\varepsilon, M} e^{\left(\frac{\log\log M}{\log M} - \kappa/2 + O(\varepsilon)\right) N_i}$$

ya que podemos absorber el factor $(4\varepsilon N_i + 1)^2$ dentro del término $e^{O(\varepsilon)N_i}$. Luego podemos elegir $\varepsilon > 0$ tal que $\mu_i(X_\kappa) \ll_{\varepsilon, M} e^{-cN_i}$, donde $c = c(\varepsilon, M, \kappa) > 0$ es una constante que no depende de $i$. Esto implica que la sucesión $(\mu_i)_i$ es tight. En efecto, dado $\eta > 0$ podemos elegir $M$ y $\kappa$ de modo que $\frac{2\log\log M}{\log M} < \kappa < \eta/2$. Por (4.13) se sigue que $\mu_i(X_{>M}) \leq \eta/2 + o(1)$, así que $\mu_i(X_{\leq M}) > 1 - \eta$ para todo $i$ suficientemente grande. Además, si $\mu \in \mathscr{M}^1(X)$ es límite de una subsucesión $(\mu_{i_n})_n$, haciendo $n \to \infty$ en la desigualdad

$$\mu_{i_n}(X_{>M}) \leq \kappa + O_{\varepsilon, M}(e^{-cN_{i_n}}) + 2\mu_{i_n}(X_{\geq H_{i_n}})$$

obtenemos por el lema 1.5 que $\mu(X_{>M}) \leq \kappa$, como queríamos ver. □

Habiendo probado que la sucesión $(\mu_i)_i$ es tight, solo hace falta probar que el único punto límite de la sucesión $(\mu_i)$ es la medida de Haar $\mu_X$ ya que en ese caso un argumento sencillo nos permitirá concluir que $\mu_i$ converge a $\mu_X$. Este resultado se deduce del lema 4.14 que demostraremos más adelante. Para probarlo necesitamos el siguiente resultado, que dice que hay una partición $\mathscr{P}$ en la que la mayoría de los elementos de $\mathscr{P}^{[-N,N]}$ pueden ser cubiertos por pocas bolas de Bowen.

**Lema 4.13.** *Para todo $M > 1$ existe una partición finita $\mathscr{P}$ de $X$ tal que para todo $\kappa \in (0, 1)$ y todo $N$, existe un conjunto $X' \subseteq X$ que cumple:*

*(i) $X'$ es la unión de $\ell$ conjuntos $S_1, \ldots, S_\ell \in \mathscr{P}^{[-N,N]}$.*

*(ii) Todo $S_j$ está contenido en la unión de a lo sumo $3^{\kappa(2N+1)}$ $N$-bolas de Bowen. En particular $X'$ es acotado.*

*(iii) $\mu(X') \geq 1 - 2\mu(X_{\geq M})\kappa^{-1}$ para toda medida de probabilidad $T$-invariante.*

*Más aún, para cualquier medida $\mu_0 \in \mathscr{M}_T^1(X)$ fija podemos elegir $\mathscr{P}$ de modo que $\mu_0(\partial S) = 0$ para todo $S \in \mathscr{P}$.*

*Demostración.* Sea $\eta > 0$ un número suficientemente chico con respecto al radio de inyectividad de $X_{\leq M}$. Para construir la partición $\mathscr{P}$, observemos primero que $\mu(\partial X_{<M}) = 0$ para toda medida $\mu \in \mathscr{M}_T^1(X)$. En efecto, $W = \partial X_{<M}$ es el conjunto de puntos de altura $M$. Si $\mu(W) > 0$, por el teorema de recurrencia de Poincaré el conjunto de puntos $x \in X$ tales que $T^n x \in W$ para infinitos $n$ tendría medida $> 0$. Sin embargo este conjunto es vacío ya que en toda $A$-órbita $xA$ hay a lo sumo 2 puntos con altura igual a $M$. Por lo tanto $\mu(W) = 0$. Sea $\mu_0 \in \mathscr{M}_T^1(X)$ una medida fija. Para cada $x \in X$, la función $r \in \mathbb{R}_{>0} \mapsto \mu_0(B_r(x))$ es creciente, así que es continua c.t.p. Se sigue que existen reales $r > 0$ arbitrariamente chicos tales que $\mu_0(\partial B_r(x)) = 0$. Por compacidad de $X_{\leq M}$ podemos cubrir este conjunto con finitas bolas $B_{r_i}(y_i)$, $1 \leq i \leq k$ con radios $r_i < \eta/2$, tales que $\mu_0(\partial B_{r_i}(y_i)) = 0$ para todo $i$. Con estas bolas obtenemos una partición de $X_{<M}$ tomando $E_i = B_{r_i}(y_i) \cap X_{<M}$ y $P_i = E_i \setminus \bigcup_{j<i} E_j$ para todo $1 \leq i \leq k$. Finalemente tomamos $\mathscr{P} = \{X_{\geq M}, P_1, \ldots, P_k\}$. Notemos que cada $P_i$ tiene diámetro $< \eta$ ya que está contenido en $B_{r_i}(y_i)$. Además es claro que $\mu_0(\partial P_i) = 0$ para todo $i$.

Notar que como $\mathscr{P}$ es más fina que la partición $\{X_{<M}, X_{\geq M}\}$, todo par de puntos $x, y \in X$ que pertenecen a un mismo conjunto $S \in \mathscr{P}^{[-N,N]}$ cumplen que

$$\mathrm{Rec}(x, X_{<M}) \cap [-N, N] = \mathrm{Rec}(y, X_{<M}) \cap [-N, N]$$

y además $d(T^n x, T^n y) < \eta$ siempre que $T^n x, T^n y \in X_{<M}$ para algún $n \in [-N, N]$. En particular el promedio $F(x) = \frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{1}_{X_{\geq M}}(T^n x)$ es constante en los elementos de $\mathscr{P}^{[-N,N]}$. Consideremos el conjunto

$$X' = \{x \in T^N X_{<M} : F(x) \leq \kappa\}.$$

Si $\mu \in \mathscr{M}_T^1(X)$, entonces $\int F \, d\mu = \mu(X_{\geq M})$ por $T$-invariancia, así que

$$\mu(X_{\geq M}) = \int F \, d\mu \geq \int\limits_{\{x:F(x)>\kappa\}} F \, d\mu \geq \kappa \mu(\{x : F(x) > \kappa\}).$$

Entonces

$$\begin{aligned} \mu(X') &= \mu(T^N X_{<M} \setminus \{x : F(x) > \kappa\}) \\ &\geq \mu(T^N X_{<M}) - \mu(\{x : F(x) > \kappa\}) \\ &\geq \mu(X_{<M}) - \kappa^{-1} \mu(X_{\geq M}) = 1 - (1 + \kappa^{-1}) \mu(X_{\geq M}). \end{aligned}$$

Como $\kappa < 1$ se sigue que $X'$ cumple la condición $(iii)$ del enunciado.

Como $F$ es constante en los elementos de $\mathscr{P}^{[-N,N]}$, el conjunto $X'$ es unión disjunta de elementos de $\mathscr{P}^{[-N,N]}$. Supongamos que $S \in \mathscr{P}^{[-N,N]}$ está contenido en $X'$ y sea $S' = T^{-N} S$. Entonces todo par de puntos $x, y \in S'$ verifica

$$\begin{aligned} \frac{1}{2N+1} \sum_{n=0}^{2N} \mathbb{1}_{X_{\geq M}}(T^n x) &= \frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{1}_{X_{\geq M}}(T^n (T^N x)) \\ &= \frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{1}_{X_{\geq M}}(T^n (T^N y)) = \frac{1}{2N+1} \sum_{n=0}^{2N} \mathbb{1}_{X_{\geq M}}(T^n y), \end{aligned}$$

y $d(T^n x, T^n y) < \eta$ siempre que $T^n x, T^n y \in X_{<M}$ para algún $n \in [0, 2N]$.

Sea $V = \{n \in [0, 2N] : T^n S' \subseteq X_{\geq M}\}$. Utilizaremos la notación $U^+ = \{\left(\begin{smallmatrix} 1 & 0 \\ t & 1 \end{smallmatrix}\right) : t \in \mathbb{R}\}$ y $U^- = \{\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right) : t \in \mathbb{R}\}$. Veamos por inducción que para todo $n \in [0, 2N]$, el conjunto $S'$ está contenido en la unión de $3^{|[0,n] \cap V|}$ conjuntos de la forma

$$x Q_n = x B^{U^+}_{c_0 \eta e^{-n}} B^{U^- A}_{c_0 \eta}, \qquad \text{con } x \in S',$$

donde $c_0 = O(1)$ y $B_r^H$ denota la bola de radio $r$ y centro 1 de un subgrupo $H \leq G$. Para $n = 0$ esto es claro. Sea ahora $0 < n \leq 2N$ y supongamos que la afirmación vale para $n - 1$, es decir,

$$S' \subseteq \bigcup_{j=1}^{3^m} x_j Q_{n-1},$$

donde $m = |[0, n-1] \cap V|$ y $x_1, \dots, x_{3^m} \in S'$. Si $n \in V$, entonces podemos cubrir a $B^{U^+}_{2\eta e^{1-n}}$ con tres bolas de radio $2\eta e^{-n}$. Luego es posible cubrir cada $x_j Q_{n-1}$ con tres conjuntos $y_i Q_n$, $3j-2 \le i \le 3j$ eligiendo los centros apropiadamente, por lo que la afirmación vale para $n$. Supongamos que en cambio $n \notin V$, de modo que $T^n S' \subseteq X_{<M}$. En tal caso $T^n S' \subseteq P_i$ para algún $i$ y en particular $\operatorname{diam}(T^n S') < \eta$. Afirmamos que en este caso $x_j Q_{n-1} \cap S' \subseteq x_j Q_n$ para todo $j$. En efecto, sean $u \in B^{U^+}_{c_0 \eta e^{1-n}}$ y $g \in B^{U^- A}_{c_0 \eta}$ tales que $x_j u g \in S'$. Sea $a = \left(\begin{smallmatrix} e^{1/2} & 0 \\ 0 & e^{-1/2} \end{smallmatrix}\right)$, de modo que $Tx = xa$. Como $T^n S'$ tiene diámetro menor que $\eta$, tenemos que

$$d(x_j a^n, x_j u g a^n) = d(1, a^{-n} u g a^n) < \eta.$$

Sean $c_1, c_2 > 0$ las constantes del lema 4.2 aplicado al compacto $K = \bigcup_{n=0}^{2N} a^{-n} X_{\le M} a^n$. Notar que $S' \subseteq T^{-N} X' \subseteq X_{\le M}$. Podemos elegir representantes $\left(\begin{smallmatrix} 1 & 0 \\ t & 1 \end{smallmatrix}\right)$ y $\left(\begin{smallmatrix} \alpha & \beta \\ 0 & 1/\alpha \end{smallmatrix}\right)$ de $u$ y $g$ respectivamente. Entonces $ug$ está representado por

$$\begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ 0 & \alpha^{-1} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ t\alpha & t\beta + \alpha^{-1} \end{pmatrix}$$

y $a^{-n} u g a^n$ por

$$\begin{pmatrix} e^{-n/2} & 0 \\ 0 & e^{n/2} \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ t\alpha & t\beta + \alpha^{-1} \end{pmatrix} \begin{pmatrix} e^{n/2} & 0 \\ 0 & e^{-n/2} \end{pmatrix} = \begin{pmatrix} \alpha & e^{-n}\beta \\ e^n t\alpha & t\beta + \alpha^{-1} \end{pmatrix}. \tag{4.15}$$

Como $c_1 \|a^{-n} u g a^n - 1\| < d(a^{-n} u g a^n, 1) < \eta$, en particular $|e^n t \alpha| < \eta / c_1$. Además $|\alpha^{-1}| \le c_0 \eta / c_1$ ya que $d(g, 1) < c_0 \eta$. Se sigue que $|e^n t| < c_0 \eta^2 c_1^{-2} < c_0 \eta / c_2$ siempre que $\eta < c_1^2 / c_2$ y luego

$$d(u, 1) < c_2 \|u - 1\| = c_2 |t| < c_0 \eta e^{-n}.$$

Por lo tanto $ug \in Q_n$, así que

$$S' \subseteq \bigcup_{j=1}^{3^m} x_j Q_n.$$

Para concluir la demostración basta con probar que $Q_{2N} \subseteq a^N B_{N,\eta} a^{-N}$, ya que en ese caso

$$S = T^N S' = S' a^N \subseteq \bigcup_{j=1}^{J} x_j Q_{2N} a^N \subseteq \bigcup_{j=1}^{J} x_j a^N B_{N,\eta},$$

donde $J = 3^{|V|} \le 3^{\kappa(2N+1)}$. Dados $u \in B^{U^+}_{c_0 \eta e^{-2N}}$ y $g \in B^{U^- A}_{c_0 \eta}$, tenemos que $a^{-n} u g a^n$ es de la forma (4.15). Podemos acotar cada una de las entradas de esta matriz menos la identidad por $\eta / c_2$. Por ejemplo, como $d(u, 1) < c_0 \eta e^{-2N}$ y $d(g, 1) < c_0 \eta$, tenemos que $|\alpha| \le 1 + \|g - 1\| < 1 + c_0 \eta \le 2$ si $\eta$ es suficientemente chico, y

$$|t| \le \|u - 1\| \le \frac{1}{c_1} d(u, 1) < \frac{c_0}{c_1} \eta e^{-2N},$$

así que $|e^n t \alpha| < \frac{2 c_0}{c_1} \eta e^{n-2N} < \eta$ si tomamos $c_0 < c_1 / 2$. Luego

$$d(a^{-n} u g a^n, 1) \le c_2 \|a^{-n} u g a^n - 1\| < \eta,$$

y por lo tanto $a^{-n} u g a^n$. Esto vale para todo $n \in [0, 2N]$, así que

$$ug \in \bigcap_{n=0}^{2N} a^n B^G_\eta a^{-n} = a^N B_{N,\eta} a^{-N}. \qquad \square$$

Usando el lema anterior podemos probar el siguiente lema, que es el último ingrediente de la demostración del teorema de Duke.

**Lema 4.14.** *Sea $(\mu_i)_i$ una sucesión de medidas de probabilidad como en 4.6. Entonces todo punto límite débil* $\mu$ de $(\mu_i)_i$ tiene entropía máxima $h_\mu(T) = 1$.*

*Demostración.* Sea $\mu$ un límite débil* de una subsucesión de $(\mu_i)_i$. Por el teorema 4.7 tenemos que $h_\mu(T) \leq 1$, así que basta ver que $h_\mu(T) \geq 1$. Fijemos $M \geq 1$, $\varepsilon > 0$ y sea $\mathscr{P}$ una partición finita de $X$ que cumple las condiciones del enunciado de 4.13 y tal que

$$\mu(\partial\mathscr{P}) = \mu\left(\bigcup_{S\in\mathscr{P}} \partial S\right) = 0.$$

Además tomamos $\kappa = \mu(X_{\geq M})^{1/2}$. Para cada $i \in \mathbb{N}$ sea $N_i = \lceil -\log \delta_i \rceil$ y sea $\mathscr{P}_i = \mathscr{P}^{[-N_i,N_i]}$. Queremos probar que $H_{\mu_i}(\mathscr{P}_i)$ es grande (i.e. que es aproximadamente 1). Por el lema 4.13 existe un subconjunto $X_i \subseteq X$ que cumple las condiciones $(i)$-$(iii)$ con $N = N_i$. Entonces todo $S \subseteq \mathscr{P}_i$ con $S \subseteq X_i$ está contenido en la unión de $\ll 3^{\kappa(2N_i+1)}$ $N_i$-bolas de Bowen. Podemos reemplazar este cubrimiento por una partición $\mathscr{R}(S)$ de $S$ en $\ll 3^{\kappa(2N_i+1)}$ conjuntos medibles cada uno de los cuales está contenido en una de las $N_i$-bolas de Bowen. Sea $\mathscr{Q}_i$ la partición de $X$ que consiste de los elementos $S \subseteq \mathscr{P}_i$ con $S \subseteq X \setminus X_i$, junto con los elementos de $\mathscr{R}(S)$ para todo $S \in \mathscr{P}_i$ con $S \subseteq X_i$. Por definición

$$H_{\mu_i}(\mathscr{Q}_i|\mathscr{P}_i) = -\sum_{\substack{B\in\mathscr{Q}_i\\ S\in\mathscr{P}_i}} \mu(B\cap S)\log\left(\frac{\mu(B\cap S)}{\mu(S)}\right).$$

Ahora, todo $B \in \mathscr{Q}_i$ está contenido en un único $S \in \mathscr{P}_i$. Si $B = S$ entonces $\mu(B \cap S) = \mu(S)$ y luego el término correspondiente al par $(B, S)$ no contribuye a la suma. Si $B \subsetneq S$, debe ser $S \subseteq X_i$. Se sigue que

$$H_{\mu_i}(\mathscr{Q}_i|\mathscr{P}_i) = -\sum_{\substack{S\in\mathscr{P}_i\\ S\subseteq X_i}}\sum_{\substack{B\in\mathscr{Q}_i\\ B\subseteq S}} \mu(B\cap S)\log\left(\frac{\mu(B\cap S)}{\mu(S)}\right) = \sum_{\substack{S\in\mathscr{P}_i\\ S\subseteq X_i}} \mu_i(S) H_{\mu_i|S}(\mathscr{Q}_i) \tag{4.16}$$

donde $\mu_i|_S$ es la medida tal que $\mu_i|_S(E) = \frac{\mu_i(E\cap S)}{\mu_i(S)}$ para todo $E$ medible. Por 1.44.$(e)$ tenemos que $H_{\mu_i|S}(\mathscr{Q}_i) \leq \log \operatorname{card}_{\mu_i}(\mathscr{Q}_i) \leq \kappa(2N_i+1)\log 3$, así que

$$H_{\mu_i}(\mathscr{Q}_i|\mathscr{P}_i) \leq \kappa(2N_i+1)\log 3. \tag{4.17}$$

Por otro lado, $H_{\mu_i}(\mathscr{Q}_i) = H_{\mu_i}(\mathscr{P}_i) + H_{\mu_i}(\mathscr{Q}_i|\mathscr{P}_i)$ ya que $\mathscr{Q}_i$ es más fina que $\mathscr{P}_i$. Luego para acotar $H_{\mu_i}(\mathscr{P}_i)$ basta acotar $H_{\mu_i}(\mathscr{Q}_i)$. En primer lugar tenemos

$$H_{\mu_i}(\mathscr{Q}_i) \geq H_{\mu_i}(\mathscr{Q}_i|\{X_i, X_i^c\}) \geq \mu(X_i) H_{\mu_i|X_i}(\mathscr{Q}_i), \tag{4.18}$$

y por la observación 1.43,

$$H_{\mu_i|X_i}(\mathscr{Q}_i) \geq -\log \sum_{\substack{B\in\mathscr{Q}_i\\ B\subseteq X_i}} \left(\frac{\mu(B)}{\mu(X_i)}\right)^2.$$

Por construcción de $\mathscr{Q}_i$, todo $B \in \mathscr{Q}_i$ con $B \subseteq X_i$ está contenido en una $N_i$-bola de Bowen. Por el lema 4.11 se sigue que

$$\bigcup_{\substack{B\in\mathscr{Q}_i\\ B\subseteq X_i}} B\times B \subseteq \bigcup_{j=1}^{J} \{(x, ya_j) : d(x,y) < \delta_i\}$$

donde $J \ll_M e^{N_i}$ y $a_1, \ldots, a_J \in A$. Por la hipótesis (2) del teorema 4.6,

$$\sum_{\substack{B\in\mathscr{Q}_i\\ B\subseteq X_i}} \mu(B)^2 \ll_\varepsilon \delta_i^{3-5\varepsilon} e^{N_i} \ll e^{(-2+5\varepsilon)N_i}.$$

Sea $C_\varepsilon > 0$ la constante implícita en esta desigualdad. Juntando estas desigualdades, obtenemos

$$\begin{aligned} H_{\mu_i}(\mathscr{Q}_i) &\geq -\mu_i(X_i)\log\left[\mu_i(X_i)^{-2}\sum_{B\in\mathscr{Q}_i, B\subseteq X_i}\mu(B)^2\right] \\ &\geq \mu_i(X_i)\left(2\log\mu_i(X_i) - \log C_\varepsilon + (2-5\varepsilon)N_i\right). \end{aligned}$$

Los primeros dos términos son $O_\varepsilon(1)$, así que para $N_i$ suficientemente grande tenemos que

$$H_{\mu_i}(\mathscr{Q}_i) \geq \mu_i(X_i)(2-6\varepsilon)N_i \geq (1-2\mu_i(X_{\geq M})\kappa^{-1})(2-6\varepsilon)N_i$$

usando la propiedad $(iii)$ del lema 4.13. Por (4.18) y (4.17) esto nos permite acotar

$$H_{\mu_i}(\mathscr{P}_i) \geq (1-2\mu_i(X_{\geq M})\kappa^{-1})(2-6\varepsilon)N_i - O(\kappa N_i).$$

Justo ahora no podemos hacer $i \to \infty$ para obtener un resultado sobre $\mu$ ya que la partición $\mathscr{P}_i$ depende de $i$. Sin embargo, podemos elegir $N_0 \in \mathbb{N}$ fijo y cubrir a $[-N_i, N_i]$ por intervalos de la forma

$$I_k = [-N + 2kN_0, -N + 2(k+1)N_0]$$

con $0 \leq k \leq \lceil N_0/N \rceil - 1$ entero. Entonces

$$\mathscr{P}_i = \bigvee_{k=0}^{\lceil N_0/N \rceil - 1} \mathscr{P}^{I_k}.$$

Por otro lado, todos los $I_k$ son traslaciones de $[-N_0, N_0]$, así que $H_{\mu_i}(\mathscr{P}^{I_k}) = H_{\mu_i}(\mathscr{P}^{[-N_0,N_0]})$ por $T$-invariancia. Entonces, por subaditividad de $H_{\mu_i}$ tenemos

$$H_{\mu_i}(\mathscr{P}_i) \leq \lceil N_0/N \rceil H_{\mu_i}(\mathscr{P}^{[-N_0,N_0]}).$$

Luego

$$H_{\mu_i}(\mathscr{P}^{[-N_0,N_0]}) \geq (1-2\mu_i(X_{\geq M})\kappa^{-1})(2-6\varepsilon)N_0 - O(\kappa N_0) - \varepsilon N_0. \tag{4.19}$$

Ahora, por hipótesis $\mu(\partial\mathscr{P}) = 0$, así que $\mu(\partial S) = 0$ para todo $S \in \mathscr{P}^{[-N_0,N_0]} = 0$. Si $\mu_{i_n} \xrightarrow{w*} \mu$, por el punto $(iv)$ del lema 1.5 obtenemos que $\mu_{i_n}(S) \to \mu(S)$ para todo tal $S$ y luego

$$H_{\mu_{i_n}}(\mathscr{P}^{[-N_0,N_0]}) \to H_\mu(\mathscr{P}^{[-N_0,N_0]}).$$

Dividiendo (4.19) por $2N_0$ y haciendo $i_n \to \infty$ obtenemos

$$H_\mu(\mathscr{P}^{[-N_0,N_0]}) \geq (1-2\mu_i(X_{\geq M})^{1/2})(1-3\varepsilon) - O(\mu(X_{\geq M})^{1/2}) - \varepsilon.$$

Por el lema 4.12, $\mu(X_{\geq M}) \leq \frac{2\log\log M}{\log M}$. Luego haciendo $M \to \infty$ y $\varepsilon \to 0$ obtenemos $h_\mu(T) \geq 1$. □

Finalmente estamos en condiciones de probar el teorema 4.6:

*Demostración de 4.6.* Supongamos que $\mu_i$ no tiende a $\mu_X$. Si elegimos una distancia $d$ en $\mathscr{M}^1(X)$ que induzca la topología débil*, esto implica que existe $\varepsilon > 0$ y una subsucesión $(\mu_{i_n})_n$ tal que $d(\mu_{i_n}, \mu_X) \geq \varepsilon$ para todo $n \in \mathbb{N}$. Por el lema 4.12 la sucesión $(\mu_i)_i$ es tight, así que $\mu_{i_n}$ posee un punto límite $\mu$ por el teorema de Prohorov. Por el lema 4.14 tenemos que $h_\mu(T) = 1$ y luego $\mu = \mu_X$ por el teorema 4.7. Pero esto se contradice con que $d(\mu_{i_n}, \mu_X) \geq \varepsilon$. Por lo tanto $\mu_i \xrightarrow{w*} \mu_X$. □

# Bibliografía


[Bil99] P. Billingsley. *Convergence of probability measures*. 2.ª ed. Wiley series in probability and statistics. Probability and statistics section. Wiley, 1999.

[BM00] M. B. Bekka y M. Mayer. *Ergodic theory and topological dynamics of group actions on homogeneous Spaces*. 1.ª ed. London Mathematical Society lecture note series 269. Cambridge University Press, 2000.

[BO07] Y. Benoist y H. Oh. «Equidistribution of rational matrices in their conjugacy classes». En: *GAFA* 17 (2007), págs. 1-32.

[Cas78] J. W. S. Cassels. *Rational Quadratic Forms*. London Mathematical Society Monographs. Academic Press, 1978.

[CF92] M. P. do Carmo y F. Flaherty. *Riemannian geometry*. 1.ª ed. Mathematics. Theory and applications. Birkhäuser, 1992.

[Con] K. Conrad. «A Multivariable Hensel's Lemma». URL: https://kconrad.math.uconn.edu/blurbs/gradnumthy/multivarhensel.pdf.

[Dav82] H. Davenport. *Multiplicative number theory*. 2nd. Graduate Texts in Mathematics 74. Springer-Verlag, 1982.

[Duk88] W. Duke. «Hyperbolic distribution problems and half-integral weight Maass forms». En: *Inventiones Mathematicae* 92 (1988), págs. 73-90.

[Ein+11] M. Einsiedler, E. Lindenstrauss, P. Michel y A. Venkatesh. «Distribution of periodic torus orbits and Duke's theorem for cubic fields». En: *Annals of Math.* 173.2 (2011), págs. 815-885.

[Ein+14] M. Einsiedler, E. Lindenstrauss, P. Michel y A. Venkatesh. «The distribution of closed geodesics on the modular surface, and Duke's theorem». En: *Enseign. Math.* (sep. de 2014).

[EKP15] M. Einsiedler, S. Kadyrov y A. Pohl. «Entropy and escape of mass for $\mathrm{SL}(3,\mathbb{Z})\backslash\mathrm{SL}(3,\mathbb{R})$». En: *Israel J. of Math.* 210 (2015), págs. 245-295.

[EL10] M. Einsiedler y E. Lindenstrauss. «Diagonal actions on locally homogeneous spaces». En: *Clay Math. Proc.* 10 (2010), págs. 155-241.

[EW11] M. Einsiedler y T. Ward. *Ergodic Theory: with a view towards Number Theory*. 1.ª ed. 259. Springer-Verlag London, 2011.

[Fol15] G. B. Folland. *A Course in Abstract Harmonic Analysis*. 2.ª ed. Textbooks in Mathematics. Chapman y Hall/CRC, 2015.

[Fol99] G. B. Folland. *Real analysis: modern techniques and their applications*. 2.ª ed. Pure and Applied Mathematics. Wiley, 1999.

[Gar95] P. Garrett. *Buildings and Classical Groups*. 1995.

[GI63] O. Goldman y N. Iwahori. «The space of $\mathfrak{p}$-adic norms». En: *Acta Math.* 109 (1963).

[Gla03] E. Glasner. *Ergodic theory via joinings*. Mathematical Surveys and Monographs 101. American Mathematical Society, 2003.

[Hal50] P. R. Halmos. *Measure Theory*. 1.ª ed. Graduate Texts in Mathematics 18. Springer-Verlag New York, 1950.

[Hel78] S. Helgason. *Differential geometry, Lie groups, and symmetric spaces*. Pure and applied mathematics 80. Academic Press, 1978.

[IR90] K. Ireland y M. Rosen. *A Classical Introduction to Modern Number Theory*. 2.ª ed. Graduate Texts in Mathematics 84. Springer-Verlag New York, 1990.

[Iwa87] H. Iwaniec. «Fourier coefficients of modular forms of half-integral weight». En: *Inventiones Mathematicae* 87 (1987), págs. 385-401.

[Lan86] E. Landau. *Elementary number theory*. 2nd. Chelsea Publishing Company, 1986.

[Lee03] J. M. Lee. *Introduction to Smooth Manifolds*. 2.ª ed. Graduate Texts in Mathematics 218. Springer-Verlag New York, 2003.

[Lin68] Y. V. Linnik. *Ergodic properties of algebraic fields*. Grundlehren der Mathematischen Wissenschaften. New York: Springer-Verlag, 1968.

[Nac65] L. Nachbin. *The Haar integral*. 1st American Edition. 1965.

[Neu99] J. Neukirch. *Algebraic Number Theory*. Grundlehren der Mathematischen Wissenschaften. 1999.

[Per12] C. Perret-Gentil. «The correspondence between binary quadratic forms and quadratic fields». Tesis de mtría. École Polytechnique Fédérale de Lausanne, 2012.

[Rag72] M. S. Raghunathan. *Discrete subgroups of Lie groups*. Results and Problems in Cell Differentiation. Springer-Verlag, 1972.

[Ser73] J. P. Serre. *A course in arithmetic*. Graduate Texts in Mathematics 7. New York: Springer-Verlag, 1973.

[Sku62] B. F. Skubenko. «The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems». En: *Izv. Akad. Nauk SSSR Ser. Mat.* 26 (1962), págs. 721-752.

[Wal00] P. Walters. *Introduction to ergodic theory*. Springer-Verlag, 2000.