The Gross-Zagier formula: a brief introduction

David Hansen

March 28, 2011

The point of this talk is to give enough background to state the Gross-Zagier formula, and describe its immediate applications. I will prove almost nothing; the goal here is for you to see the formula and all its ingredients precisely. Time permitting, I will make some comments on the proof, and on more recent generalizations.

Let $\mathfrak{H} = \{x+iy \in \mathbf{C}, y > 0\}$ be the upper half-plane, and let $\Gamma_0(N) = \left\{\gamma \in \mathrm{SL}_2(\mathbf{Z}) | \gamma \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \bmod N\right\}$ act on the upper half-plane by linear fractional transformations. We may form the quotient $Y_0(N) = \Gamma_0(N)\backslash\mathfrak{H}$, a Riemann surface with finitely many cusps. Compactifying gives a curve $X_0(N)$ which is in fact defined over $\mathbf{Q}$; the map $z \to (j(z), j(Nz)) \in \mathbf{A}^2_{\mathbf{C}}$ realizes $Y_0(N)$ as a (highly singular) plane curve with $\mathbf{Q}$-coefficients. Over a general field k of characteristic zero, the k-points of the curve $X_0(N)$ (away from the cusps) parametrize diagrams $(\phi : E \to E')$ where E/k, E'/k are elliptic curves and $\phi : E \to E'$ is a k-rational isogeny with $\ker\phi \simeq \mathbf{Z}/N\mathbf{Z}$ over $\overline{k}$. There is a canonical $\mathbf{Q}$-rational involution $w_N : X_0(N) \to X_0(N)$ which sends the diagram $(\phi : E \to E')$ to the diagram $(\widehat{\phi} : E' \to E)$.

Over $\mathbf{C}$, elliptic curves are simply quotients $\mathbf{C}/\Lambda$ for lattices $\Lambda = \omega_1\mathbf{Z} + \omega_2\mathbf{Z} \subset \mathbf{C}$, $\omega_1/\omega_2 \notin \mathbf{R}$; the Weierstrass $\wp$-function

$$\wp_\Lambda(z) = \frac{1}{z^2} + \sum_{v \in \Lambda \smallsetminus \{0\}} \left(\frac{1}{(z-v)^2} - \frac{1}{v^2} \right)$$

yields an explicit uniformization of the corresponding curve via the map $z \to (1 : \wp'_\Lambda(z) : \wp_\Lambda(z)) \in \mathbf{P}^2_{\mathbf{C}}$. Dilating ω_1 and ω_2 by a common scalar λ yields an isomorphic curve, since $\wp_{\lambda\Lambda}(z) = \lambda^{-2}\wp_\Lambda(\lambda^{-1}z)$ so we may rescale by ω_1^{-1} and consider the lattices $\Lambda = \mathbf{Z} + \tau\mathbf{Z}$, assuming without any loss that $\mathrm{Im}\tau > 0$. Finally, two distinct points $\tau, \tau' \in \mathfrak{H}$ yield homothetic lattices if and only if one is a translate of the other by an element of $\mathrm{SL}_2(\mathbf{Z})$, so the space of elliptic curves over $\mathbf{C}$ is simply the quotient $\mathrm{SL}_2(\mathbf{Z})\backslash\mathfrak{H} = X_0(1)$. The $\mathbf{C}$-points of the covering $X_0(N) \to X_0(1)$ correspond to diagrams $(\mathrm{pr} : \mathbf{C}/\Lambda' \to \mathbf{C}/\Lambda)$ for lattices $\Lambda' \subset \Lambda$ with $[\Lambda : \Lambda'] = N$, and the covering map is just the forgetful map $\big(\mathrm{pr} : \mathbf{C}/(\mathbf{Z} + \tau\mathbf{Z}) \to \mathbf{C}/(\frac{1}{N}\mathbf{Z} + \tau\mathbf{Z})\big) \to \mathbf{C}/(\mathbf{Z} + \tau\mathbf{Z})$, $\tau \in \mathfrak{H}$. The involution w_N acts by $w_N(\tau) = \frac{-1}{N\tau}$.

There is a canonical construction of algebraic points on $X_0(N)$. Let $d < 0$ be a quadratic discriminant, and let $K = \mathbf{Q}(\sqrt{d})$ be an imaginary quadratic field with Hilbert class field H_K; class field theory yields a canonical isomorphism $\mathrm{Art}_K : \mathrm{Cl}(K) \xrightarrow{\sim} \mathrm{Gal}(H_K/K)$ mapping $[\mathfrak{p}]$ to $\mathrm{Frob}_{\mathfrak{p}}$. Suppose furthermore that every prime dividing N is split in K (this is the ubiquitous **Heegner hypothesis**). Then we may find some $\mathfrak{n} \subset \mathcal{O}_K$ with $\mathcal{O}_K/\mathfrak{n} \simeq \mathbf{Z}/N\mathbf{Z}$; there are $2^{\omega(N)}$ such $\mathfrak{n}$'s, where $\omega(N)$ is the number of distinct prime divisors of N. Then for any ideal $\mathfrak{a} \subset \mathcal{O}_K$, the diagram

$\left(\mathrm{pr} : \mathbf{C}/\mathfrak{a} \to \mathbf{C}/\mathfrak{n}^{-1}\mathfrak{a}\right)$ gives a point on $X_0(N)(\mathbf{C})$. Dilating $\mathfrak{a}$ by anything in $K^\times$ gives the same elliptic curve, so this construction only depends on the image of $\mathfrak{a}$ in the ideal class group $\mathrm{Cl}(K)$ of K. Hence we get a map

$$\begin{aligned} \gamma_{\mathfrak{n}} : \mathrm{Cl}(K) &\to X_0(N)(\mathbf{C}) \\ [\mathfrak{a}] &\mapsto \left(\mathrm{pr} : \mathbf{C}/\mathfrak{a} \to \mathbf{C}/\mathfrak{n}^{-1}\mathfrak{a}\right). \end{aligned}$$

These points are actually defined over H_K and satisfy the Galois-equivariance property $\mathrm{Art}_K(\mathfrak{p}) \cdot \gamma_{\mathfrak{n}}([\mathfrak{a}]) = \gamma_{\mathfrak{n}}([\mathfrak{p}\mathfrak{a}])$ for all $\mathfrak{p}$. These are the **Heegner points**. We can be even more explicit. When $N = 1$ the Heegner hypothesis is always satisfied, and we get the usual points $\gamma(\mathfrak{a}) = \frac{-b+\sqrt{d}}{2a} \in X_0(1)$ with $-a < b \le a$ and $b^2 - 4ac = d$ for some $c \ge a$. The choices of $\mathfrak{n}$ biject with the solutions $\beta \bmod 2N$ of $\beta^2 \equiv d \bmod 4N$ (exercise), and given β there is a unique $\Gamma_0(N)$-orbit of $\gamma(\mathfrak{a})$ containing a point $\frac{-B+\sqrt{d}}{2A}$ with $N|A$ and $B \equiv \beta \bmod 2N$. This is $\gamma_{\mathfrak{n}}(\mathfrak{a})$.

Now, let $E/\mathbf{Q}$ be an elliptic curve of conductor N, say $E : y^2 = x^3 + ax + b$ for some $a, b \in \mathbf{Z}$; the conductor is just some integer dividing the discriminant $\Delta = 16(4a^3 + 27b^2)$, which measures bad reduction in a "slightly more refined way" than Δ does (e.g. it only depemds on the isogeny class of E). The version of modularity which people *prove* is an isomorphism between two ℓ-adic Galois representations; more relevantly for us, modularity means there is a (unique) modular form $f_E(z) = \sum_{n=1}^{\infty} a_E(n) e^{2\pi i n z}$ of weight 2 and level N, such that $|E(\mathbf{F}_p)| = p + 1 - a_E(p)$ for all primes p. Set

$$\|f_E\|^2 = \int_{\Gamma_0(N)\backslash\mathfrak{H}} y^2 |f_E(z)|^2 d\mu = \int_{\Gamma_0(N)\backslash\mathfrak{H}} |f(z)|^2 dx dy$$

for later use; note that this is well-defined because $f(\gamma z) = (cz+d)^2$ and $\mathrm{Im}(\gamma z) = \frac{\mathrm{Im} z}{|cz+d|^2}$, and positive because f_E is holomorphic and thus can only vanish on a countable set. Being modular also implies there is a **modular parametrization** of E, a dominant morphism $\phi_E : X_0(N) \to E$ defined over $\mathbf{Q}$. This is very deep; it comes from the embedding $X_0(N) \to \mathrm{Jac}(X_0(N))$, a construction of Shimura which yields a modular elliptic curve E' as a quotient of $\mathrm{Jac}(X_0(N))$ which is modular and with $f_{E'} = f_E$, and Faltings's isogeny theorem. There are several choices of ϕ_E, but it becomes unique if we demand that $\phi_E(\infty) = 0$ and $\phi_E^*(d\omega) = 2\pi i c f_E(z) dz$ for some $c > 0$, where $d\omega = \frac{dx}{2y}$ is a translation-invariant 1-form. In fact, ϕ_E is given under these stipulations explicitly via

$$\phi_E(z) = -2\pi i c \int_z^{i\infty} f_E(\tau) d\tau.$$

Now, remember we have those Heegner points $\gamma_{\mathfrak{n}}(\mathfrak{a}) \in X_0(N)(H_K)$ parametrized by the ideal class group of K. Composing with the modular parametrization gives a point $P_{[\mathfrak{a}],\mathfrak{n}} := \phi_E(\gamma_{\mathfrak{n}}([\mathfrak{a}])) \in E(H_K)$. It turns out that changing $\mathfrak{n}$ changes all the $P_{[\mathfrak{a}],\mathfrak{n}}$'s by either nothing or by inversion, so I will henceforth *fix* $\mathfrak{n}$ permanently and drop it from my notation. Adding up over $[\mathfrak{a}]$ with respect to the group law gives a point

$$P_K = \sum_{[\mathfrak{a}] \in \mathrm{Cl}(K)} P_{[\mathfrak{a}]}$$

which is contained in $E(K)$; indeed, for any $\sigma \in \mathrm{Gal}(H_K/K)$, the action $P_{[\mathfrak{a}]} \to \sigma P_{[\mathfrak{a}]} = P_{\mathrm{Art}_K^{-1}(\sigma)[\mathfrak{a}]}$ simply permutes the ideal classes in the summation. The Gross-Zagier theorem describes the height of this point, in terms of an L-function.

The L-function of $E/\mathbf{Q}$ is

$$L(s,E_{/\mathbf{Q}}) := \prod_{p\nmid N} \frac{1}{1-a_E(p)p^{-s}+p^{1-2s}} \prod_{p\mid N} \frac{1}{1-a_E(p)p^{-s}} = \sum_{n=1}^{\infty} a_E(n)n^{-s}.$$

Modularity implies that this is holomorphic and that $\Lambda(s,E_{/\mathbf{Q}}) := (2\pi)^{-s}N^{s/2}\Gamma(s)L(s,E_{/\mathbf{Q}})$ satisfies $\Lambda(s,E_{/\mathbf{Q}}) = \pm\Lambda(2-s,E_{/\mathbf{Q}})$. This ±1 is the **root number** $\varepsilon(E_{/\mathbf{Q}})$**.** More generally, given $K/\mathbf{Q}$ as before, there is a unique quadratic Dirichlet character χ_d of period $|d|$ with $\zeta_K(s) = \zeta_{\mathbf{Q}}(s)L(s,\chi_d)$, and we define the twisted L-function

$$L(s,E^d_{/\mathbf{Q}}) = \sum_{n=1}^{\infty} a_E(n)\chi_d(n)n^{-s} = \prod_{p\nmid N} \frac{1}{1-a_E(p)\chi_d(p)p^{-s}+\chi_d(p)^2p^{1-2s}} \prod_{p\mid N} \frac{1}{1-a_E(p)\chi_d(p)p^{-s}}.$$

The notation is justified by the fact that this is the L-function of the curve $E^d : dy^2 = x^3+ax+b$. This satisfies the same functional equation, with N replaced by Nd^2, but with a different root number, namely $\varepsilon(E^d_{/\mathbf{Q}}) = \varepsilon(E_{/\mathbf{Q}})\chi_d(-N)$. Now set

$$L(s,E_{/K}) := L(s,E_{/\mathbf{Q}})L(s,E^d_{/\mathbf{Q}}).$$

The notation is again justified by the fact that

$$L(s,E_{/K}) = \prod_{\mathfrak{p}\subset\mathscr{O}_K,\,\mathfrak{p}\nmid N\mathrm{disc}K} \frac{1}{1-a_E(\mathfrak{p})\mathbf{N}\mathfrak{p}^{-s}+\mathbf{N}\mathfrak{p}^{1-2s}} \prod_{\mathfrak{p}\mid N\mathrm{disc}K} \cdots,$$

where $a_E(\mathfrak{p}) = \mathbf{N}\mathfrak{p}+1-|E(\mathscr{O}_K/\mathfrak{p})|$. What is the root number of this L-function? We compute

$$\begin{aligned}\varepsilon(E_{/\mathbf{Q}})\varepsilon(E^d_{/\mathbf{Q}}) &= \varepsilon(E_{/\mathbf{Q}})^2\chi_d(-N)\\ &= \chi_d(-N)\\ &= \chi_d(-1)\chi_d(N)\\ &= -1,\end{aligned}$$

since $d<0$ and all the primes dividing N are split in K. This forces $L(1,E_{/K}) = 0$, and the Gross-Zagier formula computes $L'(1,E_{/K})$ as the value of a **height function**.

Given a finite extension $k/\mathbf{Q}$, let M_k be the set of all places of k and let $|\cdot|_v$ be the corresponding normalized valuation on k_v, i.e. $|x|_v = q_v^{-\mathrm{val}_v(x)}$ where q_v is the cardinality of the residue field of k_v, and $\mathrm{val}_v(\varpi_v) = 1$ on a uniformizer. We have the product formula $\prod_{v\in M_k}|x|_v = 1\,\forall x\in k$. For a point $x = (x_0:x_1:x_2)\in\mathbf{P}^2(k)$, define the height

$$h_k(x) = \frac{1}{[k:\mathbf{Q}]}\log\left(\prod_{v\in M_k}\max\{|x_0|_v,|x_1|_v,|x_2|_v\}\right).$$

Note that this is well-defined on projective space (by the product formula) and is nonnegative; the second property follows from $\prod_i\max\{a_i,b_i,\dots\}\geq\max\{\prod_i a_i,\prod_i b_i,\dots\}$ and the product formula. Note also that $h_{k'}(x) = h_k(x)$ if $k\subset k'$, so the "direct limit"

$$h(x) = \varinjlim_{k} h_k(x)$$

is well-defined on $\mathbf{P}^2(\overline{k})$. Given an elliptic curve $E \subset \mathbf{P}^2$ defined over k, and a point $P \in E(\overline{k})$, define the **canonical height**

$$h_E(P) = \lim_{n\to\infty} \frac{h(n \cdot P)}{n^2}.$$

Neron and Tate showed that this limit is well-defined, that $h_E(P)$ is a quadratic form on $E(\overline{k})$, and that $h_E(P) = 0$ if and only if $P \in E(\overline{k})_{\mathrm{tors}}$.

Theorem (Gross and Zagier). *With the above notation and assumptions, we have*

$$L'(1, E_{/K}) = \frac{32\pi^2 \left\| f_E \right\|^2}{|\mathscr{O}_K^\times|^2 \sqrt{|d|} \deg \phi_E} h_E(P_K).$$

In particular,

$$L'(1, E_{/K}) = 0 \iff P_K \,\text{is torsion in}\, E(K).$$

(Note that $\frac{h_E(x)}{\deg \phi_E}$ is an isogeny invariant.) Gross and Zagier deduce several amazing corollaries from this. Let's start with the best one.

Corollary A. *If $E/\mathbf{Q}$ is an elliptic curve with root number $\varepsilon = \varepsilon(E_{/\mathbf{Q}}) = -1$, and $L'(1, E_{/\mathbf{Q}}) \neq 0$, then $E(\mathbf{Q})$ contains elements of infinite order.*

Proof sketch. This is not explained very well anywhere, so let me try. First, by a deep theorem of Waldspurger, we may find some K satisfying the Heegner hypothesis with $L(1, E^d_{/\mathbf{Q}}) \neq 0$. Thus $L'(1, E_{/K}) = L'(1, E_{/\mathbf{Q}})L(1, E^d_{/\mathbf{Q}}) \neq 0$, so $P_K \in E(K)$ is nontorsion. Next, we need to understand the action of complex conjugation on the individual $P_{[\mathfrak{a}]}$'s. We shall do this by using the relations $w_N \cdot \gamma_{\mathfrak{n}}(\mathfrak{a}) = \gamma_{\overline{\mathfrak{n}}}(\mathfrak{a}\mathfrak{n}^{-1})$ and $\overline{\gamma_{\mathfrak{n}}(\mathfrak{a})} = \gamma_{\overline{\mathfrak{n}}}(\overline{\mathfrak{a}})$ on $X_0(N)$ together with the following lemma.

Lemma A.1. *If $E/\mathbf{Q}$ has root number ε, then for any $z \in X_0(N)(\mathbf{C})$, the point $\phi_E(z) + \varepsilon\phi_E(w_N \cdot z)$ is independent of z, and is torsion in $E(\mathbf{C})$.*

Proof. Let $f = f_E$ be the newform corresponding to E, and write $\omega_f = 2\pi i c f(z)dz$ where c is the Manin constant. By the Manin-Drinfeld theorem, the point

$$\phi_E(0) = -\int_0^{i\infty} \omega_f$$

is torsion. On the other hand, we compute

$$\begin{aligned} \int_0^{i\infty} \omega_f &= \int_z^{i\infty} \omega_f + \int_0^z \omega_f \\ &= \int_z^{i\infty} \omega_f + \int_{w_N 0}^{w_N z} w_N\omega_f \\ &= \int_z^{i\infty} \omega_f - \int_{w_N z}^{i\infty} w_N\omega_f. \end{aligned}$$

By newform theory, we know that $f(-1/Nz) = -\varepsilon z^2 N f(z)$, and $d(-1/Nz)/dz = N^{-1}z^{-2}$, so $w_N\omega_f = -\varepsilon\omega_f$. Thus $\phi_E(0) = \phi_E(z) + \varepsilon\phi_E(w_N z)$ for all $z \in X_0(N)(\mathbf{C})$.

Applying the lemma with $z = \gamma_{\overline{\mathfrak{n}}}(\overline{\mathfrak{a}})$, and noting further that $\phi_E(\overline{z}) = \overline{\phi_E(z)}$, we compute

$$\begin{aligned} \text{tors.} &= \overline{\phi_E(\gamma_{\mathfrak{n}}(\mathfrak{a}))} + \varepsilon\phi_E(w_N \cdot \gamma_{\overline{\mathfrak{n}}}(\overline{\mathfrak{a}})) \\ &= \overline{\phi_E(\gamma_{\mathfrak{n}}(\mathfrak{a}))} + \varepsilon\phi_E(\gamma_{\mathfrak{n}}(\overline{\mathfrak{a}\mathfrak{n}^{-1}})) \\ &= \overline{P_{[\mathfrak{a}]}} + \varepsilon P_{[\mathfrak{a}^{-1}\mathfrak{n}]} \\ &= \overline{P_{[\mathfrak{a}]}} + \varepsilon \mathrm{Art}_K(\mathfrak{a}^{-2}\mathfrak{n}) \cdot P_{[\mathfrak{a}]}. \end{aligned}$$

Since $\mathfrak{a}$ was arbitrary, we conclude that *if* $\widetilde{\tau} \in \mathrm{Gal}(H_K/\mathbf{Q}) = \mathrm{Gal}(H_K/K) \rtimes \mathrm{Gal}(K/\mathbf{Q})$ *acts on* K *nontrivially, then for any fixed* $\mathfrak{a}$*, there is some* $\sigma \in \mathrm{Gal}(H/K)$ *(depending on* $\mathfrak{a}$ *and* $\widetilde{\tau}$*!) such that* $\widetilde{\tau}P_{[\mathfrak{a}]} + \varepsilon\sigma P_{[\mathfrak{a}]}$ *is torsion.* Adding up the $\mathrm{Gal}(H/K)$-translates of this, we find that

$$\begin{aligned} \sum_{\rho \in \mathrm{Gal}(H/K)} \rho\widetilde{\tau}P_{[\mathfrak{a}]} + \varepsilon\rho\sigma P_{[\mathfrak{a}]} &= \sum_{\rho \in \mathrm{Gal}(H/K)} \widetilde{\tau}P_{\mathrm{Art}_K(\rho)[\mathfrak{a}]} + \varepsilon P_{\mathrm{Art}_K(\sigma\rho)[\mathfrak{a}]} \\ &= \overline{P_K} + \varepsilon P_K \end{aligned}$$

is torsion. By the parallelogram law for quadratic forms,

$$\begin{aligned} h_E(\overline{P_K} - \varepsilon P_K) + h_E(\overline{P_K} + \varepsilon P_K) &= 2h_E(P_K) + 2h_E(\overline{P_K}) \\ &= 4h_E(P_K) \\ &> 0 \end{aligned}$$

so $\overline{P_K} - \varepsilon P_K \in E(K)$ is nontorsion and is defined over $\mathbf{Q}$ iff $\varepsilon = -1$.

Corollary B. *If* $L(1, E_{/\mathbf{Q}}) \neq 0$ *and* P_K *is torsion for some* K*, then* $L(s, E^d_{/\mathbf{Q}})$ *vanishes to order at least 3 at* $s = 1$.

For example, this happens for the curve $E : y^2 = x^3 + 10x^2 - 20x + 8$ (of conductor 37) and $d = -139$. In this particular case, E^{-139} provably has algebraic rank 3, and $L(s, E_{/K})$ provably vanishes to order exactly 3 at $s = 1$.

Corollary C (Goldfeld). *There is an effective, computable constant* $c > 0$ *such that the class number of an imaginary quadratic field* $\mathbf{Q}(\sqrt{-d})$ *satisfies*

$$\begin{aligned} |\mathrm{Cl}(\mathbf{Q}(\sqrt{-d}))| &> c \log d \cdot \exp(-21\sqrt{\log\log d}) \\ &\gg_\varepsilon (\log d)^{1-\varepsilon}. \end{aligned}$$

By the way, how did Heegner's name get attached to these points? He used a proto-version of them to show, among other things, that the curve

$$py^2 = x^3 - x$$

has rational points of infinite order when p is a prime with $p \equiv 5 \,\text{or}\, 7 \bmod 8$.