A (very brief) History of the Trace Formula

James Arthur

This note is a short summary of a lecture in the series celebrating the tenth anniversary of PIMS. The lecture itself was an attempt to introduce the trace formula through its historical origins. I thank Bill Casselman for suggesting the topic. I would also like to thank Peter Sarnak for sharing his historical insights with me. I hope I have not distorted them too grievously.

As it is presently understood, the trace formula is a general identity

$$\text{(GTF)} \qquad \sum\{\text{geometric terms}\} = \sum\{\text{spectral terms}\}.$$

The spectral terms contain arithmetic information of a fundamental nature. However, they are highly inaccessible, "spectral" actually, in the nonmathematical meaning of the word. The geometric terms are quite explicit, but they have the drawback of being very complicated.

There are simple analogues of the trace formula, "toy models" one could say, which are familiar to all. For example, suppose that $A = (a_{ij})$ is a complex $(n \times n)$-matrix, with diagonal entries $\{u_i\} = \{a_{ii}\}$ and eigenvalues $\{\lambda_j\}$. By evaluating its trace in two different ways, we obtain an identity

$$\sum_{i=1}^{n} u_i = \sum_{j=1}^{n} \lambda_j.$$

The diagonal coefficients obviously carry geometric information about A as a transformation of $\mathbb{C}^n$. The eigenvalues are spectral, in the precise mathematical sense of the word.

For another example, suppose that $g \in C_c^\infty(\mathbb{R}^n)$. This function then satisfies the Poisson summation formula

$$\sum_{u \in \mathbb{Z}^n} g(u) = \sum_{\lambda \in 2\pi i \mathbb{Z}^n} \hat{g}(\lambda),$$

where

$$\hat{g}(\lambda) = \int_{\mathbb{R}^n} g(x) e^{-x\lambda} \mathrm{d}x, \qquad \lambda \in \mathbb{C}^n,$$

is the Fourier transform of g. One obtains an interesting application by letting $g = g_T$ approximate the characteristic function of the closed ball B_T of radius T about the origin. As T becomes large, the left hand side approximates the number of lattice points $u \in \mathbb{Z}^n$ in B_T. The dominant term on the right hand side is the integral

$$\hat{g}(0) = \int_{\mathbb{R}^n} g(x)\mathrm{d}x,$$

which in turn approximates $\mathrm{vol}(B_T)$. In this way, the Poisson summation formula leads to a sharp asymptotic formula for the number of lattice points in B_T.

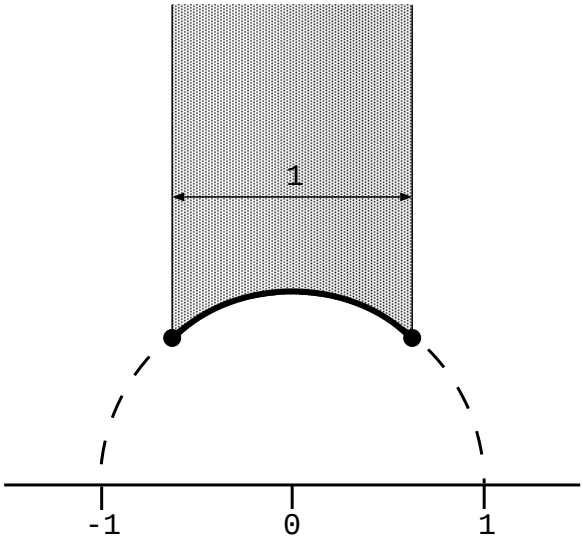
Our real starting point is the upper half plane

$$H = \big\{z \in \mathbb{C} : \ \mathrm{Im}(z) > 0\big\}.$$

The multiplicative group $SL(2,\mathbb{R})$ of (2×2) real matrices of determinant 1 acts transitively by linear fractional transformations on H. The discrete subgroup

$$\Gamma = SL(2,\mathbb{Z})$$

acts discontinuously. Its space of orbits $\Gamma\backslash H$ can be identified with a noncompact Riemann surface, whose fundamental domain is the familiar modular region.



More generally, one can take Γ to be a congruence subgroup of $SL(2,\mathbb{Z})$, such as the group

$$\Gamma(N) = \big\{\gamma \in SL(2,\mathbb{Z}) : \gamma \equiv I (\mathrm{mod}\, N)\big\}.$$

The space $\Gamma\backslash H$ comes with the hyperbolic metric

$$\mathrm{d}s^2 = \frac{\mathrm{d}x\,\mathrm{d}y}{y^2}$$

and the hyperbolic Laplacian

$$\Delta = -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right).$$

Modular forms are holomorphic sections of line bundles on $\Gamma\backslash H$. For example, a modular form of weight 2 is a holomorphic function $f(z)$ on H such that the product

$$f(z)\mathrm{d}z$$

descends to a holomorphic 1-form on the Riemann surface $\Gamma\backslash H$. The classical theory of modular forms was a preoccupation of a number of prominent nineteenth century mathematicians. It developed many strands, which intertwine complex analysis and number theory.

In the first half of the twentieth century, the theory was taken to new heights by *E. Hecke* ($\sim$ 1920–1940)[1]. Among many other things, he introduced the notion of a *cusp form*. As objects that are rapidly decreasing at infinity, cusp forms represent holomorphic eigensections of Δ (for the relevant line bundle) that are square integrable on $\Gamma\backslash H$.

The notion of an eigenform of Δ calls to mind the seemingly simpler problem of describing the spectral decomposition of Δ on the space of functions $L^2(\Gamma\backslash H)$. I do not know why this problem, which seems so natural to our modern tastes, was not studied earlier. Perhaps it was because eigenfunctions of Δ are typically not holomorphic. Whatever the case, major advances were made by *A. Selberg*. I will attach his name to the first of three sections, which roughly represent three chronological periods in the development of the trace formula.

I. Selberg

(a) **Eisenstein series for** $\Gamma\backslash H$ ($\sim$ 1950).

Eisenstein series represent the continuous spectrum of Δ on the noncompact space $\Gamma\backslash H$. In case $\Gamma = SL(2,\mathbb{Z})$, they are defined by infinite series

$$E(\lambda, z) = \sum_{(c,d)=1} \frac{(\operatorname{Im} z)^{\frac{1}{2}(\lambda+1)}}{|cz+d|^{\lambda+1}}, \qquad z \in \Gamma\backslash H,\ \lambda \in \mathbb{C},$$

that converge if $\operatorname{Re}(\lambda) > 1$. Selberg[2] introduced general techniques, which showed that $E(\lambda, z)$ has analytic continuation to a meromorphic function of $\lambda \in \mathbb{C}$, that its values at $\lambda \in i\mathbb{R}$ are analytic, and that these values exhaust the continuous spectrum of Δ on $L^2(\Gamma\backslash H)$. One can say that the function

$$E(\lambda, z), \qquad \lambda \in i\mathbb{R},\ z \in \Gamma\backslash H,$$

plays the same role for $L^2(\Gamma\backslash H)$ as the function $e^{\lambda x}$ in the theory of Fourier transforms.

(b) **Trace formula for** $\Gamma\backslash H$ ($\sim$ 1955).

Selberg's analysis of the continuous spectrum left open the question of the discrete spectrum of Δ on $L^2(\Gamma\backslash H)$. About this time, examples of square integrable eigenfunctions of Δ were constructed separately (and by very different means) by *H. Maass* and *C.L. Siegel*. Were these examples isolated anomalies, or did they represent only what was visible of a much richer discrete spectrum?

A decisive answer was provided by the trace formula Selberg created to this end. The Selberg trace formula is an identity

$$\text{(STF)} \qquad \sum_i a_i g(u_i) = \sum_j b_j \hat{g}(\lambda_j) + e(g),$$

where g is any symmetric test function in $C_c^\infty(\mathbb{R})$, $\{u_i\}$ are essentially[3] the real eigenvalues of conjugacy classes in Γ, and $\{\lambda_i\}$ are essentially[3] the discrete eigenvalues of Δ on $L^2(\Gamma\backslash H)$. The coefficients $\{a_i\}$ and $\{b_j\}$ are explicit nonzero constants, and $e(g)$ is an explicit error term (which contains both geometric and spectral data). The proof of (STF) was a tour de force. The function g gives rise to an operator on $L^2(\Gamma\backslash H)$, but the presence of a continous spectrum means that the operator is not of trace class. Selberg had first to subtract the contribution of this operator to the continuous spectrum, something he could in principle do by virtue of (a). However, the modified operator is quite complicated. It is remarkable that Selberg was able to express its trace by such a relatively simple formula.

Selberg's original application of (STF) came by choosing g so that $\hat{g}$ approximated the characteristic function of a large symmetric interval in $\mathbb{R}$. The result was a sharp asymptotic formula

$$\left|\left\{\Lambda_j = \tfrac{1}{4} - \lambda_j^2 \leq T\right\}\right| \ \sim \ \frac{\pi}{2}\text{vol}(\Gamma\backslash H)T$$

for the number of eigenvalues Λ_j in the discrete spectrum. This is an analogue of Weyl's law (which applies to compact Riemannian manifolds) for the noncompact manifold $\Gamma\backslash H$. In particular, it shows that the congruence arithmetic quotient $\Gamma\backslash H$ has a rich discrete spectrum, something subsequent experience has shown is quite unusual for noncompact Riemannian manifolds.

Ramifications ($\sim$ 1955–1960).

(i) Selberg seems to have observed *after* his discovery of (STF) that a similar but simpler formula could be proved for any *compact* Riemann

surface $\Gamma'\backslash H$. (and indeed, for any compact, locally symmetric space). For example, one could take the fundamental group Γ' to be a congruence group inside a quaternion algebra Q over $\mathbb{Q}$ with $Q(\mathbb{R}) \cong M_2(\mathbb{R})$. The trace formula in this case is similar to (STF), except that the explicit error term $e(g)$ is considerably simpler.

(ii) Selberg also observed that (STF) could be extended to the *Hecke operators*

$$\{T_p : \ p \text{ prime}\}$$

on $L^2(\Gamma\backslash H)$. These operators have turned out to be the most significant of Hecke's many contributions. They are a commuting family of operators, parametrized by prime numbers p, which also commute with Δ. The corresponding family of simultaneous eigenvalues $\{t_{p,j}\}$ carries arithmetic information. They can be regarded as the analytic embodiment of data that govern fundamental arithmetic phenomena. Selberg's generalization of (STF) includes terms on the right hand side that quantify the numbers $\{t_{p,j}\}$. It also holds more generally if $L^2(\Gamma\backslash H)$ is replaced by the space of square integrable sections of a line bundle on $\Gamma\backslash H$. In this form, it can be applied to the space of classical cusp forms of weight $2k$ on $\Gamma\backslash H$. It yields a finite closed formula for the trace of any Hecke operator on this space.

(iii) Selberg also studied generalizations of Eisenstein series and (STF) to some spaces of higher dimension.

II. Langlands

(a) **General Eisenstein series** ($\sim$ 1960–1965).

Motivated by Selberg's results, *R. Langlands* set about constructing continuous spectra for *any* locally symmetric space $\Gamma\backslash X$ of finite volume. Like the special case $\Gamma\backslash H$, the problem is to show that absolutely convergent Eisenstein series have analytic continuation to meromorphic functions, whose values at imaginary arguments exhaust the continuous spectrum. The analytic difficulties were enormous. Langlands was able to overcome them with a remarkable argument based on an interplay between spectral theory and higher residue calculus. The result was a complete description of the continuous spectrum of $L^2(\Gamma\backslash X)$ in terms of discrete spectra for spaces of smaller dimension.

(b) **Comparison of trace formulas** ($\sim$ 1970–1975).

Langlands changed the focus of applications of the trace formula. Instead of taking one formula in isolation, he showed how to establish deep results by comparing two trace formulas with each other. He treated three different kinds of comparison, following special cases that

had been studied earlier by *M. Eichler* and *H. Shimizu*, *Y. Ihara*, and *H. Saito* and *T. Shintani*. I shall illustrate each of these in shorthand, with a symbolic correpondence between associated data for which the comparison yields a reciprocity law. In each case, the left hand side represents some form of the trace formula (STF), while the right hand side represents another trace formula.

(i)
$$\begin{aligned}(\Gamma\backslash H) &\leftrightarrow (\Gamma'\backslash H)\\ \{\lambda_j, t_{p,j}\} &\leftrightarrow \{\lambda'_j, t'_{p,j}\}.\end{aligned}$$

Here $\Gamma'\backslash H$ represents a compact Riemann surface attached to a congruence quaternion group Γ'. The reciprocity law, established by Langlands in collaboration with *H. Jacquet*, is a remarkable correspondence between spectra of Laplacians on two Riemann surfaces, one noncompact and the other compact, and also a correspondence between eigenvalues of associated Hecke operators.

(ii)
$$\begin{aligned}(\Gamma\backslash H) &\leftrightarrow (\Gamma\backslash H)_p\\ \{t_{p,j}\} &\leftrightarrow \{\Phi_{p,j}\}.\end{aligned}$$

Here $(\Gamma\backslash H)_p$ represents an algebraic curve over $\mathbb{F}_p$, obtained by reduction $\bmod p$ of a $\mathbb{Z}$-scheme associated to $\Gamma\backslash H$. The relevant trace formula is the Grothendieck-Lefschetz fixed point formula, and $\{\Phi_{p,j}\}$ represent eigenvalues of the Frobenius endomorphism on the ℓ-adic cohomology of $(\Gamma\backslash H)_p$. The reciprocity law illustrated in this case gives an idea of the arithmetic significance of eigenvalues $\{t_{p,j}\}$ of Hecke operators

(iii)
$$\begin{aligned}(\Gamma\backslash H) &\leftrightarrow (\Gamma_E\backslash H_E)\\ \{\lambda_j, t_{p,j}\} &\leftrightarrow \{\lambda_{E,j}, t_{\mathfrak{p},j}\}.\end{aligned}$$

Here, $(\Gamma_E\backslash H_E)$ is a higher dimensional locally symmetric space attached to a cyclic Galois extension $E/\mathbb{Q}$, and $\mathfrak{p}$ denotes a prime ideal in $\mathcal{O}_E$ over p. The relevant formula is a twisted trace formula, attached to the diffeomorphism of $\Gamma_E\backslash H_E$ defined by a generator of the Galois group of E/F. The reciprocity law it yields (and its generalization with $\mathbb{Q}$ replaced by an arbitrary number field F) is known as cyclic base change. It has had spectacular consequences. It led to the proof of a famous conjecture of *E. Artin* on representations of Galois groups, in the special case of a two dimensional representation of a solvable Galois group. This result, known as the Langlands-Tunnell theorem, was in turn a starting point for the work of *A. Wiles* on the Shimura-Taniyama-Weil conjecture and his proof of Fermat's last theorem.

My impressionistic review of the three kinds of comparison is not to be taken too literally. For example, it is best not to fix the congruence

subgroup Γ of $SL(2,\mathbb{R})$. The correspondences are really between a (topological) projective limit

$$\varprojlim_{\Gamma}(\Gamma\backslash H)$$

and its three associated analogues. Moreover, the group $SL(2)$ should actually be replaced by $GL(2)$. Nevertheless, the basic idea is as stated, to compare a formula like (STF) with something else. One deduces relations between data on the spectral sides from a priori relations between data on the geometric sides. We recall that the geometric terms in (STF) are indexed by conjugacy classes in the discrete group Γ.

Before going to the next stage, I need to recall some other foundational ideas of Langlands. To maintain a sense of historical flow, I shall divide these remarks artificially into two time periods.

Between II(a) and II(b) ($\sim$ 1965–1970).

During this period, Langlands formulated the conjectures that came to be known as the *Langlands programme*. Many of these are subsumed in his *principle of functoriality*. This grand conjecture consists of a collection of very general, yet quite precise, relations among spectral data $\{\lambda_j, t_{p,j}\}$ attached to arbitrary locally symmetric spaces $\Gamma\backslash X$ (of congruence type). It also includes striking relations between these data and arithmetic data attached to finite dimensional, complex representations of Galois groups.

Among other things, Langlands' ideas altered definitively the language of modular forms (and its generalizations). He formulated his conjectures in terms of the adeles, a locally compact ring

$$\mathbb{A} = \mathbb{R} \times \prod_{p}^{\text{rest}} \mathbb{Q}_p,$$

which contains $\mathbb{Q}$ diagonally as a discrete subring. This point of view itself has an interesting history, which went through a series of refinements with *C. Chevalley*, *J. Tate*, *I. Gelfand* and *T. Tamagawa*. In the present setting, the basic observation is that there are natural isomorphisms

$$\begin{aligned} L^2\big(SL(2,\mathbb{Z})\backslash H\big) &\cong L^2\big(SL(2,\mathbb{Z})\backslash SL(2,\mathbb{R}/SO(2,\mathbb{R})\big) \\ &\cong L^2\big(SL(2,\mathbb{Q}\backslash SL(2,\mathbb{A})/SO(2,\mathbb{R})K_0\big), \end{aligned}$$

for the compact subgroup

$$K_0 = \prod_p SL(2, \mathbb{Z}_p)$$

of $SL(2, \mathbb{A})$. A removal of K_0 from the last quotient causes the first space to be replaced by a direct limit

$$\varinjlim_{\Gamma} L^2(\Gamma \backslash H) = L^2\Big(\varprojlim_{\Gamma} (\Gamma \backslash H) \Big).$$

If one removes $SO(2, \mathbb{R})$ from the quotient, one obtains a Hilbert space that includes the square integrable sections of line bundles that define classical cusp forms. Thus, the classical objects we have discussed can all be combined together into the single Hilbert space of square integrable functions on $SL(2, \mathbb{Q}) \backslash SL(2, \mathbb{A})$.

To treat the general case of spaces of higher dimension, one simply replaces $SL(2)$ by a general reductive algebraic group G over $\mathbb{Q}$. One then studies the irreducible decompositon of the representation of $G(\mathbb{A})$ by right translation on the Hilbert space

$$\mathcal{H} = L^2\big(G(\mathbb{Q}) \backslash G(\mathbb{A})\big).$$

Irreducible representations of $G(\mathbb{A})$ obtained in this way are known as *automorphic representations*. They carry all the information contained in the spectral decomposition.

After II(b) ($\sim$ 1975–1985).

Having established striking results by comparing the trace formula for $GL(2)$ with three other trace formulas, Langlands gave careful thought to what might happen in general. There was no general trace formula, at least initially, but it was still possible to make predictions. The result was Langlands' conjectural *theory of endoscopy*. This theory offers a general strategy for comparing trace formulas attached to arbitrary reductive groups G. It is founded largely on conjugacy classes, both in $G(\mathbb{Q})$ and any of its completions $G(\mathbb{Q}_v) \in \{G(\mathbb{R}), G(\mathbb{Q}_p)\}$. The theory is based on the critical observation that elements in $G(\mathbb{Q})$ (or $G(\mathbb{Q}_v)$) need not be conjugate even if they are conjugate over the algebraic closure $G(\bar{\mathbb{Q}})$ (or $G(\bar{\mathbb{Q}}_v)$). This phenomenon is absent in the special case $G = GL(2)$, but it would obviously be an essential consideration in any general comparison of geometric terms in trace formulas. The theory of endoscopy represents a precise measure, in both geometric and spectral terms, of the failure of geometric conjugacy to imply conjugacy.

III. Arthur[4]

(a) **The general trace formula** ($\sim$ 1975–1985).

One takes G to be a reductive group over $\mathbb{Q}$, as above. Any function $f \in C_c^\infty\big(G(\mathbb{A})\big)$ then provides a convolution operator $R(f)$ on the Hilbert space $\mathcal{H} = L^2\big(G(\mathbb{Q})\backslash G(\mathbb{A})\big)$, which in turn has an orthogonal decomposition

$$R(f) = R_{\text{disc}}(f) \oplus R_{\text{cont}}(f),$$

relative to the discrete and continuous spectra. The general trace formula (GTF) is a formula for the trace[5] of the operator $R_{\text{disc}}(f)$. It can be written as a sum of relatively simple terms, indexed by $\mathbb{Q}$-elliptic conjugacy classes in $G(\mathbb{Q})$, with more complicated "error" terms. The error terms come from hyperbolic conjugacy classes in $G(\mathbb{Q})$, which are parametrized by elliptic conjugacy classes in Levi subgroups M of G, and continuous spectra, which are parametrized by discrete spectra of Levi subgroups.

(b) **Endoscopy for classical groups** ($\sim$ 1995–present).

The problem is to classify automorphic representations of classical groups G (such as the split groups $SO(2n+1)$, $Sp(2n)$ and $SO(2n)$) in terms of automorphic representations of general linear groups $\widetilde{G} = GL(N)$. In the symbolic shorthand of II(b), the comparison takes the form

$$\begin{aligned} G(\mathbb{Q})\backslash G(\mathbb{A}) &\leftrightarrow \widetilde{G}(\mathbb{Q})\backslash \widetilde{G}(\mathbb{A}), \\ \{\lambda_j, t_{p,j}\} &\leftrightarrow \{\widetilde{\lambda}_j, \widetilde{t}_{p,j}\}. \end{aligned}$$

However, the situation here is more subtle than that of II(b). On the left, one has to take the stable trace formula for G, a refinement of the ordinary trace formula that compensates for the failure of geometric conjugacy to imply ordinary conjugacy. One also has to treat several G together, taking appropriate linear combinations of terms in their stable trace formulas. On the right, one takes the twisted trace formula of $\widetilde{G}$, relative to the standard outer automorphism $x \to {}^t x^{-1}$.

Despite the difficulties, it appears that this comparison of trace formulas will lead to precise information about automorphic representations of classical groups. I mention three of what are likely to be many applications.

(i) A classification of the automorphic representations of the split classical groups G ought to lead to a sharp analogue of Weyl's law[6] for the associated noncompact symmetric spaces

$$X_\Gamma = \Gamma\backslash X = \Gamma\backslash G(\mathbb{R})/K_{\mathbb{R}}.$$

(ii) In cases that X_Γ has a complex structure (such as for $G = GSp(2n)$), the classification gives important information about the L^2-cohomology $H^*_{(2)}(X_\Gamma)$. It leads to a decomposition of $H^*_{(2)}(X_\Gamma)$ that clearly exhibits the Hodge structure, the cup product action of a Kähler class, and the action of Hecke operators.

(iii) The theory of endoscopy for classical groups includes some significant cases of functoriality. It also places automorphic L-functions of classical groups on a par with those of $GL(N)$.

IV. The Future

(a) **Principle of functoriality** (2007–?).

Many cases of the principle of functoriality lie well beyond what is implied by the theory of endoscopy (which itself is still conjectural in general). Langlands has recently proposed a strategy for applying the trace formula (GTF) to the general principle of functoriality. The proposal includes a comparison of trace formulas that is completely different than anything attempted before. It remains highly speculative, and needless to say, is completely open.

(b) **Motives and automorphic representations** (2007–?).

As conceived by *A. Grothendieck*, motives are the essential building blocks of algebraic geometry. If one thinks of algebraic varieties (say, projective and nonsingular) as the basic objects of everyday life, motives represent the elementary particles. In a far-reaching generalization of the Shimura-Taniyama-Weil conjecture, Langlands has proposed a precise reciprocity law between general motives and automorphic representations. It amounts to a description of arithmetic data that characterize algebraic varieties in terms of eigenvalues $\{t_{p,j}\}$ of Hecke operators attached to general groups G. This conjecture is again completely open. It appears to be irrevocably intertwined with the general principle of functoriality.

Footnotes

1. *These dates, like others that follow, are not to be taken too literally. They are my attempt to approximate the relevant period of activity, and to orient the reader to the development of the subject.*
2. *These results were actually first established by H. Maass, whose work was later applied to more general discrete subgroups of* $SL(2,\mathbb{R})$ *by W. Roelke. However, Selberg's techniques have been more influential, having shown themselves to be amenable to considerable generalization.*

3. *For example, the eigenvalues* $\{\Lambda_j\}$ *are related to the numbers* $\{\lambda_j\}$ *by the formula* $\Lambda_j = \frac{1}{4} - \lambda_j^2$.
4. *I was following a suggestion to divide the history of the trace formula into three periods of development, indexed by three names!*
5. *The proof that* $R_{\mathrm{disc}}(f)$ *is of trace class is due to W. Muller.*
6. *A general noncompact form of Weyl's law has been established recently by E. Lindenstrauss* and *A. Venkatesh. In the case of classical groups above, the goal would be to establish the strongest possible error term.*