Formalising Half of a Graduate Textbook on Number Theory (Short Paper)

Manuel Eberl
University of Innsbruck, Austria

Anthony Bordg
Université Paris-Saclay, INRIA, CNRS, ENS Paris-Saclay, Laboratoire Méthodes Formelles, France

Lawrence C. Paulson
University of Cambridge, United Kingdom

Wenda Li
University of Edinburgh, United Kingdom

Abstract

Apostol's *Modular Functions and Dirichlet Series in Number Theory* [2] is a graduate text covering topics such as elliptic functions, modular functions, approximation theorems and general Dirichlet series. It relies on complex analysis, winding numbers, the Riemann ζ function and Laurent series. We have formalised several chapters and can comment on the sort of gaps found in pedagogical mathematics. Proofs are available from `https://github.com/Wenda302/Number_Theory_ITP2024`.

2012 ACM Subject Classification Theory of computation → Logic and verification

Keywords and phrases Isabelle/HOL, number theory, complex analysis, formalisation of mathematics

Digital Object Identifier 10.4230/LIPIcs.ITP.2024.31

Category Short Paper

Supplementary Material *Software (Isabelle/HOL files)*: `https://doi.org/10.5281/zenodo.12586104` [5]

Funding ERC Advanced Grant ALEXANDRIA (Project 742178)

1 Introduction

Number theory is an ideal testbed for techniques in the formalisation of mathematics: it is central to mathematics, as many Fields medals attest, and its analytic branch requires the deployment of complex analysis and approximation theory.

Apostol's popular textbook series is a good choice of source material. His *Modular Functions and Dirichlet Series* [2] follows on from his *Introduction to Analytic Number Theory* [1], most of which has already been formalised in Isabelle/HOL [4]. By formalising both volumes we create a good basis for formalising further work in analytic number theory, while at the same time investigating Apostol's actual text forensically.

Isabelle/HOL [7] is a popular proof assistant. Based on simple type theory, its advantages include best-in-class automation, a library of over four million lines of formal proofs, and a structured proof language offering a good degree of legibility. Users work within a highly sophisticated interactive development environment.

We report on ongoing work to formalise the book and build a foundation of modular forms in Isabelle/HOL. We explore the chapters that we formalised fully (1, 2, 3, 7) and the parts of Chapter 6 that were already completed, commenting on what was covered and where we had issues with the text. Except for one technical lemma that we did not need, all theorems from these chapters have been formalised. In particular, all results mentioned in this paper have been formalised.

15th International Conference on Interactive Theorem Proving (ITP 2024).
Editors: Yves Bertot, Temur Kutsia, and Michael Norrish; Article No. 31; pp. 31:1–31:7
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Prerequisites: holomorphicity, analyticity, meromorphicity

A complex function is called *holomorphic* (or *analytic*) on an open set $A \subseteq \mathbb{C}$ if its derivative exists at every point of A. In the Isabelle library, these notions are defined not only for open sets and here they do not coincide: `f holomorphic_on A` means that f is differentiable at every point of A. On the other hand, `f analytic_on A` means that f has a power series expansion at every point of A – or, equivalently, that f is holomorphic on some open superset of A. For non-open sets, the notion `analytic_on` turns out to be much more useful.

A weaker condition than holomorphicity is *meromorphicity* on a set A: the function is differentiable at every point of A except for some isolated points at which it has *poles* (i.e. it tends to infinity). It was not straightforward to extend this definition to non-open sets, and after some false starts we arrived at the following very simple definition: f is meromorphic on A if f has a Laurent series expansion at every point of A.

```
definition meromorphic_on :: "(complex ⇒ complex) ⇒ complex set ⇒ bool"
  (infixl "(meromorphic'_on)" 50) where
  "f meromorphic_on A ⟷ (∀z∈A. ∃F. (λw. f (z + w)) has_laurent_expansion F)"
```

3 Elliptic functions and complex lattices

Two complex numbers ω_1 and ω_2 such that $\omega_2/\omega_1 \notin \mathbb{R}$ generate a lattice $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ in the complex plane. If we identify all complex numbers that differ by an element of Λ we obtain a complex torus $\mathbb{T}_\Lambda$.

▶ **Definition 1** (elliptic functions). *An* elliptic *function is a meromorphic function* $\mathbb{T}_\Lambda \to \mathbb{C}$.

Apostol defines it as a meromorphic function $\mathbb{C} \to \mathbb{C}$ that is periodic in both ω_1 and ω_2. This is also how we formalise it. The simplest non-trivial elliptic function is the following:

▶ **Definition 2** (Weierstraß $\wp$ function). $\wp(\Lambda, z) = \frac{1}{z^2} + \sum_{\omega\in\Lambda\setminus\{0\}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right)$.

From this, a collection of related numbers arises:

Eisenstein series: $G_n(\Lambda) = \sum_{\omega\in\Lambda\setminus\{0\}} \omega^{-n}$.
Weierstraß invariants: $g_2(\Lambda) = 60G_4(\Lambda)$ and $g_3(\Lambda) = 140G_6(\Lambda)$.
Modular discriminant: $\Delta(\Lambda) = g_2(\Lambda)^3 - 27g_3(\Lambda)^2$.
Klein's J invariant: $J(\Lambda) = g_2(\Lambda)^3/\Delta(\Lambda)$.

To illustrate the relevance of these numbers, note the following results:

▶ **Theorem 3** (Laurent series expansion of $\wp$ at $z = 0$). $\wp(\Lambda, z) = \frac{1}{z^2} + \sum_{n\geq 1}(n+1)G_{n+2}(\Lambda)z^n$.

▶ **Theorem 4** (Differential equation for $\wp$). $[\wp'(\Lambda, z)]^2 = 4\wp^3(\Lambda, z) - g_2(\Lambda)\wp(\Lambda, z) - g_3(\Lambda)$.

▶ **Theorem 5** (Non-vanishing of Δ). $\Delta(z) \neq 0$ *for all* z.

It is convenient to rotate and scale the lattice such that $\omega_1 = 1$ and $\omega_2 = \tau$ (where $\mathrm{Im}(\tau) > 0$) so that we can describe the lattice by a single complex parameter. We can thus also write $G_n(\tau)$, $\Delta(\tau)$ etc. and view G_n, Δ, etc. as functions $\mathcal{H} \to \mathbb{C}$, where $\mathcal{H} = \{z \mid \mathrm{Im}(z) > 0\}$ is the complex upper half plane. Importantly, all functions mentioned in this section are meromorphic on $\mathcal{H}$.

The last important results in this section are the *Fourier expansions* of G_n, Δ, and J. For example, using the Riemann ζ function and writing σ_a for the divisor function $\sigma_a(n) = \sum_{d|n} d^a$ and $q = e^{2i\pi\tau}$, we have the following:

▶ **Theorem 6** (Fourier expansion of G_n)**.** *For even n, $G_n(\tau)$ has the following Fourier expansion at $\tau = i\infty$: $G_n(\tau) = 2\Big(\zeta(n) + \frac{(2i\pi)^n}{(n-1)!}\sum_{k\geq 1}\sigma_{n-1}(k)q^k\Big)$.*

We also formalised similar Fourier expansions for Δ and J. The Fourier coefficients of these do not have such simple closed forms, but we derive useful recurrences for them. These expansions show that G_n, Δ, and J are "meromorphic at $i\infty$", which will be important later.

4 Modular forms

4.1 The modular group

The Möbius transformations of the form $z \mapsto \frac{az+b}{cz+d}$ form a group under function composition. This is the projective linear group $\mathrm{PSL}(2,\mathbb{Z})$, also known as the *modular group* Γ. This group is related to the functions G_n, Δ, J above because they satisfy simple functional equations under composition with elements from the modular group, namely if $h(z) = \frac{az+b}{cz+d}$ then $G_n(h(z)) = (cz+d)^n G_n(z)$ and $\Delta(h(z)) = (cz+d)^{12}\Delta(z)$ and $J(h(z)) = J(z)$.

In Isabelle, we represent the modular group as a type `modgrp`. This is a quotient type of the set of tuples (a, b, c, d) with a, b, c, $d \in \mathbb{Z}$ and $ad - bc = 1$ modulo a relation that identifies (a, b, c, d) and $(-a, -b, -c, -d)$. We show that this is a group, which we write multiplicatively.

Two special kinds of modular transformations are shifts $T_n(z) = z + n$ and "mirror-inversions" $S(z) = -1/z$. Notably, any modular transformation can be decomposed (non-uniquely) into a product of S and T_n. We formalise this fact as an induction rule:

```
lemma modgrp_induct_S_shift [case_names id S shift]:
  fixes P:: "modgrp ⇒ bool"
  assumes "P 1  and "⋀x. P x ⟹ P (S_modgrp * x)"
      and "⋀x n. P x ⟹ P (shift_modgrp n * x)"
    shows "P x"
```

4.2 Fundamental regions

Consider a subgroup G of the modular group. We now consider two points in the upper half-plane $\mathcal{H}$ to be equivalent whenever there exists a transformation in G that maps one to the other:

▶ **Definition 7** (equivalence under a subgroup of the modular group)**.** *Let G be a subgroup of the modular group* `modgrp`*, and τ and τ' be two points in the upper half-plane H. We consider τ and τ' to be* equivalent *under G if $\tau' = f\tau$ for some f in G.*

We can designate a canonical representative for each equivalence class e.g. by picking a sub-region of $\mathcal{H}$ that contains exactly one representative of each class. The interior of a region that satisfies this is called a *fundamental region*.

▶ **Definition 8** (Fundamental region)**.** *An open subset R of $\mathcal{H}$ is a* fundamental region *of G provided that:*

- *No two distinct points of R are equivalent under G.*
- *If $\tau \in H$ then there is a point τ' in the closure of R such that τ' is equivalent to τ.*

Next we show that a particular region is indeed a fundamental region of the full modular group. We call this the *standard fundamental region* $\mathcal{R}_\Gamma$:

▶ **Theorem 9.** *The open set $\mathcal{R}_\Gamma = \{\tau \in H \mid |\tau| > 1, |\mathrm{Re}(\tau)| < \frac{1}{2}\}$ is a fundamental region for Γ.*

4.3 Removing removable singularities

One issue that arises in formalising complex analysis is that on paper, removable singularities are essentially ignored completely. For example, if we have the functions $f(z) = z$ and $g(z) = 1/z$ then a mathematician would write $f(z) \cdot g(z) = 1$. In a theorem prover like Isabelle/HOL, this does not work: at least not if f and g are functions of type $\mathbb{C} \to \mathbb{C}$.

Our solution is to introduce a special type to capture meromorphic complex functions modulo removable singularities. Since our main interest later on will be functions on the upper half plane $\mathcal{H} = \{z \mid \mathrm{Im}(z) > 0\}$, we additionally restrict the functions to that domain.

To be precise: our type `mero_uhp` consists of those functions $f : \mathbb{C} \to \mathbb{C}$ that are meromorphic on $\mathcal{H}$ and return 0 at their poles and outside $\mathcal{H}$. This captures exactly the mathematical idea of meromorphic functions on $\mathcal{H}$.

Conversion of a "normal" complex function f to the `mero_uhp` type is done by restricting f to the appropriate domain and fixing removable singularities. The latter is done with the very useful function `remove_sings`:

```
definition remove_sings :: "(complex ⇒ complex) ⇒ complex ⇒ complex" where
  "remove_sings f z = (if ∃c. f −z→ c then Lim (at z) f else 0)"
```

This function takes a complex function (assumed to be meromorphic) and returns a version of that function with all removable singularities removed and all poles totalised to 0.

With this, we can now also define basic arithmetic on `mero_uhp` and prove that it is a field and a $\mathbb{C}$-vector space, which would not be possible for the normal function type.

This type *mero_uhp* now forms the basis for our formalisation of modular forms and modular functions.

4.4 Definition of modular forms

Next we will finally define modular forms and related concepts, namely as "sufficiently nice" functions that satisfy interesting equations under composition with modular transformations.[1]

▶ **Definition 10.** *A* weakly modular form *of integer weight k w.r.t. a subgroup G of the modular group is a meromorphic function $f : \mathcal{H} \to \mathcal{C}$ that satisfies the functional equation $f(h(z)) = (cz + d)^k f(z)$ for any $h(z) = \frac{az+b}{cz+d}$ with $h \in \Gamma$.*

By adding more conditions, we can define the following concepts.

- if f is additionally meromorphic at the cusps, we call it a *meromorphic form*
- if f is even holomorphic (including at the cusps), we call it a *modular form*
- a meromorphic form of weight 0 is called a *modular function*

Here, "meromorphic at the cusps" means that $f(h(z))(cz + d)^{-k}$ has a meromorphic Fourier expansion $\sum_{n \geq n_0} a_n e^{2i\pi z}$ at $z = i\infty$ for all $h \in \Gamma$ (not just in G). For "holomorphic at the cusps", we additionally require $n_0 \geq 0$.

In Section 3 we have already seen that G_n is a modular form of weight n for $n \geq 3$, Δ is a modular form of weight 12, and J is a modular function.

Apostol does not use the terms "weakly modular form" and "meromorphic form" at all, but we find that they make the formalisation more modular: they allow e.g. the valence formula (below) to be shown directly for meromorphic forms rather than deriving them separately for modular forms and modular functions. This is a typical case where the educational approach of Apostol's textbook clashes with the needs of formalisation.

[1] For simplicity, some of our definitions in Isabelle currently only work when G is the full modular group, but this will be generalised soon.

4.5 The valence formula

The central result in our formalisation so far is the valence formula for meromorphic forms. It relates the number of zeros of a modular form to the number of its poles:

▶ **Theorem 11.** *Let f be a non-zero meromorphic form of weight k on the full modular group Γ. Then the sum of the multiplicities of the zeros of f inside the closure of $\mathcal{R}_\Gamma$ minus the sum of the multiplicities of its poles in the same set is exactly $k/12$.*

Several caveats apply here about how to count zeros and poles directly at the border of the region: any point on the border is weighted with $\frac{1}{2}$, except for the points $\pm\frac{1}{2} + \frac{\sqrt{3}}{2}i$, which are weighted with $\frac{1}{6}$. It should also be noted that $i\infty$ may also be a zero or pole and must be counted accordingly (with weight 1).

The proof of the valence formula was the most difficult to formalise so far. The basic idea is simple: we apply the argument principle and integrate along a contour that is essentially a finite version of the border of $\mathcal{R}_\Gamma$. Due to the symmetries of $\mathcal{R}_\Gamma$ and f, most of the integral cancels, only $k/12$ remains plus the contribution of the potential zero or pole at $i\infty$.

The problem is that there may be zeros or poles directly on the border itself and we need to add little "wiggles" to avoid these and account for the error made by this. This is easy to justify on paper, but not in a theorem prover. We eventually solved this problem by using the "Wiggle Framework", which the first author developed specifically for this proof (but with similar future applications in mind). It allows deforming integration contours and relating them to the original contour. We are currently planning to eventually replace this framework with a much simpler approach based on a generalised residue theorem that allows singularities on the integration path. [6]

For modular functions, the valence formula is particularly striking: it means that the number of zeros of a modular function $f(z)$ is exactly the same as the number of its poles. Moreover, since the number of zeroes in $f(z) - c$ is the same as that of $f(z)$, we can even say that f takes on all complex values equally often. In particular, J is a bijection between $\mathcal{R}'_\Gamma$ and $\mathbb{C}$ (where $\mathcal{R}'_\Gamma$ denotes the union of $\mathcal{R}_\Gamma$ and the left half of its closure).

Apostol uses this last fact to give a relatively simple proof of Picard's little theorem (a non-constant entire function takes every value in $\mathbb{C}$ with at most one exception). We formalised this as well, but it turned out to be not quite so simple: in particular, we had to first prove the stronger fact that J is a *covering* between $\mathcal{H}$ and $\mathbb{C}$, which was a reasonably simple, but somewhat tedious and definitely non-trivial proof. It is surprising that Apostol does not mention this seemingly indispensable bit of work in his proof.

Another straightforward application of the valence formula is to determine the dimension of the vector space of modular forms of weight k, the formalisation of which is ongoing work.

5 Dedekind's η function

▶ **Definition 12** (The Euler function ϕ and Dedekind's η function)**.** *Define $\phi(q) = \prod_{k\geq 1}\left(1-q^k\right)$ and $\eta(z) = e^{i\pi z/12}\phi\left(e^{2i\pi z}\right)$. They are holomorphic for $|q| < 1$ and $z \in \mathcal{H}$, respectively.*

Dedekind's η function is not a modular form in the sense that Apostol defines, but it does display interesting behaviour under the two generators $z \mapsto z + 1$ and $z \mapsto -1/z$ of the modular group:[2]

[2] Here, $\sqrt{}$ denotes the standard branch of the complex square root where $\mathrm{Re}(\sqrt{z}) \geq 0$ for all $z \in \mathbb{C}$ and $\mathrm{Im}(\sqrt{x}) > 0$ for any real $x < 0$.

▶ **Theorem 13.** *$\eta(z+1) = e^{i\pi/12}\eta(z)$ and $\eta(-1/z) = \sqrt{-iz}\,\eta(z)$.*

Using these two relations and our induction rule for the modular group, one can show the following more general equation:

▶ **Theorem 14.** *If $h(z) = \frac{az+b}{cz+d}$ is an element of the modular group, then $\eta(h(z)) = \varepsilon_h\sqrt{cz+d}\,\eta(z)$ where ε_h is a 24^{th} root of unity depending on h but not on z.*

This ε_h has an explicit (albeit complicated) formula in terms of Dedekind sums which we shall not show here. It is noteworthy that our definition of ε_h and our version of the theorem differ somewhat from Apostol's, since ours work for any value of c while he requires $c > 0$.

Interestingly, Apostol proves Theorem 14 directly using Iseki's formula: a technical lemma whose proof is four pages of dense calculations and which is never used again. We chose *not* to formalise Iseki's formula and to instead follow a simpler approach outlined in the appendix of the second edition of the book: we first follow Apostol's proofs for Theorem 13 and then obtain Theorem 14 from it.

An interesting consequence of Theorem 14 is that η^{24} is a modular form of weight 12. Combining this with the valence formula, one obtains (relatively easily) a remarkable connection: $\Delta(z) = (2\pi)^{12}\eta(z)^{24}$.

6 Discussion and related work

The tradition of formalising textbooks dates back to Jutting's formalisation of Landau's *Foundations of Analysis* using AUTOMATH [8] in 1977. The challenge is about the volume of material but also the obligation to cover everything rather than to pick and choose. Although we did not have time to formalise the entire text, we did cover half of the eight chapters.

We built upon a huge library of prior material, including Laurent series, winding numbers, Dirichlet series, polynomial factorisation and Bernoulli numbers, all of which had to interoperate. We worked under the handicap that none of us is a number theorist. Perhaps for this reason, many of the Isabelle/HOL proofs are considerably longer than Apostol's. We invested some effort in making the formal proofs clear, through Isabelle's structured proof language, hoping to retain some of the pedagogical value of the original text. The four chapters (1, 2, 3, 7) respectively consist of 12K, 10K, 4K, and 3K lines of proof scripts including comments. Together with other supporting material, the project has already exceeded 53,000 lines.

Much number theory has been formalised in other proof assistants, chiefly Lean. To our knowledge, ours was the first treatment of elliptic functions and modular forms in a theorem prover, although we are aware of more recent unpublished work by Birkbeck [3] in Lean covering mostly the definition of modular forms and Eisenstein series. This work is now also part of Mathlib 4.

7 Conclusions

The formalisation of a textbook remains challenging. Our impression was that Apostol's proofs were clear overall, and wherever they were, the formalisation process was straightforward, regardless of the mathematical tools required. There were however some gaps, mistakes, and informal arguments that took time to overcome, but none that were serious. Graduate-level analytic number theory can be formalised in Isabelle/HOL without undue effort.

References

1. Tom M. Apostol. *Introduction to Analytic Number Theory*. Springer, 1976.
2. Tom M. Apostol. *Modular Functions and Dirichlet Series in Number Theory*. Springer, 1990.
3. Chris Birkbeck. ModularForms. GitHub repository. URL: `https://github.com/CBirkbeck/ModularForms`.
4. Manuel Eberl. Nine chapters of analytic number theory in Isabelle/HOL. In *10th International Conference on Interactive Theorem Proving (ITP 2019)*, volume 141 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:19, 2019.
5. Manuel Eberl, Anthony Bordg, Lawrence C. Paulson, and Wenda Li. Formalising half of a graduate textbook on number theory (formal proof development), June 2024. `doi:10.5281/zenodo.12586104`.
6. Norbert Hungerbühler and Micha Wasem. Non-integer valued winding numbers and a generalized residue theorem. *Journal of Mathematics*, 2019:1–9, March 2019. `doi:10.1155/2019/6130464`.
7. Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. Springer, 2002.
8. L. S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH System*. PhD thesis, Eindhoven University of Technology, 1977.