

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Τυπικές Αποδείξεις Αριθμητικής Σταθερότητας στη Lean 4

Η Βιβλιοθήκη NumStability και η Πειραματική Αξιολόγηση
της Χρήσης της από Πράκτορες Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Αλέξανδρος Ο. Γεωργαντζάς

Επιβλέπουσα: Ζωή Παρασκευοπούλου, Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Οκτώβριος 2026



Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Τυπικές Αποδείξεις Αριθμητικής Σταθερότητας στη Lean 4

Η Βιβλιοθήκη NumStability και η Πειραματική Αξιολόγηση
της Χρήσης της από Πράκτορες Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Αλέξανδρος Ο. Γεωργαντζάς

Επιβλέπουσα: Ζωή Παρασκευοπούλου, Επίκουρη Καθηγήτρια Ε.Μ.Π.
Συνεπιβλέπων: Κίμων Φουντουλάκης, Αναπληρωτής Καθηγητής, University of Waterloo

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Οκτωβρίου 2026.

Ζωή Παρασκευοπούλου	Νικόλαος Παπασπύρου	Κίμων Φουντουλάκης
Επίκουρη Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής, University of Waterloo

Αθήνα, Οκτώβριος 2026

Γεώργιος Αλέξανδρος Ο. Γεωργαντζάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Γεώργιος Αλέξανδρος Ο. Γεωργαντζάς, 2026.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Κύριος στόχος της εργασίας είναι η κατασκευή της NumStability, μιας εξειδικευμένης βιβλιοθήκης στη Lean 4 για την τυπική απόδειξη θεωρημάτων αριθμητικής σταθερότητας αλγορίθμων κινητής υποδιαστολής. Μαθηματικό σημείο αναφοράς είναι το βιβλίο του N. J. Higham *Accuracy and Stability of Numerical Algorithms*· η NumStability τυποποιεί επιλεγμένο υλικό από καθένα από τα 28 κεφάλαιά του και το οργανώνει σε επαναχρησιμοποιήσιμη υποδομή.

Η βιβλιοθήκη κατασκευάστηκε κυρίως με το Codex της OpenAI και σποραδικά με το Claude της Anthropic. Η εργασία καταγράφει τις αδυναμίες μιας αρχικής αρχάριας τυποποίησης του βιβλίου, τις ανθρώπινες αποφάσεις που τις διόρθωσαν, την κατάτμηση ανά κεφάλαιο, τα skills και την αναδιοργάνωση που οδήγησε στην έκδοση `higham_v01`.

Το HighamBench εξετάζει αν η πρόσβαση στη βιβλιοθήκη μειώνει το κόστος νέων αποδείξεων για έναν πράκτορα TN, συγκρίνοντας τη συνθήκη N (Lean και Mathlib) με τη συνθήκη L (επιπλέον η NumStability). Στο HighamBench-Proof οι στόχοι δίνονται ήδη τυποποιημένοι ως σταθερές δηλώσεις Lean: ο πράκτορας αποδεικνύει 30 τέτοιους στόχους από 25 άρθρα. Στο HighamBench-E2E οι στόχοι προέρχονται από φυσική μαθηματική γλώσσα: ο πράκτορας διατυπώνει σε Lean δέκα αποτελέσματα από πέντε πηγές, με έλεγχο πιστότητας από ανεξάρτητους κριτές βασισμένους σε γλωσσικά μοντέλα, και έπειτα τα αποδεικνύει. Το κέρδος αυξάνεται μαζί με την πραγματική χρήση δηλώσεων της βιβλιοθήκης στις αποδείξεις: ο συντελεστής Spearman μεταξύ χρήσης και κέρδους χρόνου απόδειξης είναι +0,69 και +0,56 αντίστοιχα, και μεταξύ χρήσης και μείωσης των γραμμών κώδικα +0,61 για το πλήρες υποβληθέν αρχείο και +0,64 για την απόδειξη, αντίστοιχα. Η επικάλυψη της βιβλιοθήκης με τα άρθρα αριθμητικής σταθερότητας ήταν όμως περιορισμένη: μόνο 9 από τα 25 άρθρα πρόσφεραν αποτέλεσμα για το οποίο αναμενόταν βοήθεια, και σε 17 από τους 30 σταθερούς στόχους η βιβλιοθήκη δεν χρησιμοποιήθηκε καθόλου. Η γνώση της βιβλιοθήκης είναι επομένως στενή, αλλά όπου υπάρχει επικάλυψη τα αποτελέσματα είναι θετικά.

Η βιβλιοθήκη είναι το πρώτο στάδιο ενός απώτερου στόχου: ενός γενικού συστήματος τυποποίησης βιβλίων με πράκτορες TN, στο οποίο ένα skill τυποποίησης παράγει τη βιβλιοθήκη και ενημερώνεται από τις καταγραφές των εκτελέσεων και από ανθρώπινες διορθώσεις, ενώ ένα benchmark μετρά τη χρησιμότητά της.

Λέξεις-κλειδιά: αριθμητική κινητής υποδιαστολής, αριθμητική σταθερότητα, ανάλυση backward σφάλματος, Lean 4, Mathlib, τυπική επαλήθευση, πειραματική αξιολόγηση, HighamBench, Codex, αυτόματη τυποποίηση.

Abstract

The principal aim of this thesis is the construction of NumStability, a domain-specific Lean 4 library for formally proving numerical-stability theorems about floating-point algorithms. Its mathematical reference is N. J. Higham's book *Accuracy and Stability of Numerical Algorithms*; NumStability formalises selected material from each of its 28 chapters and organises it as reusable infrastructure.

The library was built mainly with OpenAI Codex and occasionally with Anthropic's Claude. The thesis documents the shortcomings of an initial, inexperienced formalisation of the book, the human decisions that corrected them, the chapter splitting, the agent skills and the reorganisation that produced the `higham_v01` release.

HighamBench examines whether access to the library reduces the cost of new proofs for an AI agent, comparing condition N (Lean and Mathlib) with condition L (plus NumStability). In HighamBench-Proof the targets are given already formalised, as fixed Lean statements: the agent proves 30 such targets from 25 papers. In HighamBench-E2E the targets come from natural mathematical language: the agent formalises ten results from five sources in Lean, with faithfulness checked by independent language-model judges, and then proves them. The gain grows with the actual use of library declarations in proofs: the Spearman coefficient between use and proof-time gain is +0.69 and +0.56 respectively, and between use and code-size reduction +0.61 for the full submitted file and +0.64 for proof code, respectively. The overlap between the library and the numerical-stability papers was limited, however: only 9 of the 25 papers offered a result for which help was expected, and in 17 of the 30 fixed targets the library was not used at all. The library's knowledge is therefore narrow, but where it overlaps the results are positive.

The library is the first stage of an ultimate goal: a general system for formalising books with AI agents, in which a formalisation skill builds the library and is updated from execution logs and human corrections, while a benchmark measures its usefulness.

Keywords: floating-point arithmetic, numerical stability, backward error analysis, Lean 4, Mathlib, formal verification, HighamBench, Codex, autoformalization.

Ευχαριστίες

Πρώτα από όλα, θέλω να ευχαριστήσω τα σημαντικότερα πρόσωπα της ζωής μου, την οικογένειά μου. Χωρίς τη στήριξη της μητέρας μου, του πατέρα μου και της αδερφής μου δεν θα κατάφερνα ποτέ να εισαχθώ σε μια τόσο μεγάλη σχολή, πόσο μάλλον να αποφοιτήσω.

Επίσης, θέλω να ευχαριστήσω την κ. Ζωή Παρασκευοπούλου, που δέχτηκε αμέσως να επιβλέψει τη διπλωματική μου, ήταν πάντοτε συνεργάσιμη, θετική και με ενθάρρυνε να προχωρήσω το project.

Τέλος, θέλω να ευχαριστήσω τον κ. Κίμωνα Φουντουλάκη, καθώς η αρχική ιδέα του project ήταν δική του. Η συνεργασία μας ήταν άψογη, και τον ευχαριστώ για τον χρόνο και την καθοδήγησή του. Είμαι επίσης ευγνώμων στον κ. Ιωάννη Εμίρη από το ΕΚΠΑ, που με έφερε σε επαφή με τον κ. Φουντουλάκη.

Περιεχόμενα

Περίληψη **5**

Abstract **7**

Ευχαριστίες **9**

1 Εισαγωγή **18**
1.1 Το ερευνητικό πρόβλημα 18
1.2 Απευθυνόμενο κοινό και στόχοι 19
1.3 Κίνητρο και ερευνητικό κενό 20
1.4 Κύρια παραδοτέα της εργασίας 20
1.5 Όρια των ισχυρισμών 22
1.6 Δομή της εργασίας 22

2 Αριθμητική κινητής υποδιαστολής και αριθμητική σταθερότητα **23**
2.1 Από τους πραγματικούς αριθμούς στον υπολογισμό 23
2.2 Το καθιερωμένο μοντέλο σχετικού σφάλματος 24
2.3 Συσσώρευση τοπικών σφαλμάτων 25
2.4 Κατάσταση του προβλήματος 26
2.5 Forward, backward και mixed σφάλματα 26
2.6 Παράδειγμα: αναδρομική άθροιση 28
2.7 Εσωτερικό γινόμενο και πράξεις πινάκων 29
2.8 Αριθμητική σταθερότητα ως ιδιότητα αλγορίθμου 30
2.9 Τι ελέγχει το HighamBench 31
2.10 Σύνοψη 31

3 Η Lean, το Codex και τα όρια εμπιστοσύνης **32**
3.1 Από τη μαθηματική αιτιολόγηση στον ελέγξιμο όρο απόδειξης 32
3.2 Η Mathlib ως μαθηματική υποδομή 33
3.3 Η επιλογή της Lean 34
3.4 Το Codex στην υποβοηθούμενη ανάπτυξη 34
3.5 Κατανομή εμπιστοσύνης μεταξύ ανθρώπου, TN και πυρήνα Lean 35
3.6 Συμπέρασμα 36

4 Σχετική βιβλιογραφία **37**
4.1 Πεδίο και επίπεδα επαλήθευσης . . . 37
4.2 Θεμελιώσεις αριθμητικής κινητής υποδιαστολής . . . 38
4.3 Αυτόματη ανάλυση συγκεκριμένων υπολογισμών . . . 39
4.4 Ανάλυση προγραμμάτων και πιστοποιημένα αποτελέσματα . . . 40
4.5 Επαναχρησιμοποιήσιμη θεωρία και ολόκληροι αλγόριθμοι . . . 42
4.6 Προηγούμενες αξιολογήσεις και μεθοδολογία . . . 43

5 Η βιβλιοθήκη NumStability **46**
5.1 Σκοπός και εύρος . . . 46
5.2 Το μοντέλο σχετικού σφάλματος του Higham . . . 46
5.3 Ακριβής άλγεβρα, Mathlib και πίνακες . . . 47
5.4 Συνθετική δομή αποδείξεων . . . 47
5.5 Σύνδεση των θεωρημάτων με συγκεκριμένους υπολογισμούς . . . 52

6 Κατασκευή της βιβλιοθήκης και ανθρώπινη καθοδήγηση **54**
6.1 Το πρώτο πέρασμα χωρίς skill και αυστηρές οδηγίες . . . 54
6.2 Από τα προβλήματα στους κανόνες του skill . . . 55
6.3 Εφαρμογή του skill στην παράλληλη κατασκευή . . . 57
6.4 Έλεγχοι, ευρήματα και όρια . . . 59

7 HighamBench: μεθοδολογία των δύο αξιολογήσεων **60**
7.1 Ερευνητικό ερώτημα και συνθήκες . . . 60
7.2 HighamBench-Proof: απόδειξη σταθερών στόχων . . . 61
7.3 HighamBench-E2E: τυποποίηση και απόδειξη . . . 63
7.4 Μέτρηση της χρήσης της NumStability . . . 66

8 HighamBench: αποτελέσματα **68**
8.1 HighamBench-Proof: απόδειξη σταθερών στόχων . . . 68
8.2 HighamBench-E2E: τυποποίηση και απόδειξη . . . 70

9 Συμπεράσματα και μελλοντική έρευνα **73**
9.1 Τι έδειξαν οι δύο αξιολογήσεις . . . 73
9.2 Επόμενα βήματα . . . 74
9.3 Απώτερος σκοπός: ένα σύστημα τυποποίησης βιβλίων . . . 74
9.4 Επέκταση σε επόμενα βιβλία . . . 75

Α΄ Δήλωση χρήσης εργαλείων τεχνητής νοημοσύνης **77**

Β΄ HighamBench-Proof: συμπληρωματικά στοιχεία **78**
Β΄.1 Κατάλογος εργασιών και πηγών . . . 78
Β΄.2 Αναλυτικά αποτελέσματα ανά εργασία . . . 82
Β΄.3 Μέτρα χρήσης της βιβλιοθήκης . . . 83
Β΄.4 Δευτερεύοντα διαγνωστικά . . . 85

Β΄.5 Ευρετήριο τεκμηρίων 86
Β΄.6 Προτροπές 87

Γ΄ HighamBench-E2E: συμπληρωματικά στοιχεία **89**
Γ΄.1 Κατάλογος εργασιών και πηγών 89
Γ΄.2 Αναλυτικά αποτελέσματα ανά εργασία 90
Γ΄.3 Μέτρα χρήσης της βιβλιοθήκης 91
Γ΄.4 Δευτερεύοντα διαγνωστικά 93
Γ΄.5 Ευρετήριο τεκμηρίων 94
Γ΄.6 Προτροπές 95

Βιβλιογραφία **98**

Κατάλογος σχημάτων

2.1 Forward και backward σφάλμα 27

3.1 Από τη μαθηματική πηγή στην ελεγμένη απόδειξη: διακριτοί ρόλοι μετάφρασης, ελέγχου πιστότητας και τυπικής επαλήθευσης. Στην παρούσα εργασία, την πιστότητα κρίνουν χωριστές, απομονωμένες εκτελέσεις του Codex. Ο άνθρωπος επιλέγει τις πηγές και σχεδιάζει τα κριτήρια, χωρίς να επαναβαθμολογεί τις επιμέρους κρίσεις. 36

4.1 Συμπληρωματικά επίπεδα μιας τυπικής ανάλυσης αριθμητικής κινητής υποδιαστολής. Τα ονόματα δηλώνουν χαρακτηριστικά παραδείγματα και όχι αποκλειστική κατάταξη· οι αριθμοί των στρωμάτων χρησιμοποιούνται στο κείμενο του κεφαλαίου. 38

5.1 Κύριες εξαρτήσεις μεταξύ των αρχιτεκτονικών στρωμάτων της NumStability στην έκδοση `higham_v01`. Κάθε βέλος δείχνει από το στρώμα που χρησιμοποιεί προς το στρώμα που χρησιμοποιείται. Η ετικέτα $m \to k$ σημαίνει ότι m δηλώσεις του πρώτου στρώματος χρησιμοποιούν k διαφορετικές δηλώσεις του δεύτερου. Όλα τα πλήθη αφορούν το σύνολο αναφοράς των 47 890 δηλώσεων του Πίνακα 5.1. Η πλήρης απομόνωση του κάτω πλαισίου λαμβάνει υπόψη και ακμές προς τις υπόλοιπες δηλώσεις του έργου. 50

5.2 Συμπτυγμένη απεικόνιση μιας επαληθευμένης ραχοκοκαλιάς της απόδειξης backward error για επίλυση μέσω Householder QR. Τα επτά πλαίσια ομαδοποιούν τις 15 δηλώσεις ανά υπομονάδα· μαζί με τις εσωτερικές εξαρτήσεις των πλαισίων η πλήρης αλυσίδα περιέχει 14 άμεσες ακμές. Κάθε βέλος έχει φορά καταναλωτή προς εξάρτηση. 51

6.1 Η αλλαγή στην τυποποίηση της Cholesky. Οι ανισότητες είναι componentwise και η γραφή είναι συνοπτική μαθηματική απόδοση των δηλώσεων Lean. Δεξιά απαιτούνται συμμετρία του A, $(n+1)u < 1$, μη αρνητικοί υπολογισμένοι οδηγοί και μη μηδενική υπολογισμένη διαγώνιος. 55

6.2 Γενικό χρονολόγιο της κατασκευής της NumStability, από τον Μάρτιο έως τον Σεπτέμβριο του 2026. Πάνω από τον άξονα δίνονται οι φάσεις και κάτω οι κύριες ανθρώπινες αποφάσεις και τα ορόσημα κάθε μήνα. 57

6.3 Ο κύκλος του skill σε κάθε τμήμα κάθε ροής. Οι ανθρώπινες αποφάσεις του Πίνακα 6.1 καθορίζουν τι επιλέγεται, πώς κατασκευάζεται και με ποιους ελέγχους γίνεται δεκτό. 59

7.1 Κατηγορίες αναμενόμενης βοήθειας T1, T2 και T3 62
7.2 Η ροή μιας εργασίας στο HighamBench-E2E. Τα έντονα πλαίσια χρονομετρούνται ως εργασία του πράκτορα· τα γκρίζα εκτελούνται εκτός του χρόνου του από ανεξάρτητους ελεγκτές. 64
7.3 Οι δύο διαδρομές ελέγχου πιστότητας 64

8.1 HighamBench-Proof: μέση άμεση χρήση της βιβλιοθήκης στην απόδειξη (DP) έναντι των κερδών της L. Τα χρώματα και τα σύμβολα δηλώνουν την αναμενόμενη κατηγορία και όχι μετρημένη χρήση. 69
8.2 Κόστος τυποποίησης και απόδειξης στο HighamBench-E2E 71
8.3 HighamBench-E2E: διακριτές δηλώσεις της βιβλιοθήκης στην αποδεκτή διατύπωση (DS, πάνω σειρά), στον όρο απόδειξης και στα τοπικά βοηθητικά λήμματα (DP, μεσαία σειρά) και στην ένωσή τους (DU, κάτω σειρά), έναντι των κερδών της τυποποίησης, της απόδειξης και του συνόλου αντίστοιχα. Σε αυτό το σύνολο DU = DP σε κάθε εργασία. 72

9.1 Το σύστημα τυποποίησης βιβλίων 75

Κατάλογος πινάκων

4.1 Αναπαραγωγή του σώματος του Πίνακα 2 των Titolo κ.ά. (2025), ο οποίος συγκρίνει δυνατότητες εργαλείων ανάλυσης σφάλματος στρογγυλοποίησης χειρότερης περίπτωσης. Η δημοσίευση διατίθεται με άδεια CC BY 4.0. 42

5.1 Ποσοτικές ενδείξεις εσωτερικής συνοχής της τελικής έκδοσης. Η χρήση από διαφορετική υπομονάδα εξαρτάται και από τα όρια των αρχείων, ενώ οι μετρικές έχουν διαφορετική σημασία και δεν συνδυάζονται σε έναν ενιαίο δείκτη συνοχής. 49

6.1 Ευρήματα της αρχικής τυποποίησης και ανθρώπινες αποφάσεις που μετατράπηκαν σε απαιτήσεις του skill για τις επόμενες εκτελέσεις. 56

7.1 Οι δύο αξιολογήσεις που εκτελέστηκαν. Ο ορισμός της άμεσης χρήσης της βιβλιοθήκης στην απόδειξη, ο κανόνας μέτρησης των γραμμών κώδικα και ο τύπος των νέων tokens είναι κοινοί· όλες οι υπόλοιπες διαστάσεις διαφέρουν. 61

7.2 Οι 16 σημασιολογικοί έλεγχοι του άμεσου κριτή και του κριτή αντίστροφης μετάφρασης στο HighamBench-E2E, με τη σειρά της προτροπής τους. 65

8.1 HighamBench-Proof: μέσοι όροι των 30 εργασιών και κέρδος από τον λόγο των συνολικών τιμών. Κάθε τιμή εργασίας είναι ο μέσος όρος τριών αποδεκτών επαναλήψεων. 68

8.2 HighamBench-Proof: αναμενόμενη βοήθεια έναντι παρατηρούμενης άμεσης χρήσης της βιβλιοθήκης στην απόδειξη. 69

8.3 HighamBench-E2E: αθροίσματα των δέκα εργασιών ανά στάδιο και συνολικά. Το κόστος της εξοικείωσης προσμετράται μία φορά, μόνο στις δύο αντίστοιχες συνολικές γραμμές. 70

8.4 HighamBench-E2E: συντελεστές Spearman ανά στάδιο, για τις δέκα εργασίες. . 71

9.1 Βιβλία για την επέκταση του συστήματος τυποποίησης 76

B′.1 HighamBench-Proof: περιγραφή των εργασιών και ακριβής θέση στις πηγές. . 78

B′.2 HighamBench-Proof: χρόνος (s) και νέα tokens (χιλιάδες) ανά εργασία· μέσοι όροι τριών επαναλήψεων. 82

B′.3 HighamBench-Proof: αναμενόμενη κατηγορία, μέση άμεση χρήση της βιβλιοθήκης στην απόδειξη (DP) και γραμμές υποβληθέντος κώδικα. 83

Β΄.4 HighamBench-Proof: συντελεστές Spearman των δευτερευόντων μέτρων χρήσης με τα κέρδη της L ($n = 30$). Οι ισχυρότερες συσχετίσεις ορισμένων ευρύτερων μέτρων δεν τα καθιστούν ακριβέστερα μέτρα άμεσης βοήθειας, επειδή μπορεί να αντανακλούν εν μέρει τις διαφορές των κατηγοριών. 84
Β΄.5 HighamBench-Proof: ομαδοποίηση των εργασιών κατά το πλήθος των τελικών αποδείξεων της L με εξάρτηση από τη βιβλιοθήκη. Οι ομαδικοί μέσοι όροι δίνουν ίσο βάρος σε κάθε εργασία. 84
Β΄.6 HighamBench-Proof: οι τρεις επαναλήψεις της `P01-T2`. Η χρήση της βιβλιοθήκης δεν ανατέθηκε τυχαία· οι δύο επαναλήψεις με χρήση είχαν κατά μέσο όρο +76,8% κέρδος χρόνου και +64,2% κέρδος γραμμών, ενώ η επανάληψη χωρίς χρήση −12,4% και −24,8%. Τρεις παρατηρήσεις δεν τεκμηριώνουν σταθερή επίδραση. 84
Β΄.7 HighamBench-Proof: γραμμές κώδικα της περιοχής της απόδειξης του στόχου, χωρίς κενές γραμμές και σχόλια. 85
Β΄.8 HighamBench-Proof: ιστορικές φυσικές γραμμές του πλήρους υποβληθέντος αρχείου. 86
Β΄.9 HighamBench-Proof: τεκμήρια και διαδρομές τους. 87
Β΄.10 HighamBench-Proof: αποτυπώματα των αρχειοθετημένων προτύπων προτροπών. 88

Γ΄.1 Οι δέκα εργασίες του HighamBench-E2E, με τη σειρά των πηγών. 89
Γ΄.2 HighamBench-E2E: ενεργός χρόνος (s) και νέα tokens (χιλιάδες) ανά εργασία. . 90
Γ΄.3 Νέα tokens ανά εργασία και στάδιο 90
Γ΄.4 HighamBench-E2E: γραμμές κώδικα της τελικής δήλωσης και της απόδειξης, χρήση της βιβλιοθήκης από την L και υποβολές (δήλωση/απόδειξη). Όλες οι πρώτες υποβολές έγιναν δεκτές. 91
Γ΄.5 HighamBench-E2E: διακριτές δηλώσεις της βιβλιοθήκης στη διατύπωση (DS), στον όρο απόδειξης (DP) και στην ένωσή τους (DU). Και τα δέκα σύνολα DS περιέχονται στα αντίστοιχα DP, οπότε DU = DP. 91
Γ΄.6 HighamBench-E2E: συντελεστές Spearman των DS και DP με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens ($n = 10$). 92
Γ΄.7 HighamBench-E2E: αναφορές σε δηλώσεις της NumStability στο τελικό αρχείο της L, με επίλυση των χώρων ονομάτων (πάνω), και οι συντελεστές Spearman τους με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens (κάτω). Για παράδειγμα, στην `P14-T2` 13 διακριτά ονόματα αναφέρονται 55 φορές. 92
Γ΄.8 HighamBench-E2E: οικογένειες δηλώσεων στη διατύπωση και στον όρο απόδειξης της L. Η στήλη *n* δίνει το πλήθος ΟΔ στη διατύπωση και DP στην απόδειξη (Πίνακας Γ΄.4). 93
Γ΄.9 HighamBench-E2E: δευτερεύοντα μέτρα πηγαίου κώδικα και εξαρτήσεων της διατύπωσης (πάνω) και οι συσχετίσεις τους με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens (κάτω). Κανένα δεν μετρά άμεση επαναχρησιμοποίηση στην απόδειξη. 94
Γ΄.10 HighamBench-E2E: τεκμήρια και κρυπτογραφικά αποτυπώματα. 94
Γ΄.11 HighamBench-E2E: αποτυπώματα των σταθερών προτροπών. 95

Κεφάλαιο 1

Εισαγωγή

1.1 Το ερευνητικό πρόβλημα

Η αριθμητική κινητής υποδιαστολής αποτελεί τον πρακτικό μηχανισμό με τον οποίο οι υπολογιστές προσεγγίζουν τους πραγματικούς αριθμούς. Η προσέγγιση αυτή είναι αναγκαία, αλλά ελαττωματική: κάθε αποθήκευση και κάθε βασική πράξη μπορεί να εισάγει σφάλμα στρογγυλοποίησης. Ένας αλγόριθμος που είναι ορθός στην ακριβή αριθμητική ενδέχεται, επομένως, να επιστρέψει αποτέλεσμα το οποίο απέχει πολύ από το σωστό όταν εκτελεστεί σε μηχάνημα πεπερασμένης ακρίβειας. Το πρόβλημα δεν αντιμετωπίζεται απλώς με αύξηση της διαθέσιμης ακρίβειας σε επίπεδο υλικού. Απαιτείται ανάλυση που να διαχωρίζει την εγγενή ευαισθησία του μαθηματικού προβλήματος από τον τρόπο με τον οποίο ο συγκεκριμένος αλγόριθμος διαδίδει τα τοπικά σφάλματα (Goldberg, 1991; N. J. Higham, 2002; Wilkinson, 1963).

Η ανάλυση αριθμητικής σταθερότητας παρέχει τη γλώσσα αυτού του διαχωρισμού. Ένα υπολογισμένο αποτέλεσμα μπορεί να συγκριθεί άμεσα με το ακριβές αποτέλεσμα, δίνοντας ένα *forward error*, ή να αναζητηθεί μια μικρή διαταραχή των δεδομένων εισόδου για την οποία το υπολογισμένο αποτέλεσμα αποτελεί ακριβή λύση, δίνοντας ένα *backward error*. Η δεύτερη οπτική, κεντρική στην παράδοση των Wilkinson και Higham, επιτρέπει συχνά μια καθαρή σύνδεση ανάμεσα στη σταθερότητα του αλγορίθμου και στην κατάσταση (conditioning) του προβλήματος. Ωστόσο, οι αποδείξεις αυτών των φραγμάτων απαιτούν ιδιαίτερη προσοχή: συνθέτουν αλληλουχίες τοπικών σφαλμάτων και εξαρτώνται από πλευρικές συνθήκες, όπως $nu < 1$, η αντιστρεψιμότητα όσων πινάκων πρόκειται να αντιστραφούν και η συμβατότητα της επιλεγμένης νόρμας με τις αντίστοιχες πράξεις. Στην άτυπη παρουσίαση τέτοιες προϋποθέσεις μπορεί να θεωρούνται αυτονόητες· στη Lean 4 πρέπει να διατυπωθούν ρητά.

Οι συνήθεις μαθηματικές αποδείξεις ελέγχονται αρχικά από τον συγγραφέα και τον επιβλέποντα και, όταν δημοσιεύονται, από ομότιμους κριτές (*peer reviewers*) και αναγνώστες. Η τυπική επαλήθευση, με τη βοήθεια ενός βοηθού αποδείξεων όπως η Lean 4, προσθέτει μια ισχυρή μηχανική εγγύηση: η μαθηματική δήλωση διατυπώνεται σε αυστηρή τυπική γλώσσα και η απόδειξη γίνεται δεκτή μόνο αν ο έμπιστος λογικός πυρήνας ελέγξει τον αντίστοιχο όρο απόδειξης (Avigad, de Moura κ.ά., 2026; De Moura & Ullrich, 2021). Επομένως, η ορθότητα της αποθηκευμένης απόδειξης δεν εξαρτάται από την αξιοπιστία του Codex ή των τακτικών που τη δημιούργησαν.

Η αξία αυτής της μηχανικής εγγύησης αυξάνεται όταν δεν εφαρμόζεται σε μεμονωμένα θεωρήματα, αλλά σε ένα συνεκτικό σώμα αποτελεσμάτων. Η τυποποίηση επιλεγμένων αλυσίδων του Higham υποχρεώνει τους ορισμούς, τις υποθέσεις και τις εξαρτήσεις τους να γίνουν ρητές, ώστε ασάφειες ή αβλεψίες της άτυπης παρουσίασης να μπορούν να εντοπιστούν κατά τη σύνθεση των αποδείξεων. Δεν παράγεται, επομένως, μόνο ένα σύνολο μηχανικά ελεγμένων δηλώσεων, αλλά μια υποδομή που μπορεί να δοκιμάζεται, να διορθώνεται και να επεκτείνεται σε πολυπλοκότερες αναλύσεις. Η `higham_v01` αποτελεί την τελική έκδοση αυτής της βιβλιοθήκης για την παρούσα εργασία.

Η τυπική ορθότητα συμπληρώνεται από την πιστότητα προς τη μαθηματική πηγή: εξετάζεται αν η δήλωση Lean αποδίδει το επιδιωκόμενο αποτέλεσμα, επιτρέποντας λογικά ισοδύναμες αναδιατυπώσεις και ορθές γενικεύσεις (Meek κ.ά., 2026; Poiroux κ.ά., 2025). Η διάκριση και η κατανομή του ελέγχου μεταξύ ανθρώπου, πράκτορα και πυρήνα αναπτύσσονται στην Ενότητα 3.5.

Για να εξεταστεί η χρησιμότητα της βιβλιοθήκης σε νέες αποδείξεις, η εργασία εισάγει το HighamBench, ένα ενιαίο πλαίσιο πειραματικής αξιολόγησης. Συγκρίνει την εργασία ενός πράκτορα με και χωρίς πρόσβαση στη NumStability, τόσο όταν η δήλωση Lean είναι δοσμένη όσο και όταν ο ίδιος πρέπει πρώτα να τη διατυπώσει από τη μαθηματική πηγή.

1.2 Απευθυνόμενο κοινό και στόχοι

Η NumStability απευθύνεται κυρίως σε ερευνητές αριθμητικής ανάλυσης και δημιουργούς τυπικών αποδείξεων που χρησιμοποιούν τη Lean, σε δημιουργούς επαληθευμένων αριθμητικών αλγορίθμων, καθώς και σε ανθρώπους ή πράκτορες τεχνητής νοημοσύνης που χρειάζονται έτοιμους ορισμούς και λήμματα για νέες αποδείξεις. Μπορεί επίσης να αξιοποιηθεί από συντηρητές βιβλιοθηκών τυπικών μαθηματικών, όταν εξετάζουν ποια αποτελέσματα μπορούν να γενικευθούν ή να μεταφερθούν σε κοινή υποδομή της Mathlib. Στην παρούσα μορφή της, η NumStability δεν παρέχει εκτελέσιμες αριθμητικές ρουτίνες για άμεση ενσωμάτωση σε εφαρμογές. Πρόκειται για βιβλιοθήκη τυπικών ορισμών και θεωρημάτων, δηλαδή υποδομή για την κατασκευή και τον έλεγχο αποδείξεων.

Οι στόχοι διακρίνονται σε τέσσερα επίπεδα:

1. Στο επίπεδο της παρούσας διπλωματικής εργασίας, κύριος στόχος είναι η κατασκευή της βιβλιοθήκης. Συμπληρωματικός στόχος είναι η αξιολόγηση μέσω του HighamBench, η οποία εξετάζει αν η πρόσβαση στη βιβλιοθήκη μειώνει το κόστος νέων αποδείξεων, τόσο για δοσμένους στόχους Lean όσο και για νέες, πιστές τυποποιήσεις αποτελεσμάτων από την πηγή τους.
2. Στο επίπεδο της βιβλιοθήκης, στόχος είναι ένα επαναχρησιμοποιήσιμο ενδιάμεσο στρώμα ανάμεσα στη γενική υποδομή της Mathlib και σε συγκεκριμένα θεωρήματα αριθμητικών αλγορίθμων: μοντέλα στρογγυλοποίησης, λήμματα γ_n, κατάσταση προβλήματος, διαταραχές και θεωρήματα σταθερότητας.
3. Στον μακροπρόθεσμο επιστημονικό ορίζοντα, στόχος είναι η μετάβαση από ένα μαθηματικό θεώρημα σε μηχανικά ελεγμένη εγγύηση για συγκεκριμένο αλγόριθμο πεπερασμένης ακρί-

βειας, με προοπτική και τους τυχαιοποιημένους αριθμητικούς αλγορίθμους (Ενότητα 9.4).

4. Στο επίπεδο του ευρύτερου έργου, απώτερος στόχος είναι ένα γενικό σύστημα τυποποίησης βιβλίων με πράκτορες TN, του οποίου η NumStability και το HighamBench αποτελούν το πρώτο στάδιο (Ενότητα 9.3). Ο Higham είναι το πρώτο βιβλίο· το ίδιο σύστημα προορίζεται για τα επόμενα, με πρώτο το βιβλίο υψηλοδιάστατης πιθανότητας του Vershynin (2018) (Ενότητα 9.4).

1.3 Κίνητρο και ερευνητικό κενό

Τα τελευταία χρόνια η τυποποίηση μαθηματικών στη Lean παρουσιάζεται όλο και περισσότερο ως κοινή επιστημονική υποδομή και όχι μόνο ως απομονωμένη άσκηση επαλήθευσης (Tao, 2025). Η πρωτοβουλία Palomar καταγράφει αμετάβλητες εκδόσεις αποθετηρίων και διαχωρίζει τον μηχανικό έλεγχο της Lean από τον σημασιολογικό έλεγχο της άτυπης περιγραφής. Ο δεύτερος ανατίθεται σε μεγάλο γλωσσικό μοντέλο και οι ίδιοι οι συντελεστές διευκρινίζουν ότι δεν ισοδυναμεί με ανθρώπινη αξιολόγηση από ομοτίμους (Avigad, Ballard κ.ά., 2026; Palomar Registry, 2026; Tao, 2026). Παράλληλα, η CSLib διατυπώνει ρητά τον στόχο δημιουργίας μιας αντίστοιχης επαναχρησιμοποιήσιμης βάσης για την επιστήμη υπολογιστών (Barrett κ.ά., 2026). Τον Σεπτέμβριο του 2026, η OpenAI συνόδευσε την ανακοίνωση μιας προτεινόμενης λύσης για το πρόβλημα Navier–Stokes με τυποποίηση στη Lean. Η ανακοίνωση αυτή δεν χρησιμοποιείται εδώ ως ανεξάρτητη επικύρωση της μαθηματικής λύσης, αλλά ως πρόσφατη ένδειξη της ζήτησης για αποτελέσματα που διατίθενται μαζί με μηχανικά ελέγξιμα τυπικά αντικείμενα (OpenAI, 2026e).

Η επαναχρησιμοποίηση προσφέρει επίσης ευκαιρίες εντοπισμού σφαλμάτων. Ο Kontorovich (2026), συνοψίζοντας συνομιλία του με τον Christian Szegedy, παρουσιάζει —χωρίς να τον υιοθετεί— τον ισχυρισμό ότι ένα σφάλμα σε χρήσιμο ορισμό ή θεώρημα τείνει να αποκαλύπτεται όταν μεταγενέστερα αποτελέσματα δεν μπορούν να στηριχθούν σε αυτό όπως αναμενόταν (Kontorovich, 2025, χρονικό σημείο 26:45–28:27). Πρόκειται για πρόσθετο μηχανισμό εντοπισμού και διόρθωσης, όχι για εγγύηση αρχικής πιστότητας.

Ο ειδικός στόχος της παρούσας εργασίας είναι διαφορετικός από την πιστοποίηση μιας μεμονωμένης αριθμητικής έκφρασης ή την κατασκευή πλήρους σημασιολογίας του προτύπου IEEE 754. Είναι η δημιουργία μιας επαναχρησιμοποιήσιμης βιβλιοθήκης στη Lean που εκφράζει το συνθετικό ύφος του Higham: από το αφηρημένο μοντέλο στρογγυλοποίησης έως τις παραγοντοποιήσεις και τα φράγματα διαταραχών. Η αναλυτική σύγκριση με τα προϋπάρχοντα έργα γίνεται στο Κεφάλαιο 4.

Η σύνδεση της Lean 4 με τη Mathlib, ο μεταπρογραμματισμός και τα εργαλεία ανάπτυξης επιτρέπουν στη βιβλιοθήκη να λειτουργεί τόσο ως αποθετήριο ελεγμένων λημμάτων όσο και ως υποδομή για υποβοηθούμενη από TN εργασία (de Moura & Ullrich, 2021; The mathlib Community, 2020; Yang κ.ά., 2023). Η επιλογή αιτιολογείται συνοπτικά στο Κεφάλαιο 3.

1.4 Κύρια παραδοτέα της εργασίας

Η υλοποίηση των παραπάνω στόχων αποτυπώνεται σε τρία βασικά παραδοτέα.

1.4.1 Τυπική βιβλιοθήκη και τελική έκδοση

Η NumStability οργανώνει ένα αφηρημένο `FPModel`, βασικές στρογγυλοποιημένες πράξεις και αλυσίδες θεωρημάτων για αθροίσεις, εσωτερικά γινόμενα, πράξεις πίνακα–διανύσματος και πίνακα–πίνακα, τριγωνικές επιλύσεις και τμήματα παραγοντοποιήσεων. Αξιοποιεί τους πίνακες, τις νόρμες και την ακριβή άλγεβρα της Mathlib αντί να δημιουργεί δεύτερη, ασύμβατη θεμελίωση.

Η τελική έκδοση `higham_v01` διαχωρίζει την υποδομή κατά μαθηματικό πεδίο από την αντιστοίχιση προς τα κεφάλαια του Higham. Η αρχιτεκτονική και οι έλεγχοι της έκδοσης παρουσιάζονται στο Κεφάλαιο 5. Η βιβλιοθήκη, τα δεδομένα των αξιολογήσεων και τα τεκμήρια της κατασκευής διατίθενται στο δημόσιο αποθετήριο `NumStability`, στο οποίο παραπέμπουν οι υποσημειώσεις της μορφής `commit:διαδρομή`.[1]

1.4.2 Τεκμηριωμένη διαδικασία κατασκευής και ελέγχου

Η εργασία καταγράφει πώς οι ανθρώπινες διορθώσεις της αρχικής τυποποίησης μετατράπηκαν σε κανόνες για τους πράκτορες, πώς οργανώθηκε η παράλληλη εργασία ανά κεφάλαιο και ποια ευρήματα προέκυψαν από τους μεταγενέστερους ελέγχους. Το Κεφάλαιο 6 παρουσιάζει αυτή τη διαδικασία μαζί με τα τεκμήριά της.

1.4.3 Οι δύο αξιολογήσεις του HighamBench

Το HighamBench οργανώνεται ως πείραμα δύο συνθηκών. Στη συνθήκη N ο πράκτορας διαθέτει τη Lean 4 και τη Mathlib χωρίς τη NumStability, ενώ στη συνθήκη L λαμβάνει επιπλέον σταθερή έκδοση της βιβλιοθήκης και σύντομη οδηγία να την αξιοποιήσει. Εκτελέστηκαν δύο αξιολογήσεις. Στο HighamBench-Proof ο πράκτορας αποδεικνύει 30 σταθερούς στόχους Lean από 25 δημοσιευμένα άρθρα, με τρεις επαναλήψεις ανά συνθήκη. Στο HighamBench-E2E ο πράκτορας διατυπώνει δέκα αποτελέσματα από πέντε πηγές σε δηλώσεις Lean, που ελέγχονται από ανεξάρτητους κριτές πιστότητας βασισμένους σε γλωσσικά μοντέλα, και έπειτα αποδεικνύει την οριστική δήλωση. Και στις δύο μετρώνται ο ενεργός χρόνος, τα tokens, το μέγεθος του κώδικα και η πραγματική χρήση δηλώσεων της βιβλιοθήκης στις αποδεκτές αποδείξεις.

Στις δύο αξιολογήσεις η L χρειάστηκε συνολικά 35,0% και 22,5% λιγότερο χρόνο αντίστοιχα, και το όφελος συσχετίζεται θετικά με την πραγματική χρήση της βιβλιοθήκης: ο συντελεστής Spearman μεταξύ χρήσης και κέρδους χρόνου απόδειξης είναι +0,69 και +0,56 αντίστοιχα, και μεταξύ χρήσης και μείωσης γραμμών κώδικα +0,61 για το πλήρες αρχείο του HighamBench-Proof και +0,64 για την απόδειξη του HighamBench-E2E. Τα κέρδη συγκεντρώνονται όμως στις εργασίες που καλύπτει η βιβλιοθήκη, ενώ στο HighamBench-E2E η L χρησιμοποίησε περισσότερα tokens (Κεφάλαιο 8).[2,3]

[1] https://github.com/AlexGeorgantzas/NumStability. Οι αφαιρέσεις κειμένου πηγών και η αντικατάσταση τοπικών διαδρομών και ονομάτων υπολογιστών καταγράφονται στα αρχεία `REDACTIONS.md`. Το ιστορικό ανάπτυξης διατηρείται στο https://github.com/AlexGeorgantzas/lean-numerical-stability.

[2] *Αναφορά της αξιολόγησης των 30 σταθερών στόχων.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:paper_benchmark/highambench/reports/full-report-scope30/report.tex`.

[3] *Αναφορά της αξιολόγησης των δέκα εργασιών τυποποίησης και απόδειξης.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:Documentation/report.tex`.

1.5 Όρια των ισχυρισμών

Για την ακριβή αποτίμηση των αποτελεσμάτων απαιτείται σαφής οριοθέτηση των ισχυρισμών. Πρώτον, το `FPModel` είναι σχεδιασμένο για κλασική ανάλυση σχετικού σφάλματος πάνω σε πραγματικές τιμές. Δεν μοντελοποιεί πλήρως τη σημασιολογία του IEEE 754 (Ενότητα 2.1.1). Δεύτερον, η παρουσία επιλεγμένου υλικού από όλα τα κεφάλαια του βιβλίου δεν συνιστά πλήρη τυποποίηση κάθε σελίδας, άσκησης ή αλγορίθμου. Τρίτον, ένα θεώρημα που μεταφέρει μια ήδη υποτεθειμένη αφηρημένη προδιαγραφή δεν είναι ισοδύναμο με απόδειξη ότι η πραγματική στρογγυλοποιημένη υλοποίηση ικανοποιεί την προδιαγραφή.

Τέταρτον, οι έλεγχοι πιστότητας έγιναν από πράκτορες: δεν έχει γίνει συστηματική ανθρώπινη αντιπαραβολή των δηλώσεων ολόκληρης της βιβλιοθήκης με τις πηγές τους. Ούτε οι κρίσεις πιστότητας του HighamBench-E2E επανεξετάστηκαν από άνθρωπο. Πέμπτον, οι δύο αξιολογήσεις περιορίζονται σε 30 και 10 εργασίες αντίστοιχα, σε συγκεκριμένες διαμορφώσεις μοντέλων και στα δηλωμένα όρια πόρων. Οι συγκρίσεις, επομένως, δεν γενικεύονται σε όλους τους πράκτορες TN ή σε όλα τα θεωρήματα αριθμητικής σταθερότητας.

1.6 Δομή της εργασίας

Το Κεφάλαιο 2 εισάγει την αριθμητική κινητής υποδιαστολής, το καθιερωμένο μοντέλο, τις ποσότητες γ_n, την κατάσταση προβλήματος και τις αναλύσεις forward και backward σφάλματος. Το Κεφάλαιο 3 παρουσιάζει τους βοηθούς αποδείξεων, αιτιολογεί την επιλογή της Lean και ορίζει την κατανομή εμπιστοσύνης μεταξύ ανθρώπου, Codex και πυρήνα. Το Κεφάλαιο 4 τοποθετεί το έργο απέναντι στις κύριες βιβλιοθήκες, αναλυτές και συγκριτικές αξιολογήσεις.

Το Κεφάλαιο 5 περιγράφει το μοντέλο, τις συνθετικές αλυσίδες και την αρχιτεκτονική της NumStability. Το Κεφάλαιο 6 ανασυνθέτει τη διαδικασία κατάτμησης ανά κεφάλαιο, το βασικό αρχείο καθοδήγησης και τους διαδοχικούς ελέγχους. Το Κεφάλαιο 7 ορίζει τη μεθοδολογία των δύο αξιολογήσεων του HighamBench, τις συνθήκες, τις μετρήσεις και τον έλεγχο πιστότητας, και το Κεφάλαιο 8 παρουσιάζει τα αποτελέσματά τους· τα πλήρη αριθμητικά στοιχεία και οι προτροπές βρίσκονται στα Παραρτήματα Β΄ και Γ΄. Τέλος, το Κεφάλαιο 9 συνθέτει τα συμπεράσματα, τους περιορισμούς και τις κατευθύνσεις μελλοντικής έρευνας και παρουσιάζει το σύστημα τυποποίησης βιβλίων στο οποίο εντάσσεται η εργασία, με πρώτο επόμενο βιβλίο αυτό του Vershynin.

Κεφάλαιο 2

Αριθμητική κινητής υποδιαστολής και αριθμητική σταθερότητα

2.1 Από τους πραγματικούς αριθμούς στον υπολογισμό

Η κλασική αριθμητική ανάλυση συχνά περιγράφει έναν αλγόριθμο ως ακολουθία πράξεων πάνω στους πραγματικούς αριθμούς. Ένας πραγματικός υπολογιστής, αντίθετα, διαθέτει πεπερασμένο σύνολο αναπαραστάσιμων τιμών. Ακόμη και όταν τα δεδομένα εισόδου είναι αναπαραστάσιμα, ένα ενδιάμεσο ακριβές αποτέλεσμα συνήθως δεν είναι και πρέπει να στρογγυλοποιηθεί. Η διάκριση αυτή απαιτεί τη χωριστή περιγραφή δύο όψεων του ίδιου υπολογισμού:

1. τη μαθηματική συνάρτηση ή το ακριβές αποτέλεσμα που ζητείται να υπολογιστεί·
2. τον υπολογισμό πεπερασμένης ακρίβειας, δηλαδή την ακολουθία πράξεων που πράγματι εκτελεί ο αλγόριθμος, μαζί με τις ενδιάμεσες στρογγυλοποιήσεις και το τελικό υπολογισμένο αποτέλεσμα.

Η ανάλυση σφαλμάτων συνδέει τις δύο περιγραφές. Δεν αρκεί να δηλωθεί ότι «υπάρχει σφάλμα στρογγυλοποίησης». Χρειάζεται ποσοτικό όριο, σαφές μέτρο μεγέθους, εμφανείς προϋποθέσεις και πληροφορία για το πώς το σφάλμα εξαρτάται από το πρόβλημα και τον αλγόριθμο. Η παρουσίαση ακολουθεί τη σύγχρονη σύνθεση του N. J. Higham (2002), η οποία συνδέει τα forward και backward σφάλματα με την κατάσταση του προβλήματος και τη σταθερότητα. Ειδικότερα, η backward ανάλυση σφάλματος, ένα από τα κεντρικά εργαλεία αυτής της σύνθεσης, ανήκει στην παράδοση που συστηματοποίησε και καθιέρωσε ο Wilkinson (1963).

2.1.1 Ένα ιδεατό σύστημα κινητής υποδιαστολής

Σε ένα τυπικό θεωρητικό σύστημα, ένας μη μηδενικός κανονικοποιημένος αριθμός κινητής υποδιαστολής γράφεται σχηματικά ως

$$y = \pm m\, \beta^{e}, \tag{2.1}$$

όπου β είναι η βάση, m η significand (ή mantissa) με πεπερασμένο πλήθος ψηφίων και e εκθέτης

σε πεπερασμένο εύρος. Για έναν σταθερό εκθέτη, δύο διαδοχικές significands διαφέρουν κατά μία μονάδα της τελευταίας αποθηκευμένης θέσης. Όταν όμως ο εκθέτης αυξηθεί κατά ένα, το αντίστοιχο απόλυτο βήμα πολλαπλασιάζεται επί β· στο δυαδικό σύστημα, δηλαδή, διπλασιάζεται. Έτσι, η απόλυτη απόσταση ανάμεσα σε γειτονικούς αναπαραστάσιμους αριθμούς αυξάνεται περίπου ανάλογα με το μέγεθός τους, ενώ η σχετική απόσταση παραμένει της ίδιας τάξης (N. J. Higham, 2002, Λήμμα 2.1). Συνεπώς, μακριά από underflow, η φυσική πρώτη προσέγγιση είναι σχετική και όχι απόλυτη: η στρογγυλοποίηση ενός αριθμού μεγέθους 10^{10} επιτρέπεται να έχει μεγαλύτερο απόλυτο σφάλμα από τη στρογγυλοποίηση ενός αριθμού μεγέθους 10^{-3}, αλλά παρόμοιο σχετικό σφάλμα.

Η παράμετρος *unit roundoff*, που συμβολίζεται εδώ με u, συνοψίζει τη χειρότερη σχετική επίδραση μιας σωστής στρογγυλοποίησης στο ιδεατό μοντέλο. Σε σύστημα βάσης β με t ψηφία και στρογγυλοποίηση στον πλησιέστερο αριθμό, $u = \frac{1}{2}\beta^{1-t}$. Επομένως, το u χαρακτηρίζει την επιλεγμένη αριθμητική εργασίας, αφού εξαρτάται από τον μορφότυπο, την ακρίβεια και τον κανόνα στρογγυλοποίησης· δεν είναι μία μοναδική ιδιότητα ολόκληρου του φυσικού υπολογιστή. Το ίδιο μηχάνημα, για παράδειγμα, έχει διαφορετικό u όταν χρησιμοποιεί binary32 και διαφορετικό όταν χρησιμοποιεί binary64. Το άνω όριο u προσεγγίζεται όταν η ακριβής τιμή βρίσκεται πολύ κοντά στο μέσο δύο γειτονικών αριθμών κινητής υποδιαστολής: τότε η στρογγυλοποίηση τη μετακινεί περίπου κατά το μισό του μεταξύ τους κενού. Το σχετικό σφάλμα πλησιάζει περισσότερο το u κοντά στο κάτω άκρο μιας περιοχής σταθερού εκθέτη. Στη διατύπωση του Higham το όριο για τη στρογγυλοποίηση μιας ακριβούς τιμής είναι αυστηρό, $|\delta| < u$, επομένως η χειρότερη περίπτωση συνήθως προσεγγίζεται χωρίς να επιτυγχάνεται ακριβώς (N. J. Higham, 2002, Θεώρημα 2.2).

Το IEEE 754 προσδιορίζει πολύ περισσότερα από αυτό το ιδεατό σχήμα: μορφότυπους, ειδικές τιμές, προσημασμένα μηδενικά, υποκανονικούς αριθμούς, τρόπους στρογγυλοποίησης, εξαιρέσεις και απαιτήσεις πράξεων (Institute of Electrical and Electronics Engineers [IEEE], 2019). Η διάκριση είναι σημαντική για την παρούσα εργασία. Το μοντέλο κατά Higham προσφέρει ισχυρή αφαίρεση για την ανάλυση αλγορίθμων, αλλά δεν ισοδυναμεί με πλήρη λειτουργική σημασιολογία του προτύπου.

2.2 Το καθιερωμένο μοντέλο σχετικού σφάλματος

Για μια βασική πράξη $\mathrm{op} \in \{+, -, \times, /\}$, το καθιερωμένο μοντέλο γράφεται

$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1 + \delta), \qquad |\delta| \le u. \tag{2.2}$$

Η ίδια μορφή χρησιμοποιείται για τη στρογγυλοποίηση μιας ακριβούς τιμής, $\mathrm{fl}(x) = x(1 + \delta)$.

Η εξίσωση (2.2) λειτουργεί ως αναλυτικό συμβόλαιο, δηλαδή ως αφηρημένη εγγύηση για το αποτέλεσμα κάθε βασικής πράξης. Η ανάλυση μπορεί να χρησιμοποιήσει μόνο το γεγονός ότι υπάρχει κάποιο δ με $|\delta| \le u$ που συνδέει το υπολογισμένο αποτέλεσμα με το ακριβές· δεν χρειάζεται να γνωρίζει τη δυαδική αναπαράσταση του αποτελέσματος ή τις λεπτομέρειες της υλοποίησης. Το δ είναι αναλυτικός μάρτυρας αυτής της σχέσης, όχι ποσότητα που αποθηκεύει ή υπολογίζει το μηχάνημα. Το συμβόλαιο δεν προσδιορίζει ποια τιμή πήρε το δ ούτε απαιτεί να

είναι η ίδια σε δύο πράξεις. Σε μια ακολουθία υπολογισμών εμφανίζονται διαφορετικές ποσότητες $\delta_1, \dots, \delta_n$, καθεμία με $|\delta_i| \le u$. Η ουσία της ανάλυσης είναι να ελεγχθεί η σύνθεσή τους χωρίς να γίνει η αντικατάσταση κάθε γινομένου από ανεξάρτητο σενάριο χειρότερης περίπτωσης.

2.2.1 Απόλυτα και componentwise όρια

Από την εξίσωση (2.2) προκύπτει αμέσως

$$|\mathrm{fl}(z) - z| \le u\,|z|\,. \tag{2.3}$$

Για διανύσματα και πίνακες, η ίδια ιδέα εφαρμόζεται συχνά ανά συνιστώσα. Αν A είναι πίνακας, η γραφή $|A|$ σημαίνει τον πίνακα των απόλυτων τιμών των στοιχείων του και μια ανισότητα όπως $|\Delta A| \le u\,|A|$ νοείται componentwise: για κάθε ζεύγος δεικτών i, j ισχύει

$$\left|\Delta a_{ij}\right| \le u\,\left|a_{ij}\right|. \tag{2.4}$$

Κάθε στοιχείο μπορεί να έχει διαφορετική διαταραχή. Ένα στοιχείο μεγάλου μέτρου επιτρέπεται να έχει αναλογικά μεγαλύτερη απόλυτη μεταβολή, αλλά η σχετική κλίμακα παραμένει η ίδια· αν $a_{ij} = 0$, η ανισότητα επιβάλλει $\Delta a_{ij} = 0$. Τα componentwise φράγματα διατηρούν πληροφορία για την κλίμακα κάθε στοιχείου, ενώ τα normwise φράγματα συμπυκνώνουν την πληροφορία σε έναν αριθμό. Κανένα από τα δύο δεν είναι καθολικά ισχυρότερο: η κατάλληλη μορφή εξαρτάται από το πρόβλημα και τον τρόπο με τον οποίο θα χρησιμοποιηθεί το φράγμα.

2.3 Συσσώρευση τοπικών σφαλμάτων

Η βασική τεχνική δυσκολία εμφανίζεται όταν πολλαπλασιάζονται πολλοί παράγοντες $1 + \delta_i$ ή τα αντίστροφά τους. Η επανειλημμένη ανάπτυξη δημιουργεί όρους όλων των τάξεων. Ο Higham τους συμπυκνώνει σε μια ποσότητα θ_n και ορίζει

$$\gamma_n = \frac{nu}{1 - nu}, \qquad nu < 1. \tag{2.5}$$

Μια τυπική μορφή του θεμελιώδους λήμματος είναι η εξής: αν $|\delta_i| \le u$ και $\rho_i \in \{-1, 1\}$, τότε, υπό την προϋπόθεση $nu < 1$,

$$\prod_{i=1}^{n} (1 + \delta_i)^{\rho_i} = 1 + \theta_n, \qquad |\theta_n| \le \gamma_n. \tag{2.6}$$

Η μορφή με $\rho_i = -1$ είναι ουσιώδης για backward αναλύσεις, όπου μια τοπική σχέση χρειάζεται να αντιστραφεί. Η προϋπόθεση $nu < 1$ εξασφαλίζει ότι ο παρονομαστής είναι θετικός και ότι το συγκεντρωτικό φράγμα χειρότερης περίπτωσης παραμένει πεπερασμένο. Για μικρό nu, ισχύει η πρώτη προσέγγιση $\gamma_n \approx nu$. Καθώς όμως το nu πλησιάζει το 1, η τιμή του γ_n αυξάνεται απότομα και το συγκεκριμένο όριο παύει να είναι πληροφοριακό. Αυτό δεν συνεπάγεται ότι το πραγματικό σφάλμα είναι αναγκαστικά μεγάλο, αλλά ότι απαιτείται λεπτομερέστερη ανάλυση. Η ακριβής

ρητή μορφή είναι προτιμότερη στην τυπική απόδειξη.

2.4 Κατάσταση του προβλήματος

Η *κατάσταση* ή ευαισθησία (conditioning) είναι ιδιότητα του μαθηματικού προβλήματος και των συγκεκριμένων δεδομένων, όχι του αλγορίθμου. Για μια συνάρτηση f, μια τυπική τοπική μορφή σχετικού condition number είναι

$$\kappa_f(x) := \lim_{\varepsilon \to 0^+} \sup_{\|\Delta x\| \le \varepsilon \|x\|} \frac{\|f(x+\Delta x) - f(x)\|}{\varepsilon \, \|f(x)\|}. \tag{2.7}$$

Όταν το f είναι διαφορίσιμο και οι παρονομαστές είναι μη μηδενικοί, η παράγωγος δίνει την αντίστοιχη τοπική έκφραση. Η επιλογή του μοντέλου διαταραχών είναι μέρος του ορισμού: normwise, componentwise, δομημένα ή μη δομημένα μοντέλα μπορούν να οδηγήσουν σε διαφορετικά condition numbers. Αφού καθοριστούν το μοντέλο διαταραχών, η νόρμα και η κλιμάκωση, μικρότερο condition number είναι γενικά επιθυμητό, επειδή σημαίνει μικρότερη ενίσχυση των διαταραχών. Δεν υπάρχει όμως ένα καθολικό αριθμητικό κατώφλι που να διακρίνει τα προβλήματα καλής από εκείνα της κακής κατάστασης.

Για το γραμμικό σύστημα $Ax = b$ με αντιστρέψιμο A, το γνωστό normwise condition number είναι

$$\kappa(A) = \|A\| \left\|A^{-1}\right\|. \tag{2.8}$$

Για επαγόμενη νόρμα πίνακα ισχύει $\kappa(A) \ge 1$, επομένως μια τιμή κοντά στο 1 είναι η καλύτερη περίπτωση. Μεγάλο $\kappa(A)$ σημαίνει ότι υπάρχουν μικρές επιτρεπτές διαταραχές που μπορούν να προκαλέσουν μεγάλη μεταβολή στη λύση. Δεν σημαίνει ότι κάθε δεδομένη είσοδος θα εμφανίσει τη χειρότερη περίπτωση ούτε ότι ο αλγόριθμος είναι ασταθής. Αντίστροφα, μικρό condition number δεν σώζει έναν αλγόριθμο που εισάγει πολύ μεγαλύτερο backward σφάλμα από αυτό που δικαιολογεί η ακρίβεια. Στην πράξη ενδιαφέρει ο συνδυασμός των δύο μεγεθών: για έναν backward σταθερό υπολογισμό με τοπικά σφάλματα τάξης u, ένας χρήσιμος πρώτος δείκτης είναι το κu σε σχέση με την απαιτούμενη ακρίβεια (N. J. Higham, 2002, κεφ. 1).

2.5 Forward, backward και mixed σφάλματα

Έστω ακριβές πρόβλημα $y = f(x)$ και υπολογισμένο αποτέλεσμα $\hat{y}$. Το *forward σφάλμα* μετρά απευθείας την απόσταση $\|\hat{y} - y\|$, απόλυτα ή σχετικά.

Το *backward σφάλμα* αναζητά μια διαταραχή Δx τέτοια ώστε

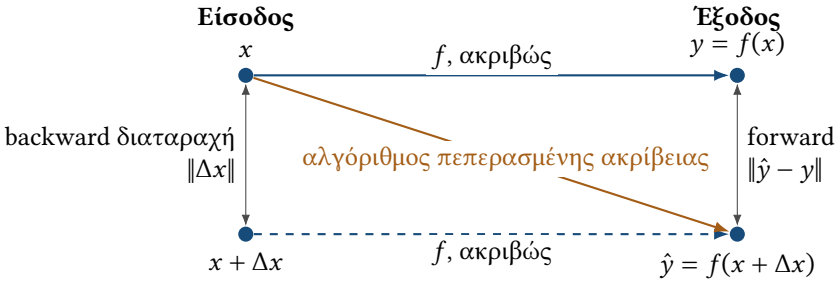
$$\hat{y} = f(x + \Delta x). \tag{2.9}$$

Ο αλγόριθμος είναι backward σταθερός, με την κατάλληλη έννοια, όταν μπορεί να επιλεγεί μικρή Δx, συγκρίσιμη με τα τοπικά σφάλματα της αριθμητικής. Η ακριβής έννοια του «μικρή»

πρέπει να δηλώνει νόρμα, κλιμάκωση και δομή της διαταραχής. Για ορισμένα προβλήματα απαιτείται *mixed* ανάλυση: το υπολογισμένο αποτέλεσμα είναι κοντά στην ακριβή έξοδο ενός κοντινού προβλήματος, χωρίς να ταυτίζεται ακριβώς μαζί της.

Το Σχήμα 2.1 διακρίνει τη διαταραχή στην είσοδο από το σφάλμα στην έξοδο. Η κατάσταση του προβλήματος συνδέει τα δύο μεγέθη: σε πρώτη τάξη,

$$\text{σχετικό forward σφάλμα} \lesssim \text{condition number} \times \text{σχετικό backward σφάλμα}. \tag{2.10}$$



Σχήμα 2.1: Δύο οπτικές του ίδιου υπολογισμένου αποτελέσματος. Η διαγώνιος παριστά τον αλγόριθμο και τα οριζόντια βέλη την ακριβή συνάρτηση. Οι κατακόρυφες αποστάσεις είναι απόλυτα normwise μεγέθη. Η κάτω διαδρομή δείχνει μία επιτρεπτή διαταραχή Δx, όταν υπάρχει τέτοια ώστε $\hat{y} = f(x + \Delta x)$ (N. J. Higham, 2002, κεφ. 1).

2.5.1 Residual και backward σφάλμα γραμμικού συστήματος

Η περίπτωση του γραμμικού συστήματος χρησιμοποιείται εδώ ως αντιπροσωπευτικό παράδειγμα της γενικής σχέσης ανάμεσα σε residual, backward σφάλμα, condition number και forward σφάλμα. Η ίδια διάκριση εφαρμόζεται και σε άλλα αριθμητικά προβλήματα, με διαφορετικό κάθε φορά ορισμό του residual και των επιτρεπτών διαταραχών.

Για μια προσεγγιστική λύση $\hat{x}$ του $Ax = b$, το residual είναι

$$r = b - A\hat{x}. \tag{2.11}$$

Αν θεωρηθεί ότι μόνο το δεξί μέλος διαταράσσεται, τότε $A\hat{x} = b - r$: το residual δίνει άμεσα backward διαταραχή $\Delta b = -r$. Αν επιτρέπονται διαταραχές και στο A και στο b, ένα κλασικό normwise μέτρο έχει τη μορφή

$$\eta(\hat{x}) = \frac{\|r\|}{\|A\|\,\|\hat{x}\| + \|b\|}, \tag{2.12}$$

υπό τις κατάλληλες συμβατές νόρμες και μη εκφυλιστικές περιπτώσεις (N. J. Higham, 2002, Θεώρημα 7.1, σ. 120). Το μικρό residual μόνο του δεν συνεπάγεται μικρό forward σφάλμα για πίνακα A κακής κατάστασης. Σε συνδυασμό με την $\kappa(A)$, όμως, οδηγεί σε forward εκτίμηση.

Η residual οπτική είναι ιδιαίτερα κατάλληλη για τυπική επαλήθευση, επειδή χωρίζει δύο επίπεδα. Το επίπεδο του αλγορίθμου αποδεικνύει μια σχέση για τον residual ή μια παραγοντο-

ποίηση διαταραχής. Το επίπεδο των διαταραχών μεταφέρει αυτή τη σχέση σε forward σφάλμα υπό προϋποθέσεις για την κατάσταση του προβλήματος. Η σύνθεση είναι ασφαλής μόνο αν οι διαστάσεις, οι νόρμες και οι προϋποθέσεις μικρότητας των δύο επιπέδων ταιριάζουν ακριβώς.

2.6 Παράδειγμα: αναδρομική άθροιση

Έστω $x_1, \dots, x_n \in \mathbb{R}$ και ο ακριβής στόχος

$$s_n = \sum_{i=1}^{n} x_i. \tag{2.13}$$

Ο απλός αλγόριθμος, που αθροίζει από αριστερά προς τα δεξιά, σχηματίζει $\hat{s}_1 = x_1$ και

$$\hat{s}_k = \mathrm{fl}(\hat{s}_{k-1} + x_k), \qquad k = 2, \dots, n. \tag{2.14}$$

Πρόκειται για την κλασική αναδρομική άθροιση (N. J. Higham, 1993, §2, εξ. (2.1)).

Κάθε βήμα ικανοποιεί $\hat{s}_k = (\hat{s}_{k-1} + x_k)(1+\delta_k)$ με $|\delta_k| \le u$. Η πλήρης ανάπτυξη δείχνει ότι οι πρώτοι όροι πολλαπλασιάζονται από περισσότερους παράγοντες σφάλματος από τους τελευταίους. Η συμπύκνωση με το λήμμα γ_n δίνει το κλασικό forward φράγμα

$$|\hat{s}_n - s_n| \le \gamma_{n-1} \sum_{i=1}^{n} |x_i|, \qquad (n-1)u < 1. \tag{2.15}$$

Η ανισότητα (2.15) είναι το κλασικό, ανεξάρτητο από τη σειρά, forward φράγμα της αναδρομικής άθροισης (N. J. Higham, 1993, §2, εξ. (2.6)). Συγκρίνει το αποτέλεσμα $\hat{s}_n$ που υπολόγισε ο αλγόριθμος με το ακριβές αποτέλεσμα s_n για τα ίδια αρχικά δεδομένα $x_1, \dots, x_n$. Ο παράγοντας γ_{n-1} συνδυάζει το unit roundoff u της επιλεγμένης αριθμητικής εργασίας με το πλήθος των στρογγυλοποιημένων προσθέσεων. Το u χαρακτηρίζει το τοπικό μοντέλο πράξης, ενώ το γ_{n-1} είναι εκτίμηση χειρότερης περίπτωσης για τη συσσώρευση των τοπικών σφαλμάτων στα $n-1$ βήματα του συγκεκριμένου αλγορίθμου.

Αν $s_n \neq 0$, η διαίρεση της ανισότητας (2.15) με $|s_n|$ δίνει το σχετικό forward φράγμα

$$\frac{|\hat{s}_n - s_n|}{|s_n|} \le \gamma_{n-1} \underbrace{\frac{\sum_{i=1}^{n} |x_i|}{\left|\sum_{i=1}^{n} x_i\right|}}_{\kappa_{\mathrm{sum}}(x)}. \tag{2.16}$$

Ο λόγος $\kappa_{\mathrm{sum}}(x)$ προκύπτει ακριβώς κατά τη μετατροπή του απόλυτου φράγματος σε σχετικό. Είναι ο σχετικός **componentwise condition number** της άθροισης στα συγκεκριμένα δεδομένα και μετρά πόσο μπορεί να μεταβληθεί το άθροισμα όταν κάθε x_i υποστεί μικρή σχετική διαταραχή. Αν $s_n = 0$, το σχετικό forward σφάλμα δεν ορίζεται και παραμένει χρήσιμο το απόλυτο φράγμα (2.15) (N. J. Higham, 1993, §2, σ. 785).

Ως άμεση εξειδίκευση του παραπάνω condition number, η επίδραση της ακύρωσης φαίνεται ήδη για $x_1 = 1$ και $x_2 = -1 + \varepsilon$, όπου $0 < \varepsilon \ll 1$. Το ακριβές άθροισμα είναι $s_2 = \varepsilon$, ενώ $|x_1| + |x_2| = 2 - \varepsilon$.

Επομένως,

$$\kappa_{\mathrm{sum}}(x) = \frac{2-\varepsilon}{\varepsilon} \approx \frac{2}{\varepsilon}. \tag{2.17}$$

Οι δύο μεγάλοι όροι σχεδόν αλληλοαναιρούνται και αφήνουν ένα πολύ μικρό αποτέλεσμα. Μικρές σχετικές μεταβολές των αρχικών όρων μπορούν συνεπώς να είναι μεγάλες σε σύγκριση με αυτό το αποτέλεσμα.

Η ίδια ανάπτυξη των τοπικών σφαλμάτων δίνει και τη διαφορετική backward ερμηνεία

$$\hat{s}_n = \sum_{i=1}^{n} x_i(1+\eta_i), \qquad |\eta_i| \le \gamma_{n-1}, \tag{2.18}$$

με δυνατότητα οξύτερων φραγμάτων που εξαρτώνται από τη θέση του κάθε όρου. Η εξίσωση (2.18) είναι backward δήλωση επειδή μεταβάλλει νοητά τα δεδομένα εισόδου αντί να συγκρίνει το υπολογισμένο και το ακριβές αποτέλεσμα: το $\hat{s}_n$ είναι το ακριβές άθροισμα των κοντινών αριθμών $x_i(1+\eta_i)$ (N. J. Higham, 2002, §4.2, σ. 82).

Οι δύο εκτιμήσεις περιγράφουν διαφορετικούς παράγοντες. Το backward φράγμα ελέγχει το μέγεθος των διαταραχών εισόδου που μπορούν να αποδοθούν στον αλγόριθμο, ενώ το $\kappa_{\mathrm{sum}}(x)$ ελέγχει πόσο ενισχύει αυτές τις διαταραχές το μαθηματικό πρόβλημα. Έτσι, ένας αλγόριθμος μπορεί να έχει μικρό backward σφάλμα και παρ' όλα αυτά μεγάλο σχετικό forward σφάλμα όταν το άθροισμα είναι κακής κατάστασης λόγω έντονης ακύρωσης.

Η σειρά άθροισης, η pairwise summation και οι αντισταθμισμένοι αλγόριθμοι αλλάζουν τη δομή του δέντρου σφαλμάτων και μπορούν να δώσουν διαφορετικά φράγματα. Για την αναδρομική άθροιση, η NumStability συνδέει ρητά το θεώρημα με τη συγκεκριμένη σειρά πράξεων. Ο ορισμός `fl_recursiveSum` είναι αριστερή αναδίπλωση (left fold) `Fin.foldl` με πράξη την `fp.fl_add`· χάρη στον πρόσθετο νόμο `fl_add_zero` (Ενότητα 5.2), τα θεωρήματα `recursiveSum_backward_error` και `recursiveSum_forward_error_bound` χρησιμοποιούν το γ_{n-1}, υπό την προϋπόθεση `gammaValid`. Πρόκειται επομένως για αποτέλεσμα συνδεδεμένο με ένα ρητά ορισμένο υπολογιστικό σχήμα ως προς το αφηρημένο `FPModel`, όχι για πιστοποίηση συγκεκριμένης υλοποίησης του IEEE 754.[1]

2.7 Εσωτερικό γινόμενο και πράξεις πινάκων

Για $x, y \in \mathbb{R}^n$, το ακριβές εσωτερικό γινόμενο είναι $x^{\mathsf{T}}y = \sum_{i=1}^{n} x_i y_i$. Στο συνηθισμένο μοντέλο, τόσο οι πολλαπλασιασμοί όσο και οι προσθέσεις στρογγυλοποιούνται. Με τις συμβάσεις μέτρησης πράξεων του Higham, προκύπτει το componentwise forward φράγμα

$$\left|\mathrm{fl}(x^{\mathsf{T}}y) - x^{\mathsf{T}}y\right| \le \gamma_n \, |x|^{\mathsf{T}} |y|, \tag{2.19}$$

υπό την αντίστοιχη προϋπόθεση $nu < 1$ (N. J. Higham, 2002, κεφ. 3). Η δεξιά πλευρά είναι

[1] *Υπολογιστικό σχήμα και forward/backward φράγματα της αναδρομικής άθροισης.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/Summation/Recursive/Core.lean`.

το γ_n επί το άθροισμα των απόλυτων τιμών των επιμέρους γινομένων. Αν το ακριβές εσωτερικό γινόμενο προκύπτει από ακύρωση, το απόλυτο φράγμα παραμένει χρήσιμο αλλά το σχετικό forward σφάλμα μπορεί να είναι μεγάλο.

Η εξίσωση (2.19) λειτουργεί ως παραγωγός για ανώτερα αποτελέσματα. Αν κάθε γραμμή του $A \in \mathbb{R}^{m\times n}$ πολλαπλασιαστεί με το $x \in \mathbb{R}^n$ μέσω του ίδιου executor εσωτερικού γινομένου, τότε

$$|\mathrm{fl}(Ax) - Ax| \le \gamma_n \, |A| \, |x| \,. \tag{2.20}$$

Αντίστοιχα, για τον κλασικό πολλαπλασιασμό πινάκων,

$$|\mathrm{fl}(AB) - AB| \le \gamma_n \, |A| \, |B| \,, \tag{2.21}$$

όπου n είναι η εσωτερική διάσταση. Αυτές οι σχέσεις δείχνουν γιατί η συνθετική οργάνωση είναι χρήσιμη (Ενότητα 5.4.2).

Η μετάβαση από componentwise φράγματα σε νόρμες απαιτεί ακριβείς ιδιότητες της επιλεγμένης νόρμας. Για συμβατές νόρμες, για παράδειγμα, $\|Ax\| \le \|A\| \, \|x\|$. Η τυπική απόδειξη οφείλει να γνωρίζει ποια νόρμα πίνακα χρησιμοποιείται, πώς συνδέεται με τη νόρμα διανύσματος και αν η μονοτονικότητα ως προς componentwise απόλυτες τιμές είναι διαθέσιμη. Για αυτό η NumStability χρησιμοποιεί, όπου είναι δυνατό, τη θεωρία πινάκων και νορμών της Mathlib (Ενότητα 3.2).

2.8 Αριθμητική σταθερότητα ως ιδιότητα αλγορίθμου

Ο όρος «σταθερός αλγόριθμος» χρησιμοποιείται με περισσότερες από μία συγγενείς έννοιες. Εκτός από τη backward και τη mixed σταθερότητα (Ενότητα 2.5), ένας forward σταθερός αλγόριθμος επιστρέφει αποτέλεσμα κοντά στην ακριβή λύση, με φράγμα συγκρίσιμο με εκείνο που επιτρέπει η κατάσταση του προβλήματος. Η σωστή δήλωση εξαρτάται από το πρόβλημα· δεν υπάρχει ένα ενιαίο κατηγόρημα που να καλύπτει χωρίς παραμέτρους όλες τις περιπτώσεις (N. J. Higham, 2002, κεφ. 1).

Αυτή η παρατήρηση καθορίζει ποια στοιχεία πρέπει να εξετάζονται όταν αξιολογείται η ισχύς ενός θεωρήματος σταθερότητας. Ανάλογα με το πρόβλημα, η δήλωση χρειάζεται να καταγράφει:

- το είδος της επιτρεπτής διαταραχής·
- τη νόρμα ή componentwise τάξη·
- την ακριβή κλιμάκωση του φράγματος·
- τις προϋποθέσεις αντιστρεψιμότητας, rank ή nonbreakdown·
- το συγκεκριμένο εκτελέσιμο αντικείμενο που παρήγαγε το αποτέλεσμα.

2.9 Τι ελέγχει το HighamBench

Το μαθηματικό υπόβαθρο καθορίζει το περιεχόμενο των εργασιών του HighamBench: σύνθεση τοπικών σφαλμάτων, φράγματα για διανύσματα και πίνακες, backward διαταραχές και μεταφορά τους σε forward εκτιμήσεις μέσω της κατάστασης. Η αξιολόγηση εξετάζει αν η πρόσβαση στη βιβλιοθήκη διευκολύνει νέες αποδείξεις με αυτά τα εργαλεία, υπό τις κατάλληλες προϋποθέσεις μικρότητας, rank και συμβατότητας νορμών. Οι πειραματικές συνθήκες και οι μετρήσεις παρουσιάζονται στο Κεφάλαιο 7.

2.10 Σύνοψη

Η αριθμητική σταθερότητα δεν είναι συνώνυμη της μικρής αριθμητικής απόκλισης σε μια δοκιμαστική είσοδο. Είναι ποσοτική σχέση μεταξύ του τοπικού μοντέλου στρογγυλοποίησης, της δομής του αλγορίθμου, του backward σφάλματος και της κατάστασης του προβλήματος. Οι ποσότητες γ_n συμπυκνώνουν τη συσσώρευση πολλών τοπικών σφαλμάτων, ενώ οι componentwise και normwise μορφές μεταφέρουν αυτή την πληροφορία σε πράξεις γραμμικής άλγεβρας. Το επόμενο κεφάλαιο εξετάζει ποιο λογικό και τεχνολογικό περιβάλλον είναι κατάλληλο για να γίνουν αυτές οι αλυσίδες μηχανικά ελεγμένες, επαναχρησιμοποιήσιμες και αξιολογήσιμες.

Κεφάλαιο 3

Η Lean, το Codex και τα όρια εμπιστοσύνης

3.1 Από τη μαθηματική αιτιολόγηση στον ελέγξιμο όρο απόδειξης

Στην τυπική επαλήθευση (Ενότητα 1.1), η απόδειξη αναπαρίσταται ως αντικείμενο που μπορεί να ελεγχθεί από έναν μικρό και σαφώς οριοθετημένο πυρήνα. Σε έναν βοηθό αποδείξεων, όπως η Lean, ο χρήστης δηλώνει ορισμούς και προτάσεις σε τυπική γλώσσα, ενώ οι τακτικές και η αυτοματοποίηση βοηθούν στην κατασκευή της απόδειξης. Η τελική αποδοχή δεν βασίζεται στην αυθεντία του χρήστη, αλλά στον επανέλεγχο του όρου σύμφωνα με τους κανόνες του λογικού πυρήνα.

Μια συμβατική απόδειξη σε χαρτί μπορεί να συμπτύσσει σε μία γραμμή μια ακολουθία μετασχηματισμών ανισοτήτων, να παραλείπει τις προϋποθέσεις ενός λήμματος ή να χρησιμοποιεί σιωπηρά κάποια υπόθεση. Κατά την τυποποίηση σε έναν βοηθό αποδείξεων, κάθε δομικό μαθηματικό στοιχείο πρέπει είτε να αιτιολογηθεί είτε να αναχθεί σε ήδη ελεγμένο αποτέλεσμα. Όταν ο πυρήνας δεχθεί τον όρο απόδειξης, η τυπική πρόταση έχει αποδειχθεί ως προς τους δηλωμένους ορισμούς και τις υποθέσεις. Η τυποποίηση καθιστά σαφές ποια ακριβώς πρόταση αποδεικνύεται, από ποιες υποθέσεις και μέσω ποιων εξαρτήσεων.

Βασική εννοιολογική περιγραφή αυτού του μηχανισμού είναι η αντιστοιχία Curry–Howard: οι προτάσεις αντιμετωπίζονται ως τύποι και οι αποδείξεις ως όροι των αντίστοιχων τύπων. Η αντιστοιχία αυτή διατυπώνεται στο πλαίσιο της θεωρίας τύπων. Μια πρόταση της μορφής $P \to Q$ αντιστοιχεί σε τύπο συναρτήσεων που λαμβάνουν έναν όρο του τύπου P και επιστρέφουν έναν όρο του τύπου Q. Οι ποσοδείκτες, οι εξαρτημένοι τύποι και οι επαγωγικοί ορισμοί επιτρέπουν να εκφραστούν όχι μόνο απλές προτασιακές σχέσεις αλλά και δομές όπως διανύσματα συγκεκριμένης διάστασης, πίνακες, αλγόριθμοι και ιδιότητες που εξαρτώνται από τις παραμέτρους τους (Avigad, de Moura κ.ά., 2026, κεφ. 3). Για τη συγκεκριμένη εργασία, αυτή η ενιαία αντιμετώπιση προγράμματος, μαθηματικού αντικειμένου και απόδειξης είναι χρήσιμη: ο ίδιος φορμαλισμός μπορεί να περιγράψει το ακριβές αποτέλεσμα ενός αλγορίθμου, ένα αφηρημένο μοντέλο στρογγυλοποίησης και το ζητούμενο φράγμα για το forward ή backward σφάλμα.

3.1.1 Πυρήνας, elaborator και τακτικές

Σε μια δήλωση Lean, το τμήμα μετά την άνω και κάτω τελεία εκφράζει την πρόταση που πρέπει να αποδειχθεί. Στη θεωρία τύπων η πρόταση αυτή αντιμετωπίζεται ως τύπος, ενώ κατά την κατασκευή της απόδειξης εμφανίζεται ως ο τρέχων στόχος. Για παράδειγμα:

```
theorem keepLeft (P Q : Prop) : P -> Q -> P := by
  intro hP hQ
  exact hP
```

Αρχικά ο στόχος είναι $P \to Q \to P$. Η τακτική `intro` εισάγει τις υποθέσεις `hP : P` και `hQ : Q`, αφήνοντας ως νέο στόχο το P. Η `exact hP` τον κλείνει. Οι τακτικές δεν αποτελούν διαφορετική λογική από τον πυρήνα: βοηθούν να κατασκευαστεί σταδιακά ένας όρος του δηλωμένου τύπου.

Ο χρήστης μπορεί επίσης να παραλείπει ορίσματα και πληροφορίες τύπων ή να χρησιμοποιεί συμβολισμούς και κλάσεις τύπων. Ο elaborator, δηλαδή ο μηχανισμός που επεξεργάζεται και αποσαφηνίζει τη γραφή του χρήστη, συμπληρώνει αυτά τα στοιχεία και επιλύει μεταμεταβλητές, δηλαδή προσωρινά κενά στον υπό κατασκευή όρο. Οι μηχανισμοί απλοποίησης και επίλυσης μπορούν να εκτελούν σύνθετη αναζήτηση, αλλά ο τελικός πλήρης όρος ελέγχεται ξανά από τον πυρήνα. Το ίδιο ισχύει για ένα εξωτερικό σύστημα επίλυσης μόνο όταν επιστρέφει όρο απόδειξης ή πιστοποιητικό που ελέγχεται στη Lean. Ένας πράκτορας όπως το Codex μπορεί να προτείνει κώδικα, αλλά δεν αποφασίζει ο ίδιος αν ο κώδικας αποτελεί απόδειξη.

3.2 Η Mathlib ως μαθηματική υποδομή

Η Mathlib είναι μια ενιαία, συνεργατικά αναπτυσσόμενη βιβλιοθήκη που οργανώνει μεγάλα τμήματα άλγεβρας και ανάλυσης σε επαναχρησιμοποιήσιμες ιεραρχίες (The mathlib Community, 2020). Βασικός σχεδιαστικός πυλώνας της NumStability είναι να χτίζει πάνω σε αυτή την υποδομή και να αποφεύγει τον επαναορισμό εννοιών που υπάρχουν ήδη. Έτσι, πεπερασμένα αθροίσματα, πίνακες, γραμμικοί μετασχηματισμοί και νόρμες παραμένουν συμβατά με το ευρύτερο οικοσύστημα της Lean.

Παράδειγμα είναι η νόρμα `complexVecLpNorm`. Σε ισοδύναμη γραφή χωρίς τα σύμβολα συντομογραφίας της Lean, ο ορισμός είναι:

```
abbrev CVec (n : Nat) := Fin n -> Complex

noncomputable def complexVecLpNorm {n : Nat}
    (p : ENNReal) (x : CVec n) : Real :=
  norm (WithLp.toLp p x)
```

Ο τύπος `CVec n` παριστά ένα μιγαδικό διάνυσμα με n συνιστώσες. Ο δεύτερος ορισμός δεν εισάγει ανεξάρτητη θεωρία νορμών: εφαρμόζει την υπάρχουσα νόρμα της Mathlib στο αντικείμενο `WithLp.toLp p x`. Έτσι μπορούν να χρησιμοποιηθούν απευθείας οι υποδομές της Mathlib για χώρους L^p, συνεχείς γραμμικούς μετασχηματισμούς και νόρμες τελεστών.[1]

[1] *Ορισμός της complexVecLpNorm πάνω στην υποδομή της Mathlib.* Τεκμήριο στο αποθετήριο NumStability: `67344c880: historical-library-snapshot/NumStability/Analysis/Norms.lean`.

Η επιστημονική συνεισφορά της NumStability βρίσκεται στο επόμενο στρώμα: στα μοντέλα τοπικού σφάλματος, στα λήμματα γ_n και στα θεωρήματα για forward, backward και mixed σφάλματα. Η συγκεκριμένη χρήση τύπων πινάκων και νορμών αναπτύσσεται στο Κεφάλαιο 5.

3.3 Η επιλογή της Lean

Η επιλογή της Lean υπαγορεύθηκε από τις ανάγκες του παρόντος έργου και δεν αποτελεί ισχυρισμό γενικής υπεροχής έναντι της Rocq/Coq ή άλλων βοηθών αποδείξεων. Η Rocq διαθέτει ώριμη εξειδικευμένη υποδομή κινητής υποδιαστολής, όπως οι Flocq και LAProof (Boldo & Melquiond, 2011; Kellison κ.ά., 2023).

Η χρήση της Lean στις συνεργατικές τυποποιήσεις σύγχρονης έρευνας που παρουσιάστηκαν στην Ενότητα 1.3 αποτελεί ένδειξη υιοθέτησης σε ερευνητική κλίμακα και αποτέλεσε σημαντικό κίνητρο. Η συστηματική σύγκριση των σχετικών υπαρχόντων οικοσυστημάτων κινητής υποδιαστολής ανήκει στο Κεφάλαιο 4.

3.4 Το Codex στην υποβοηθούμενη ανάπτυξη

Στην παρούσα εργασία, ο όρος Codex δηλώνει τον πράκτορα της OpenAI και όχι μία αμετάβλητη έκδοση μοντέλου. Η πολιτική κατά την ανάπτυξη ήταν να χρησιμοποιείται το ισχυρότερο διαθέσιμο μοντέλο με το υψηλότερο διαθέσιμο επίπεδο συλλογισμού. Στις δύο αξιολογήσεις του HighamBench τα μοντέλα καθορίστηκαν ρητά: το HighamBench-Proof χρησιμοποίησε το GPT-5.6 Sol με επίπεδο `xhigh` (OpenAI, 2026b), ενώ το HighamBench-E2E χρησιμοποίησε το GPT-6 Sol με επίπεδο `high` για τον πράκτορα και το GPT-6 Astra για τους κριτές πιστότητας (OpenAI, 2026c, 2026d).[2,3] Επειδή το ακριβές αναγνωριστικό μοντέλου δεν καταγράφηκε σε κάθε καταχώριση Git της βιβλιοθήκης, οι ειδικές ονομασίες χρησιμοποιούνται μόνο όπου υπάρχει τέτοιο τεκμήριο· στο υπόλοιπο κείμενο χρησιμοποιείται ο γενικότερος όρος Codex.

Το Codex ήταν το κύριο εργαλείο ανάπτυξης: σύμφωνα με τα τοπικά αρχεία των νημάτων του, χρησιμοποίησε αρχικά το GPT-5.5 και, από τα τέλη Ιουλίου 2026, το GPT-5.6 Sol (OpenAI, 2026b). Το ιστορικό του αποθετηρίου καταγράφει επίσης μικρότερη συμμετοχή του Claude: αρχικά με Sonnet και Opus 4.6, αργότερα με Opus 4.8 και Fable 5, και στις νεότερες καταγεγραμμένες συνεισφορές με Opus 5 (Anthropic, 2026). Η συμμετοχή αυτή αφορά τόσο πρώιμες τυποποιήσεις όσο και επιλεγμένους μεταγενέστερους ελέγχους, ταξινομήσεις και εργασίες αναδιοργάνωσης. Στους ελέγχους της ευρύτερης βιβλιοθήκης αξιοποιήθηκαν αποτελέσματα και των δύο πρακτόρων, ενώ οι έλεγχοι πιστότητας των εργασιών του HighamBench εκτελέστηκαν με μοντέλα της OpenAI μέσω του Codex. Η αναφορά του πράκτορα δηλώνει την προέλευση ενός τεκμηρίου και όχι ανεξάρτητη ανθρώπινη πιστοποίησή του.

[2] *Μοντέλο και επίπεδο συλλογισμού των καταγεγραμμένων προσπαθειών της αξιολόγησης των 30 εργασιών.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:paper_bencmark/highambench/reports/full-report-scope30/report.tex`.

[3] *Μοντέλα του πράκτορα και των κριτών της αξιολόγησης των δέκα εργασιών.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/protocol/EXECUTION.md`.

3.4.1 Το skill ως επαναχρησιμοποιήσιμη διαδικασία

Στην ορολογία του Codex, ένα *skill* είναι ένας κατάλογος με υποχρεωτικό αρχείο `SKILL.md`, το οποίο περιέχει μεταδεδομένα ενεργοποίησης και οδηγίες για τον πράκτορα. Μπορεί επίσης να περιλαμβάνει βοηθητικά προγράμματα, αρχεία αναφοράς και άλλα επαναχρησιμοποιήσιμα μέσα (OpenAI, 2026a). Με τη δομή αυτή μπορεί να επαναχρησιμοποιείται μια διαδικασία ως προδιαγραφή που ερμηνεύει ο πράκτορας, αντί να αποτυπώνεται εξ ολοκλήρου σε πρόγραμμα με πλήρως προκαθορισμένη ροή. Δεν αντικαθιστά τον κώδικα: τα ντετερμινιστικά βήματα, όπως η μεταγλώττιση και οι έλεγχοι αποδοχής, παραμένουν προγράμματα. Το skill που καθοδήγησε την κατασκευή της NumStability και τα επιμέρους στάδιά του αναπτύσσονται στο Κεφάλαιο 6.

3.5 Κατανομή εμπιστοσύνης μεταξύ ανθρώπου, TN και πυρήνα Lean

Η παρουσία βοηθού αποδείξεων αλλάζει το είδος της εμπιστοσύνης που απαιτείται· δεν την εξαλείφει. Η κρίσιμη διάκριση βρίσκεται ανάμεσα στη μετάφραση και στην απόδειξη. Κατά τη μετάφραση από μια φυσική μαθηματική διατύπωση σε Lean, το μοντέλο (ή και ένας άνθρωπος) μπορεί να παραγάγει μια τυπική δήλωση που μεταγλωττίζεται, αλλά είναι ασθενέστερη, ισχυρότερη ή διαφορετική από την πηγή. Ο έλεγχος τύπων από μόνος του δεν μετρά αυτή τη σημασιολογική πιστότητα (Meek κ.ά., 2026; Poiroux κ.ά., 2025).

Από τη στιγμή που οριστικοποιηθούν η δήλωση Lean, οι ορισμοί, οι εισαγωγές και η πολιτική επιτρεπόμενων αξιωμάτων, η αναζήτηση της απόδειξης μπορεί να ανατεθεί στον πράκτορα TN χωρίς να θεωρείται έμπιστος ο εσωτερικός συλλογισμός του. Η έξοδός του είναι ένας υποψήφιος όρος απόδειξης. Ο όρος γίνεται δεκτός μόνο αν περάσει τον έλεγχο του πυρήνα και τους μηχανικούς ελέγχους για κενά απόδειξης και μη επιτρεπόμενες παραδοχές. Ο άνθρωπος ή ο σημασιολογικός κριτής ελέγχει τη σχέση της πηγής με την τυπική δήλωση· ο πράκτορας αναζητά απόδειξη· ο πυρήνας ελέγχει το τελικό τυπικό αντικείμενο (Σχήμα 3.1).

Το σχήμα αποτυπώνει ένα γενικό κανονιστικό μοντέλο εκτέλεσης: ο έλεγχος πιστότητας αποτελεί χωριστό ρόλο από τη μετάφραση και την αναζήτηση απόδειξης και μπορεί να ανατεθεί είτε σε άνθρωπο είτε σε χωριστό LLM σημασιολογικό κριτή. Στην παρούσα εργασία εφαρμόστηκε η δεύτερη παραλλαγή. Οι κρίσεις πιστότητας στις δύο αξιολογήσεις του HighamBench είναι επομένως κρίσεις πρακτόρων και όχι ανθρώπινη πιστοποίηση (Κεφάλαιο 7). Ο ανθρώπινος ρόλος αφορούσε την επιλογή του συνόλου πηγών και τον σχεδιασμό του πρωτοκόλλου και των κριτηρίων.[4]

Η εφαρμογή των μηχανικών ελέγχων και η πλήρης ροή της συγκριτικής αξιολόγησης παρουσιάζονται αντίστοιχα στα Κεφάλαια 6 και 7.

[4] *Πρωτόκολλο ελέγχου πιστότητας*. Τεκμήριο στο αποθετήριο NumStability: `8db102c12:faithfulness/METHODOLOGY.md`.



Σχήμα 3.1: Από τη μαθηματική πηγή στην ελεγμένη απόδειξη: διακριτοί ρόλοι μετάφρασης, ελέγχου πιστότητας και τυπικής επαλήθευσης. Στην παρούσα εργασία, την πιστότητα κρίνουν χωριστές, απομονωμένες εκτελέσεις του Codex. Ο άνθρωπος επιλέγει τις πηγές και σχεδιάζει τα κριτήρια, χωρίς να επαναβαθμολογεί τις επιμέρους κρίσεις.

3.6 Συμπέρασμα

Η Lean 4 διαχωρίζει την αναζήτηση απόδειξης από τον έλεγχο του τελικού όρου, ενώ η Mathlib παρέχει την κοινή μαθηματική υποδομή της NumStability. Το Codex και τα skills οργανώνουν την αναζήτηση και την επαναχρησιμοποίηση διαδικασιών, χωρίς να μετακινούν το τελικό όριο αποδοχής έξω από τον πυρήνα και τους δηλωμένους ελέγχους. Το επόμενο κεφάλαιο τοποθετεί τη βιβλιοθήκη και το HighamBench μέσα στη σχετική βιβλιογραφία.

Κεφάλαιο 4

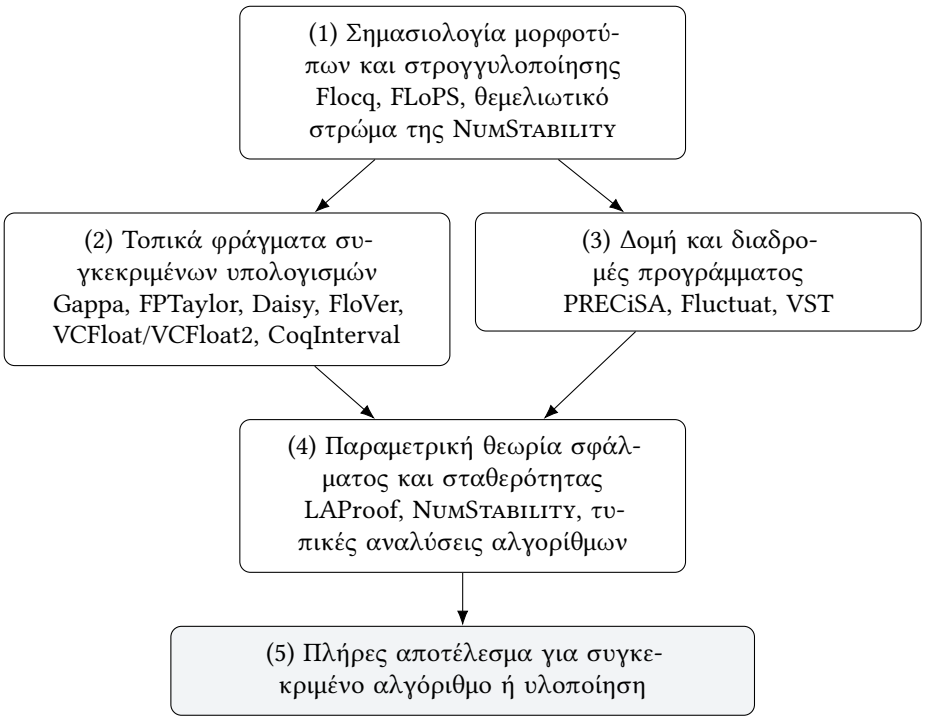
Σχετική βιβλιογραφία

4.1 Πεδίο και επίπεδα επαλήθευσης

Η τυπική μελέτη της αριθμητικής κινητής υποδιαστολής περιλαμβάνει έργα με διαφορετική είσοδο, έξοδο και κλίμακα εγγύησης. Στο θεμελιώδες επίπεδο, βιβλιοθήκες ορίζουν τους αναπαραστάσιμους αριθμούς ενός μορφοτύπου, δηλαδή τη βάση, την ακρίβεια του σημαντικού (*significand*, συχνά και *mantissa*), την περιοχή εκθετών, τις ειδικές τιμές και τους κανόνες στρογγυλοποίησης. Σε επόμενα επίπεδα, αναλυτές υπολογίζουν φράγματα για συγκεκριμένες εκφράσεις ή προγράμματα, ελεγκτές πιστοποιούν προτεινόμενα φράγματα και βιβλιοθήκες θεωρημάτων οργανώνουν παραμετρικά αποτελέσματα για ολόκληρες οικογένειες αλγορίθμων. Ένα αριθμητικό φράγμα για σταθερή έκφραση και ένα θεώρημα backward stability για αυθαίρετη διάσταση απαντούν επομένως σε διαφορετικά ερωτήματα.

Το Σχήμα 4.1 συνοψίζει τον τρόπο με τον οποίο αυτά τα επίπεδα μπορούν να συνδυαστούν. Δεν υπονοεί ότι ένα εργαλείο ενός ανώτερου στρώματος περιέχει ή αντικαθιστά τις δυνατότητες των εργαλείων που εμφανίζονται χαμηλότερα. Ένα πλήρες αποτέλεσμα για κώδικα μπορεί να χρειάζεται συγχρόνως ακριβή σημασιολογία του μορφοτύπου, τοπικά φράγματα, απόδειξη της δομής του προγράμματος και ένα γενικό θεώρημα διάδοσης του σφάλματος.

Χαρακτηριστικό πλήρες παράδειγμα αυτής της διαστρωμάτωσης είναι το VerifiedLeapfrog, του οποίου η απόδειξη αναπτύσσεται στη Rocq/Coq και αφορά υλοποίηση σε C της μεθόδου leapfrog για ένα πρόβλημα αρχικών τιμών, σε αριθμητική binary32. Η απόδειξη διατρέχει όλα τα στρώματα του σχήματος. Στο στρώμα 1, το μοντέλο IEEE 754 της Flocq ορίζει τη σημασιολογία των πράξεων. Στο στρώμα 2, το VCFloat, μαζί με το CoqInterval, φράσσει το τοπικό σφάλμα στρογγυλοποίησης ενός βήματος. Στο στρώμα 3 χρησιμοποιείται το VST (Verified Software Toolchain), μια λογική προγραμμάτων για τη C της οποίας η ορθότητα έχει αποδειχθεί στη Rocq ως προς την τυπική σημασιολογία της γλώσσας· με αυτό αποδεικνύεται ότι το πρόγραμμα C συμπεριφέρεται όπως ένα συναρτησιακό μοντέλο του αλγορίθμου. Στο στρώμα 4, η ανάλυση του ίδιου του έργου φράσσει το σφάλμα διακριτοποίησης με τη βιβλιοθήκη πραγματικής ανάλυσης Coquelicot και συνθέτει τα τοπικά σφάλματα όλων των βημάτων. Τέλος, το θεώρημα του στρώματος 5 φράσσει το συνολικό σφάλμα της υλοποίησης (Kellison & Appel, 2022). Το παράδειγμα δεν σημαίνει ότι



Σχήμα 4.1: Συμπληρωματικά επίπεδα μιας τυπικής ανάλυσης αριθμητικής κινητής υποδιαστολής. Τα ονόματα δηλώνουν χαρακτηριστικά παραδείγματα και όχι αποκλειστική κατάταξη· οι αριθμοί των στρωμάτων χρησιμοποιούνται στο κείμενο του κεφαλαίου.

κάθε έργο χρειάζεται όλα τα στρώματα, αλλά δείχνει ότι το σχήμα αντιστοιχεί σε πραγματική αρχιτεκτονική απόδειξης.

4.2 Θεμελιώσεις αριθμητικής κινητής υποδιαστολής

Η Flocq είναι βιβλιοθήκη της Rocq που παρέχει τυπική σημασιολογία για μορφοτύπους διαφορετικής βάσης και ακρίβειας, τρόπους στρογγυλοποίησης και σχέσεις μεταξύ των πράξεων κινητής υποδιαστολής και των πραγματικών αριθμών (Boldo & Melquiond, 2011). Οι αποδείξεις της ελέγχονται από τον πυρήνα της Rocq και χρησιμοποιούνται ως θεμέλιο από μεταγενέστερα έργα.

Στη Lean, η πρόσφατη βιβλιοθήκη FLoPS τυποποιεί προτεινόμενες αναπαραστάσεις χαμηλής ακρίβειας του προτύπου IEEE P3109 και συνδέει τη δυαδική κωδικοποίηση με την αφηρημένη σημασιολογία τους (Chang κ.ά., 2026). Η FLoPS είναι συνεπώς σημαντικό τεκμήριο ότι αναπτύσσεται στη Lean εξειδικευμένη υποδομή συγκεκριμένων μορφοτύπων, χωρίς να έχει τον ίδιο στόχο με μια βιβλιοθήκη γενικής ανάλυσης σταθερότητας.

Η Flocq, η FLoPS και η NumStability μοιράζονται την ιδέα της επαναχρησιμοποιήσιμης, μηχανικά ελεγμένης υποδομής, αλλά εξυπηρετούν διαφορετικούς στόχους. Η Flocq σχεδιάστηκε για να ενοποιήσει προηγούμενες τυποποιήσεις της αριθμητικής κινητής υποδιαστολής στη Rocq, καθεμία από τις οποίες ευνοούσε μία χρήση, όπως την επαλήθευση προγραμμάτων, τις ιδιότητες υψηλού επιπέδου ή την αυτοματοποίηση. Αποτελεί έτσι κοινό και γενικό θεμέλιο για αποδείξεις αλγορίθμων κινητής υποδιαστολής (Boldo & Melquiond, 2011). Η FLoPS επιδιώκει να αποτελέσει

μαθηματικά ακριβή αναφορά για το πρότυπο P3109, το οποίο προορίζεται για υλικό και λογισμικό μηχανικής μάθησης. Κατά την ανάπτυξή της εντοπίστηκαν σφάλματα στο προσχέδιο του προτύπου, τα οποία αναφέρθηκαν στην ομάδα εργασίας του, ενώ η επαληθευμένη διαδικασία στρογγυλοποίησης και κωδικοποίησης προτείνεται ως αναφορά για τον έλεγχο ισοδυναμίας σχεδίων υλικού (Chang κ.ά., 2026).

Και οι δύο βιβλιοθήκες περιέχουν αποτελέσματα για αλγορίθμους και όχι μόνο για μεμονωμένες πράξεις. Η Flocq περιλαμβάνει αποδείξεις για τους FastTwoSum και TwoSum, τη διάσπαση του Veltkamp και τον αλγόριθμο του Dekker για το ακριβές γινόμενο (Boldo & Melquiond, 2026), ενώ η FLoPS μελετά τους FastTwoSum και ExtractScalar σε μορφοτύπους P3109 (Chang κ.ά., 2026). Πρόκειται κυρίως για σύντομους αλγορίθμους με σταθερό πλήθος πράξεων, των οποίων η ορθότητα εξαρτάται από ιδιότητες του συγκεκριμένου μορφοτύπου, όπως η βάση, η ακρίβεια και η απουσία υπερχείλισης. Τα θεωρήματα έχουν συνήθως τη μορφή ακριβών ταυτοτήτων: για παράδειγμα, το σφάλμα στρογγυλοποίησης μιας πρόσθεσης αναπαρίσταται ακριβώς στον ίδιο μορφότυπο και υπολογίζεται από τον αλγόριθμο.

Η NumStability στοχεύει σε διαφορετικό είδος αποτελέσματος. Μελετά κυρίως αλγορίθμους για είσοδο αυθαίρετου μεγέθους, όπως η άθροιση n όρων και οι παραγοντοποιήσεις πινάκων $n \times n$, στο αφηρημένο μοντέλο του Higham, όπου κάθε πράξη ικανοποιεί $|\delta| \leq u$ (Ενότητα 5.2). Τα θεωρήματά της είναι φράγματα forward, backward και mixed σφάλματος, σε componentwise ή normwise μορφή, εκφρασμένα μέσω των ποσοτήτων γ_n, νορμών και condition numbers. Ένα τέτοιο θεώρημα ισχύει για κάθε αριθμητική που ικανοποιεί το πρότυπο, αλλά δεν περιγράφει τι συμβαίνει όταν οι υποθέσεις του παραβιάζονται, για παράδειγμα σε υπερχείλιση ή υποχείλιση. Τα όρια δεν είναι απόλυτα: η NumStability περιέχει και ακριβή αποτελέσματα για πεπερασμένους μορφοτύπους, όπως τις συνθήκες υπό τις οποίες ο διορθωτικός όρος του FastTwoSum είναι ακριβής σε πεπερασμένο δυαδικό μορφότυπο (N. J. Higham, 2002, εξ. (4.7), σ. 84).[1]

Οι βιβλιοθήκες αυτές δεν περιέχουν η μία την άλλη. Η Flocq ανήκει σε άλλον βοηθό αποδείξεων, ενώ η FLoPS, αν και είναι επίσης γραμμένη στη Lean, δεν αποτελεί εξάρτηση της NumStability. Επιπλέον, το κύριο αφηρημένο μοντέλο της NumStability δεν έχει γενικά παραχθεί από κάθε συγκεκριμένο πεπερασμένο μορφότυπο. Πρόκειται για συμπληρωματικά στρώματα: ένα θεώρημα εξειδίκευσης, που αποδεικνύει ότι ένας μορφότυπος ικανοποιεί το πρότυπο του Higham σε συγκεκριμένη περιοχή τιμών, επιτρέπει να εφαρμοστούν τα γενικά θεωρήματα της NumStability στους υπολογισμούς που παραμένουν σε αυτή την περιοχή.

4.3 Αυτόματη ανάλυση συγκεκριμένων υπολογισμών

Τα εργαλεία της ενότητας παρουσιάζονται για να χαρτογραφηθεί το γειτονικό πεδίο, όχι επειδή επιτελούν την ίδια εργασία με τη NumStability. Όπως φαίνεται και στο Σχήμα 4.1, ανήκουν σε άλλο στρώμα της τυπικής ανάλυσης, εκείνο των τοπικών φραγμάτων (στρώμα 2). Συνήθως δέχονται μια συγκεκριμένη έκφραση ή ένα συγκεκριμένο πρόγραμμα μαζί με φραγμένο πεδίο

[1] *Ακρίβεια του διορθωτικού όρου του FastTwoSum σε πεπερασμένο μορφότυπο με στρογγυλοποίηση στον πλησιέστερο άρτιο.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/Summation/Compensated/FastTwoSum.lean`.

εισόδων και παράγουν αριθμητικό φράγμα για αυτόν τον υπολογισμό.

4.3.1 Το εργαλείο Gappa

Το Gappa είναι εξειδικευμένο σύστημα απόδειξης αριθμητικών προτάσεων με ακριβείς και στρογγυλοποιημένους τελεστές (Daumas & Melquiond, 2010). Από πεδία εισόδων και έναν ρητά δηλωμένο υπολογισμό μπορεί να αποδείξει φράγματα τιμών, απόλυτα ή σχετικά σφάλματα, ακριβείς πράξεις και ιδιότητες αναπαραστασιμότητας. Χρησιμοποιεί αριθμητική διαστημάτων με κατευθυνόμενη στρογγυλοποίηση, αναδιατυπώσεις και, όταν χρειάζεται, υποδιαίρεση του πεδίου. Μπορεί επίσης να παραγάγει πιστοποιητικό που ελέγχεται με συνοδευτικές βιβλιοθήκες της Rocq ή του HOL Light (Daumas & Melquiond, 2010, §§ 3.2–3.3).

4.3.2 Ο αναλυτής FPTaylor και το πλαίσιο Daisy

Ο FPTaylor παράγει αυστηρά φράγματα forward σφάλματος για σταθερές αριθμητικές εκφράσεις σε φραγμένο πεδίο εισόδων. Χρησιμοποιεί συμβολική ανάπτυξη ως προς τα τοπικά σφάλματα και καθολική βελτιστοποίηση, διατηρώντας εξαρτήσεις που χάνονται σε μια απλή διαστηματική διάδοση (Solovyev κ.ά., 2018). Ο βασικός μηχανισμός αφορά υπολογιστικό γράφο χωρίς γενικούς βρόχους και διακλαδώσεις, ενώ παλαιότερη έκδοση μπορούσε να παραγάγει αποδείξεις για έλεγχο στο HOL Light.

Το Daisy συνδυάζει περισσότερες τεχνικές ανάλυσης και μετασχηματισμού, μεταξύ των οποίων, όπως και στο Gappa, η αριθμητική διαστημάτων, καθώς και η επίλυση λογικών περιορισμών, η επιλογή ακρίβειας και η αναδιατύπωση προγραμμάτων (Darulova κ.ά., 2018). Νεότερες επεκτάσεις καλύπτουν περιορισμένες διαδικασίες και φραγμένες συλλογές, αλλά αυτό δεν ισοδυναμεί με παραμετρικό θεώρημα αυθαίρετου μεγέθους. Η συνήθης έξοδός του δεν ελέγχεται από βοηθό αποδείξεων.

4.4 Ανάλυση προγραμμάτων και πιστοποιημένα αποτελέσματα

Τα εργαλεία της προηγούμενης ενότητας δέχονται κυρίως μια σταθερή έκφραση και υπολογίζουν ένα φράγμα για αυτήν. Τα εργαλεία της παρούσας ενότητας επεκτείνουν αυτή τη δυνατότητα προς δύο διαφορετικές κατευθύνσεις. Το Fluctuat και το PRECiSA αναλύουν ολόκληρα προγράμματα, με βρόχους, διακλαδώσεις ή κλήσεις συναρτήσεων, και ανήκουν στο στρώμα 3 του Σχήματος 4.1. Ο FloVer και το VCFloat2 αφορούν, όπως τα προηγούμενα εργαλεία, φράγματα συγκεκριμένων υπολογισμών (στρώμα 2), αλλά το αποτέλεσμά τους ελέγχεται από τον πυρήνα ενός βοηθού αποδείξεων. Η διάκριση δεν είναι απόλυτη, αφού και το Gappa μπορεί να παραγάγει πιστοποιητικό για βοηθό αποδείξεων.

4.4.1 Fluctuat

Το Fluctuat εφαρμόζει αφηρημένη ερμηνεία σε προγράμματα C και Ada και υπολογίζει περιοχές πραγματικών και πεπερασμένης ακρίβειας τιμών, σφάλματα και την προέλευση της απώλειας ακρίβειας (Goubault & Putot, 2013). Οι αφηρημένοι τομείς, οι συνενώσεις και οι τεχνικές διεύρυνσης

επιτρέπουν την ανάλυση βρόχων, πινάκων και διακλαδώσεων σε κλίμακα πηγαίου κώδικα. Η συνήθης εκτέλεσή του, ωστόσο, δεν παράγει ανεξάρτητο πιστοποιητικό Rocq, PVS ή Lean για κάθε αποτέλεσμα. Η ισχύς του βρίσκεται στην ορθή και διαγνωστική ανάλυση προγραμμάτων, όχι στην παραγωγή μόνιμης βιβλιοθήκης θεωρημάτων.

4.4.2 Ο ελεγκτής FloVer και το VCFloat2

Ο FloVer διαχωρίζει την αναζήτηση φράγματος από την πιστοποίησή του. Δέχεται ένα πιστοποιητικό με περιοχές τιμών και τοπικά σφάλματα και ελέγχει στη Rocq ή στο HOL4 ότι η διάδοση του σφάλματος είναι ορθή και ότι δεν προκύπτει διαίρεση με μηδέν ή υπερχείλιση στο υποστηριζόμενο τμήμα (Becker κ.ά., 2018). Δεν ανακαλύπτει ο ίδιος το καλύτερο φράγμα. Η σύνδεσή του με το Daisy αποκάλυψε σφάλμα στη διαχείριση της διαίρεσης, αποδεικνύοντας πρακτικά την αξία ενός μικρού, επαληθευμένου ελεγκτή.

Το VCFloat2, αντίθετα, εκτελεί την ανάλυση μέσα στη Rocq: αναπαριστά τον υπολογισμό, παράγει τοπικές υποχρεώσεις και κατασκευάζει όρο απόδειξης που ελέγχεται από τον πυρήνα (Appel & Kellison, 2024). Μπορεί να συνδυαστεί με το VST ώστε μια αριθμητική έκφραση να συνδεθεί με πρόγραμμα C. Παραμένει όμως αναγκαία η χωριστή μαθηματική απόδειξη που μετατρέπει τα τοπικά φράγματα σε καθολικό forward ή backward αποτέλεσμα για τον αλγόριθμο.

4.4.3 PRECiSA 4.0

Το PRECiSA 4.0 είναι ανοικτού κώδικα στατικός αναλυτής της NASA για το απόλυτο σφάλμα χειρότερης περίπτωσης σε αριθμητικά προγράμματα (Titolo κ.ά., 2025). Δέχεται προγράμματα σε υποσύνολο της PVS και, μέσω χωριστής διεπαφής, σε FPCore. Από ένα πρόγραμμα και φραγμένες περιοχές εισόδων κατασκευάζει μία διαδρομή για την ιδεατή πραγματική εκτέλεση και μία για την εκτέλεση κινητής υποδιαστολής. Σχηματίζει κατόπιν εκφράσεις σφάλματος, τις μεγιστοποιεί αυστηρά με το Kodiak και παράγει πιστοποιητικά που ελέγχονται στην PVS.

Όταν οι δύο εκτελέσεις επιλέγουν διαφορετικό κλάδο μιας συνθήκης, το PRECiSA συνυπολογίζει και τη διαφορά μεταξύ των αποτελεσμάτων των κλάδων. Υποστηρίζει, με ρητούς περιορισμούς, μη αναδρομικές κλήσεις, φραγμένα σχήματα επανάληψης και συλλογές δεδομένων, ενώ η ανάλυση περιορίζεται κυρίως στη στρογγυλοποίηση στον πλησιέστερο αριθμό.

Ο Πίνακας 4.1 κλείνει την ενότητα με τη σύγκριση των εργαλείων όπως ακριβώς την παρουσίασαν οι συγγραφείς του PRECiSA.

Η NumStability δεν προστίθεται ως ακόμη μία στήλη, επειδή δεν είναι αυτόματος αναλυτής του ίδιου τύπου. Το PRECiSA υπερέχει στη σχεδόν αυτόματη ανάλυση ενός συγκεκριμένου προγράμματος, των κλάδων του και ενός δεδομένου πεδίου εισόδων. Η NumStability στοχεύει σε άλλο αποτέλεσμα: θεωρήματα αυθαίρετης διάστασης, υπαρξιακές backward ερμηνείες, δομές πινάκων, νόρμες και κατάσταση του προβλήματος (conditioning). Ένα φράγμα του τύπου που υπολογίζει το PRECiSA, αφού διατυπωθεί και αποδειχθεί στη Lean, μπορεί να χρησιμοποιηθεί ως προκείμενη ενός γενικού θεωρήματος της NumStability.

	PRECiSA	FPTaylor	Daisy	VCFloat	Fluctuat	Gappa
proof certificates	✓	✓	✗	✓	✗	✓
conditionals	✓	✗	✗	✗	✓	✗
instability detection	✓	✗	✗	✗	✓	✗
instability analysis	✓	✗	✗	✗	✗	✗
function calls	✓	✗	✗	✗	✓	✗
bounded loops	✓	✗	✗	✗	✓	✗
widening	✓	✗	✗	✗	✓	✗
data collections	✓	✗	✓	✗	✓	✗
rounding modes	✗	✓	✗	✗	✗	✗
fixed-point arith.	✗	✗	✓	✗	✗	✓

Πίνακας 4.1: Αναπαραγωγή του σώματος του Πίνακα 2 των Titolo κ.ά. (2025), ο οποίος συγκρίνει δυνατότητες εργαλείων ανάλυσης σφάλματος στρογγυλοποίησης χειρότερης περίπτωσης. Η δημοσίευση διατίθεται με άδεια CC BY 4.0.

4.5 Επαναχρησιμοποιήσιμη θεωρία και ολόκληροι αλγόριθμοι

4.5.1 Roux

Ο Roux (2016) τυποποιεί στη Rocq φράγματα forward σφάλματος για αθροίσεις και εσωτερικά γινόμενα και τα συνθέτει σε μηχανικά ελεγμένο επαρκές κριτήριο για positive definiteness μέσω παραγοντοποίησης Cholesky. Η επιτυχία του ελέγχου πιστοποιεί το αποτέλεσμα, ενώ η αποτυχία του είναι απλώς μη αποφασιστική. Το έργο αποτελεί καθαρό παράδειγμα μετάβασης από τοπικά λήμματα σε δομή πινάκων και αλγοριθμικό συμπέρασμα.

4.5.2 LAProof

Η LAProof είναι η πλησιέστερη δημοσιευμένη βιβλιοθήκη προς τον στόχο της NumStability. Παρέχει στη Rocq επαναχρησιμοποιήσιμα forward, backward και mixed αποτελέσματα για άθροιση, εσωτερικό γινόμενο, γινόμενο πίνακα–διανύσματος και γινόμενο πινάκων, με μοντέλο που λαμβάνει υπόψη την υποχείλιση (Kellison κ.ά., 2023). Ως ολοκληρωμένο παράδειγμα, ένα γενικό θεώρημα για γινόμενο πίνακα–διανύσματος συνδέεται μέσω VCFloat2 και VST με υλοποίηση C αραιού γινομένου σε μορφή CSR. Αυτό τεκμηριώνει την αξία του επαναχρησιμοποιήσιμου ενδιάμεσου στρώματος.

4.5.3 Βιβλιοθήκες και πράκτορες TN

Οι δημοσιεύσεις του Roux και της LAProof δεν περιλαμβάνουν πειραματική σύγκριση της προσπάθειας απόδειξης με και χωρίς πρόσβαση στη βιβλιοθήκη (Kellison κ.ά., 2023; Roux, 2016). Αυτό δεν αποτελεί αδυναμία τους: η επαναχρησιμοποίηση είναι βασικός στόχος της ανάπτυξής τους, ενώ η αξιολόγηση αυτόνομων πρακτόρων TN δεν ήταν το αντικείμενό τους. Και η νεότερη FLoPS δεν αναφέρει κατασκευή από τέτοιον πράκτορα (Chang κ.ά., 2026).

Η παρούσα εργασία προτείνει να μετρηθεί αυτό το όφελος με πράκτορες που επαναλαμβάνουν τις ίδιες εργασίες υπό σταθερούς περιορισμούς. Το ιδιαίτερο ερώτημα είναι αν ο πράκτορας

βρίσκει και αξιοποιεί την κατάλληλη υποδομή, ώστε να τυποποιεί και να αποδεικνύει νέα αποτελέσματα με μικρότερο κόστος.

4.6 Προηγούμενες αξιολογήσεις και μεθοδολογία

Το FPBench και η γλώσσα FPCore παρέχουν κοινή υποδομή συγκριτικής αξιολόγησης εργαλείων που αναλύουν ή μετασχηματίζουν συγκεκριμένες αριθμητικές εκφράσεις, όπως εργαλεία της ίδιας γενικής κατηγορίας με τα Gappa, FPTaylor και Daisy (Damouche κ.ά., 2017). Περιγράφουν κοινή αναπαράσταση του υπολογισμού, των πεδίων εισόδων, της ακρίβειας και της προέλευσης κάθε δοκιμασίας, αλλά δεν αποτελούν σύνολο αξιολόγησης για την απόδειξη παραμετρικών θεωρημάτων αυθαίρετης διάστασης. Για την παρούσα αξιολόγηση λειτουργούν ως μεθοδολογικό προηγούμενο ως προς την αμετάβλητη προδιαγραφή και την είσοδο των δοκιμασιών.

Τα σύνολα αξιολόγησης τυπικών αποδείξεων διαφέρουν ως προς το τι δίνεται ως δεδομένο, τι ζητείται από το σύστημα και τι μετράται.

Το miniF2F περιέχει σταθερές τυπικές δηλώσεις 488 προβλημάτων από τους διαγωνισμούς AMC, AIME και IMO και από μαθήματα λυκείου και πανεπιστημίου, σε Metamath, Lean και εν μέρει σε Isabelle και HOL Light (Zheng κ.ά., 2022). Το PutnamBench περιέχει 1 692 σταθερές τυποποιήσεις 640 προβλημάτων του διαγωνισμού Putnam σε Lean 4, Isabelle και, για μεγάλο υποσύνολο, Coq, μαζί με την αγγλική διατύπωσή τους (Tsoukalas κ.ά., 2024). Και στα δύο ζητείται η απόδειξη της δοσμένης δήλωσης και μετράται η επιτυχία εντός n προσπαθειών (pass@n). Από αυτά το HighamBench-Proof υιοθετεί τη σταθερή τυπική δήλωση κάθε εργασίας, ενώ και οι δύο αξιολογήσεις υιοθετούν τον έλεγχο της επιτυχίας από τον πυρήνα.

Το ProofNet συνδυάζει φυσική γλώσσα και Lean: καθένα από τα 371 προπτυχιακά θεωρήματά του συνοδεύεται από δήλωση Lean 3, διατύπωση και απόδειξη σε φυσική γλώσσα. Εξετάζει τόσο την αυτόματη τυποποίηση της διατύπωσης όσο και την τυπική απόδειξη (Azerbayev κ.ά., 2023). Είναι το πλησιέστερο προηγούμενο για το HighamBench-E2E ως προς το σημείο εκκίνησης, αφού και αυτό ξεκινά από τη διατύπωση σε φυσική γλώσσα.

Το LeanDojo εξάγει από τη Mathlib 98 734 θεωρήματα και αποδείξεις, με σημειωμένες τις προκείμενες κάθε απόδειξης, κυρίως σε Lean 3 και με υποστήριξη και για Lean 4. Ζητείται η απόδειξη με ανάκτηση προκειμένων από τη βιβλιοθήκη και μετράται το Pass@1, ενώ ένας ειδικός διαχωρισμός των δεδομένων απαιτεί προκείμενες που δεν χρησιμοποιήθηκαν στην εκπαίδευση (Yang κ.ά., 2023). Το έργο δείχνει ότι η εύρεση των κατάλληλων προκειμένων μέσα σε μια μεγάλη βιβλιοθήκη αποτελεί από μόνη της δυσκολία. Γι' αυτό στο HighamBench η πρόσβαση στη NumStability δεν θεωρείται ισοδύναμη με τη χρήση της και ούτε θεωρείται εκ των προτέρων πλεονεκτική.

Το CAM-Bench περιέχει 1 000 σταθερές δηλώσεις Lean 4 από τη βελτιστοποίηση, την αριθμητική γραμμική άλγεβρα και την αριθμητική ανάλυση, μαζί με την αρχική άσκηση, το ανακτημένο πλαίσιο της πηγής και άτυπη διατύπωση. Ζητείται η απόδειξη της δοσμένης δήλωσης· μετράται το pass@32 και, για πράκτορες, το πλήθος των εργασιών που λύνονται εντός οκτώ ωρών (Long κ.ά., 2026). Όπως το HighamBench-Proof, ζητά την απόδειξη σταθερής δήλωσης και συνδέει

κάθε δήλωση με την πηγή από την οποία προέρχεται.

Το TheoremBench περιέχει 1 142 εργασίες Lean 4 από 83 κλασικά θεωρήματα, σε δύο μορφές: χωρίς και με ρητές προκείμενες από προηγούμενα αποτελέσματα. Μετρά, μεταξύ άλλων, το pass@k, την κάλυψη των θεωρημάτων και το μήκος των αποδείξεων, και δείχνει ότι οι ρητές προκείμενες βελτιώνουν ουσιωδώς την επίδοση (Pham κ.ά., 2026). Η LAMP αξιολογείται σε 90 θεωρήματα Lean 4 της συνδυαστικής λέξεων με σύστημα πολλών πρακτόρων, που έχει πρόσβαση μέσω MCP σε οντολογία του πεδίου. Η αφαίρεση της αρχιτεκτονικής που στηρίζεται σε αυτά τα εργαλεία μειώνει την επιτυχία κατά περίπου 12 ποσοστιαίες μονάδες (Srinivasan R & Patawar, 2026). Και τα δύο συγκρίνουν την ίδια εργασία με και χωρίς πρόσθετη γνώση· το HighamBench εφαρμόζει την ίδια λογική με μεταβλητή την πρόσβαση στη βιβλιοθήκη.

Όλα τα παραπάνω ζητούν την απόδειξη μιας έτοιμης τυπικής δήλωσης, και μόνο το ProofNet εξετάζει επιπλέον την τυποποίηση· στα υπόλοιπα, η φυσική γλώσσα, όπου υπάρχει, απλώς συνοδεύει τη δήλωση. Η κύρια μέτρηση είναι η επιτυχία εντός ορισμένου αριθμού προσπαθειών ή ορισμένου χρόνου και όχι η ταχύτητα. Οι εργασίες αυτές τεκμηριώνουν επίσης ότι το διαθέσιμο πλαίσιο επηρεάζει την απόδειξη.

Το HighamBench διαφέρει ως προς τη σύγκριση με και χωρίς ολόκληρη τη βιβλιοθήκη του πεδίου και την έμφαση στο κόστος: χρόνο, tokens και μέγεθος κώδικα. Το HighamBench-Proof απομονώνει την απόδειξη δοσμένης δήλωσης, ενώ το HighamBench-E2E εξετάζει επιπλέον τη διατύπωση από τη μαθηματική πηγή. Η μεθοδολογία εκτέλεσης περιγράφεται στο Κεφάλαιο 7.

Δύο πρόσφατα έργα είναι ιδιαίτερα συναφή με την κλίμακα και τη διαδικασία της παρούσας εργασίας. Το CSLib αποτελεί πρώιμη πρωτοβουλία για την κατασκευή μιας συνεκτικής βιβλιοθήκης επιστήμης υπολογιστών στη Lean (Barrett κ.ά., 2026). Κοινό σημείο με τη NumStability είναι η επαναχρησιμοποιήσιμη υποδομή για την ανάλυση αλγορίθμων. Το CSLib παρουσιάζει κυρίως αποδείξεις ορθότητας και υπολογιστικής πολυπλοκότητας, ενώ η NumStability εστιάζει στη διάδοση σφαλμάτων στρογγυλοποίησης και στην αριθμητική σταθερότητα. Ιδιαίτερα συναφής είναι η προτεινόμενη αμοιβαία ενίσχυση βιβλιοθήκης και TN: οι τυποποιήσεις παρέχουν αφαιρέσεις και δεδομένα στα συστήματα απόδειξης, ενώ τα συστήματα αυτά επιταχύνουν την επέκταση της βιβλιοθήκης. Η εργασία παρουσιάζει πρώιμη υλοποίηση και προσωρινό οδικό χάρτη, όχι όμως ζευγαρωμένη μέτρηση της επίδρασης της πρόσβασης στη βιβλιοθήκη.

Οι Meek κ.ά. (2026) παρέχουν την κύρια μεθοδολογική βάση για τον έλεγχο πιστότητας του HighamBench-E2E. Παρουσιάζουν την υποβοηθούμενη από πράκτορα TN τυποποίηση συγγράμματος αριθμητικών μεθόδων για διαφορικές εξισώσεις και αξιολογούν τρεις χωριστές ιδιότητες: την πιστότητα προς την πηγή, την επαναχρησιμοποίηση της Mathlib και την επαναχρησιμοποίηση προηγούμενων αποτελεσμάτων του ίδιου έργου. Για την πρώτη ιδιότητα χρησιμοποιούν δύο συμπληρωματικές διαδρομές. Ο άμεσος κριτής συγκρίνει την αρχική διατύπωση και τα συμφραζόμενά της με τη δήλωση Lean, εξετάζοντας και τις δύο κατευθύνσεις λογικής συνεπαγωγής. Στη δεύτερη διαδρομή, ένας τυφλός μεταφραστής αποδίδει τη δήλωση Lean σε φυσική γλώσσα χωρίς πρόσβαση στην πηγή και διαφορετικός κριτής συγκρίνει αυτή την αντίστροφη μετάφραση με το πρωτότυπο.

Η διαδικασία εντόπισε ελλιπείς πολυμερείς δηλώσεις, πρόσθετες υποθέσεις και περιορισμούς

παραμέτρων που δεν γίνονται ορατοί μόνο από τη μεταγλώττιση. Δύο συγγραφείς με εμπειρία στη Lean επανεξέτασαν ανεξάρτητα στρωματοποιημένο δείγμα είκοσι δηλώσεων και συμφώνησαν με τον άμεσο κριτή στις δεκαοκτώ· οι δύο αστοχίες αφορούσαν την ερμηνεία δηλώσεων μετά τον μετασχηματισμό υπαρξιακών ποσοδεικτών (Skolemization) (Meek κ.ά., 2026, §§ 3.2.1, 4.2.1). Το HighamBench-E2E προσαρμόζει τη διπλή αυτή αρχή, επεκτείνοντας τον έλεγχο στους εισαγόμενους ορισμούς και προσθέτοντας ανεξάρτητο επιδιαιτητή για τις ανοικτές περιπτώσεις. Αντίθετα με τους Meek κ.ά. (2026), δεν πραγματοποιήθηκε ανθρώπινη επανεξέταση δείγματος των κρίσεων. Η πλήρης διαδικασία περιγράφεται στο Κεφάλαιο 7.

Κεφάλαιο 5

Η βιβλιοθήκη NumStability

5.1 Σκοπός και εύρος

Η NumStability είναι βιβλιοθήκη Lean 4 για την τυπική ανάλυση σφαλμάτων στρογγυλοποίησης και αριθμητικής σταθερότητας αριθμητικών αλγορίθμων. Κύριο σημείο αναφοράς είναι η δεύτερη έκδοση του βιβλίου του N. J. Higham (2002). Ο σχεδιασμός οργανώνει επαναχρησιμοποιήσιμα μαθηματικά στρώματα: από τον κανόνα σφάλματος μιας στοιχειώδους πράξης έως θεωρήματα σταθερότητας σύνθετων διαδικασιών. Τα όρια του εύρους της διατυπώνονται στις Ενότητες 1.5 και 5.5.

5.2 Το μοντέλο σχετικού σφάλματος του Higham

Η ακριβής πραγματική αριθμητική χρησιμοποιείται ως μαθηματικό σημείο αναφοράς. Κάθε βασική πράξη ικανοποιεί το καθιερωμένο μοντέλο (2.2), στη μη αυστηρή εκδοχή $|\delta| \leq u$ του Higham. Στη Lean η δομή `FPModel` περιλαμβάνει έναν πραγματικό u, απόδειξη ότι $0 \leq u$, και τις πράξεις `fl_add`, `fl_sub`, `fl_mul`, `fl_div` και `fl_sqrt`. Για τη διαίρεση απαιτείται μη μηδενικός παρονομαστής. Για την τετραγωνική ρίζα ο νόμος είναι

$$\mathrm{fl}_{\sqrt{}}(x) = \sqrt{x}(1+\delta), \qquad |\delta| \leq u, \quad x \geq 0. \tag{5.1}$$

Η τετραγωνική ρίζα ξεχωρίζει επειδή δεν ανήκει στις τέσσερις δυαδικές πράξεις που απαριθμούνται στην εξίσωση (2.4) του Higham.

Ο ορισμός του μοντέλου και το ενοποιημένο θεώρημα για τις τέσσερις βασικές πράξεις βρίσκονται στο θεμελιώδες αρχείο του μοντέλου.[1]

Η δομή περιέχει ακόμη τον πρόσθετο ακριβή νόμο

$$\mathrm{fl}(0 + x) = x. \tag{5.2}$$

Ο νόμος αυτός δεν προκύπτει από την εξίσωση (2.2), η οποία θα επέτρεπε $x(1+\delta)$. Στην ανα-

[1] *Ορισμός του FPModel και νόμοι βασικών πράξεων.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/FloatingPoint/Model.lean`.

δρομική άθροιση ο Higham αρχικοποιεί ρητά το μερικό άθροισμα με $s = 0$ και έπειτα εκτελεί τις ενημερώσεις $s = s + x_i$, ενώ η ανάλυση προσμετρά $n - 1$ στρογγυλοποιημένες προσθέσεις (N. J. Higham, 2002, §§ 4.1–4.2). Η υλοποίηση της βιβλιοθήκης ως αναδίπλωση εκτελεί τυπικά και την πρώτη κλήση $\mathrm{fl}(0 + x_1)$. Ο πρόσθετος νόμος την καθιστά ακριβή και επιτρέπει να ανακτηθεί το φράγμα γ_{n-1}. Αποτελεί επομένως ρητή ενίσχυση του αφηρημένου μοντέλου, όχι συνέπεια του γενικού νόμου σχετικού σφάλματος.

Τα θεωρήματα σφάλματος που διατυπώνονται ως προς αυτό το μοντέλο δέχονται μια παράμετρο `fp : FPModel`. Κατά συνέπεια ισχύουν για οποιαδήποτε αριθμητική που ικανοποιεί τους δηλωμένους νόμους, όχι μόνο για μία προκαθορισμένη τιμή του u.

Οι συναρτήσεις `fl_*` του `FPModel` δέχονται και επιστρέφουν πραγματικούς αριθμούς. Αυτό δεν σημαίνει ότι ένας υπολογιστής αποθηκεύει αυθαίρετους πραγματικούς αριθμούς· είναι επιλογή του μαθηματικού μοντέλου για την τιμή μιας υπολογισμένης ποσότητας. Οι βοηθητικές διαταραχές, τα υπόλοιπα και οι παράγοντες θ_n μπορούν έτσι να ορίζονται ως ακριβείς πραγματικές ποσότητες που χρησιμοποιούνται μόνο στην απόδειξη. Σε ένα backward αποτέλεσμα, για παράδειγμα, αρκεί να υπάρχει μια μικρή διαταραχή της εισόδου που εξηγεί την υπολογισμένη έξοδο.

5.3 Ακριβής άλγεβρα, Mathlib και πίνακες

Η βιβλιοθήκη διαχωρίζει τη στρογγυλοποιημένη εκτέλεση από την ακριβή μαθηματική υποδομή. Οι πράξεις με πρόθεμα `fl_` ανήκουν στο πρώτο στρώμα. Για την ακριβή άλγεβρα χρησιμοποιούνται, όπου είναι διαθέσιμα, οι τύποι και τα θεωρήματα της Mathlib (Ενότητα 3.2), όπως οι ορίζουσες (The mathlib Community, 2020). Η βιβλιοθήκη διατηρεί και τοπικούς βοηθητικούς ορισμούς όπου αυτοί χρειάζονται.

Χαρακτηριστικό παράδειγμα επαναχρησιμοποίησης της Mathlib είναι οι συντομογραφίες για πραγματικούς πίνακες $m \times n$

$$\texttt{RMat}\ m\ n := \mathrm{Matrix}(\mathrm{Fin}\, m)(\mathrm{Fin}\, n)\,\mathbb{R}, \qquad \texttt{RSqMat}\ n := \texttt{RMat}\ n\ n.$$

Πρόκειται για συντομογραφίες του τύπου `Matrix` της Mathlib και όχι για νέο τοπικό τύπο πίνακα. Έτσι οι πράξεις και τα γενικά θεωρήματα πινάκων μπορούν να χρησιμοποιηθούν χωρίς μετατροπές αναπαράστασης.[2]

5.4 Συνθετική δομή αποδείξεων

5.4.1 Ο ρόλος του γ_n

Η σύμπτυξη των παραγόντων $(1 + \delta_i)$ σε έναν παράγοντα $1 + \theta_n$ με $|\theta_n| \le \gamma_n$ (εξισώσεις (2.5) και (2.6)) απαιτεί την προϋπόθεση $nu < 1$, την οποία στη Lean κωδικοποιεί η πρόταση `gammaValid`. Λήμματα σύνθεσης και μονοτονίας επιτρέπουν να συνδυάζονται πλήθη στρογγυλοποιήσεων χωρίς να επαναλαμβάνεται η ίδια πραγματική άλγεβρα. Παράλληλα, κάνουν ελέγξιμο τον αριθμό

[2] *Συντομογραφίες πινάκων βασισμένες στη Mathlib.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Analysis/MatrixAlgebra.lean`.

πράξεων πίσω από μια σταθερά όπως γ_{n+1}.

5.4.2 Παράδειγμα σύνθεσης: από την άθροιση στον πολλαπλασιασμό πινάκων

Σε μια αναδρομική άθροιση κάθε μερικό άθροισμα εισάγει νέο τοπικό δ_i, και τα λήμματα της βιβλιοθήκης συμπτύσσουν αυτούς τους παράγοντες σε componentwise διαταραχές.

Η `fl_dotProduct` ορίζει ρητά τη σειρά των στρογγυλοποιημένων πράξεων στο αφηρημένο `FPModel`: κάθε γινόμενο υπολογίζεται με `fl_mul` και τα γινόμενα συσσωρεύονται με `fl_add`. Το `dotProduct_backward_error` κατασκευάζει από αυτές ακριβώς τις πράξεις τις componentwise διαταραχές του backward αποτελέσματος.[3] Στη συνέχεια η `fl_matVec` χρησιμοποιεί το εσωτερικό γινόμενο σε κάθε γραμμή και η `fl_matMul` χρησιμοποιεί το γινόμενο πίνακα–διανύσματος σε κάθε στήλη.[4,5]

Η επαναχρησιμοποίηση έχει αριθμητική, όχι μόνο οργανωτική, αξία. Αν κάθε ανώτερο αρχείο επανέγραφε τη λογιστική των σφαλμάτων, μικρές αποκλίσεις στον αριθμό πράξεων θα μπορούσαν να παράγουν ασύμβατες σταθερές. Η κοινή αλυσίδα κάνει επίσης σαφές σε ποιο χαμηλότερο επίπεδο πρέπει να βελτιωθεί ένα φράγμα.

Δεν ακολουθούν όλες οι μαθηματικές διαδικασίες μία μοναδική γραμμική αλυσίδα. Οι τριγωνικές επιλύσεις, για παράδειγμα, χρησιμοποιούν απευθείας τα λήμματα στρογγυλοποίησης και αναδρομικής αφαίρεσης. Η κοινή αρχή είναι ότι ένα ανώτερο θεώρημα πρέπει να επαναχρησιμοποιεί το κατάλληλο χαμηλότερο αποτέλεσμα και να διατηρεί ρητές τις μαθηματικές του υποθέσεις, όπως η μη ιδιομορφία, η θετική οριστικότητα ή η `gammaValid`.

5.4.3 Ποσοτικά τεκμήρια εσωτερικής συνοχής

Η παραπάνω συνθετική δομή ελέγχθηκε ποσοτικά στην έκδοση `higham_v01`, στο commit `045daf28056a6e4358d5de7c22c7a9d7acc2e80e`. Για να μη συγχέονται οι δηλώσεις που γράφτηκαν στο έργο με τις βοηθητικές δηλώσεις που παράγει η Lean, το κύριο σύνολο αναφοράς περιορίστηκε συντηρητικά σε 47 890 δημόσιες δηλώσεις των οποίων το όνομα επιβεβαιώθηκε στην καταγεγραμμένη θέση εισαγωγής του πηγαίου κώδικα. Οι ακμές προέρχονται από τους πλήρως επεξεργασμένους τύπους και τα σώματα των δηλώσεων: μια ακμή $A \to B$ σημαίνει ότι η δήλωση A χρησιμοποιεί άμεσα τη δήλωση B. Δεν πρόκειται επομένως για απλή αναζήτηση ονομάτων στο κείμενο ή για συσχέτιση αρχείων που έχουν κοινές εισαγωγές.

[3] *Ορισμός και backward θεώρημα του εσωτερικού γινομένου.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/DotProduct.lean`.

[4] *Επαναχρησιμοποίηση του εσωτερικού γινομένου στο γινόμενο πίνακα–διανύσματος.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/MatVec.lean`.

[5] *Επαναχρησιμοποίηση του γινομένου πίνακα–διανύσματος στο γινόμενο πινάκων.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/MatMul.lean`.

Μετρική	Αναλογία	Ποσοστό
Άμεση χρήση από άλλη δήλωση του έργου	35 889/47 890	74,94%
Χρήση από δήλωση διαφορετικής υπομονάδας	12 584/47 890	26,28%
Συμμετοχή σε κατευθυντική αλυσίδα τουλάχιστον δύο ακμών	46 811/47 890	97,75%
Πλήρης απομόνωση στον γράφο δηλώσεων του έργου	143/47 890	0,30%
Δηλώσεις του στρώματος `Source` που χρησιμοποιούν κατώτερο επαναχρησιμοποιήσιμο στρώμα	22 062/30 289	72,84%

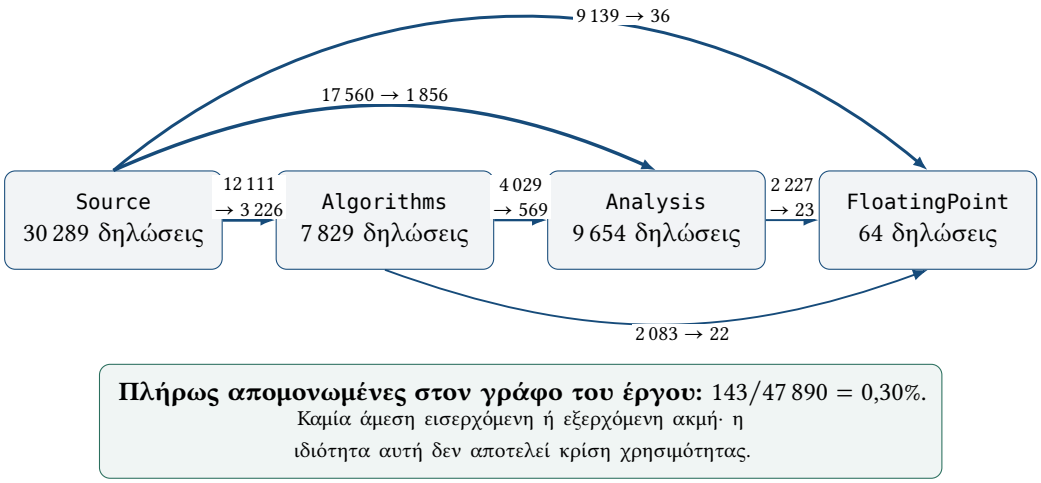
Πίνακας 5.1: Ποσοτικές ενδείξεις εσωτερικής συνοχής της τελικής έκδοσης. Η χρήση από διαφορετική υπομονάδα εξαρτάται και από τα όρια των αρχείων, ενώ οι μετρικές έχουν διαφορετική σημασία και δεν συνδυάζονται σε έναν ενιαίο δείκτη συνοχής.

Ως *υπομονάδα* μετράται ένα αρχείο Lean και ως *στρώμα* η αρχιτεκτονική κατηγορία του, όπως `Algorithms` ή `Analysis`. Στον Πίνακα 5.1, άμεση χρήση σημαίνει τουλάχιστον μία εισερχόμενη ακμή. Η πολυβηματική αλυσίδα απαιτεί κατευθυντική διαδρομή τουλάχιστον δύο ακμών, προς τις εξαρτήσεις ή προς τους καταναλωτές της δήλωσης· δεν σημαίνει συμμετοχή όλων των δηλώσεων σε μία κοινή συνεκτική ομάδα. Η πλήρης απομόνωση αποκλείει τόσο εισερχόμενες όσο και εξερχόμενες ακμές προς οποιαδήποτε άλλη δήλωση του έργου, όχι μόνο του συνόλου αναφοράς. Η απουσία εσωτερικού καταναλωτή δεν καθιστά μια δήλωση περιττή: μπορεί να αποτελεί τελικό αποτέλεσμα για εξωτερική χρήση.[6]

Το Σχήμα 5.1 συνοψίζει τις έξι κύριες σχέσεις μεταξύ των στρωμάτων, χρησιμοποιώντας μόνο ακμές μεταξύ δηλώσεων του συνόλου αναφοράς. Προκύπτει από τα αρχειοθετημένα δεδομένα του ελέγχου. Δεν απεικονίζει το βοηθητικό στρώμα `Root` ούτε τις εξαρτήσεις αντίστροφης φοράς, οι οποίες αποτελούν το 4,26% των 359 542 άμεσων ζευγών μεταξύ δηλώσεων του συνόλου· επομένως η στρωμάτωση δεν είναι απόλυτη.[7]

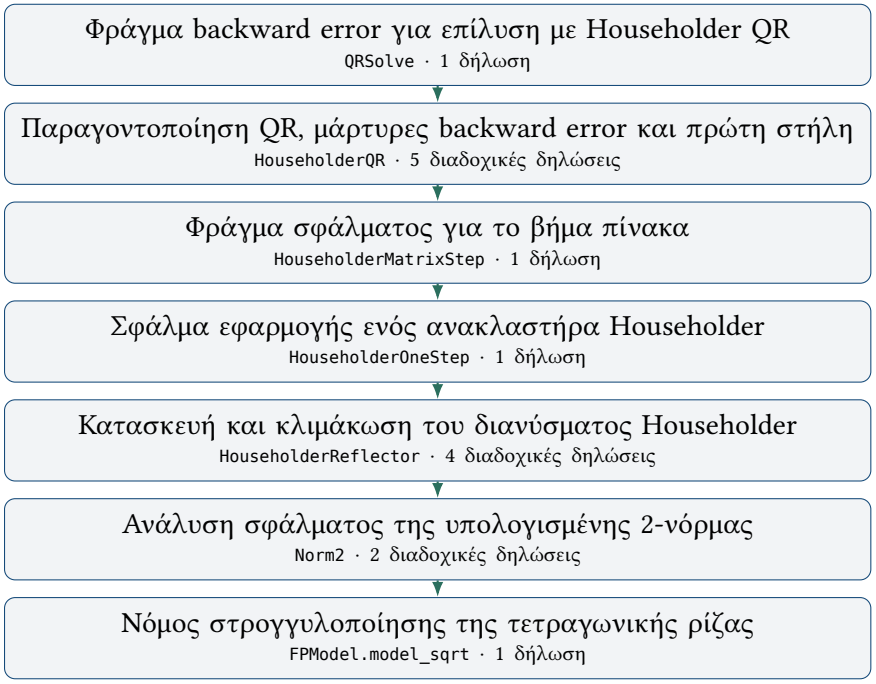
[6] *Ορισμοί και σύνολα αναφοράς των μετρικών συνοχής.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:library-audit/higham_v01-045daf2/current_graph/metrics/incoming_and_cross_boundary_coverage_contract.csv`.

[7] *Παράγωγος πίνακας εξαρτήσεων ανά στρώμα για το σύνολο αναφοράς και το script που τον παράγει.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:library-audit/higham_v01-045daf2-derived/reference_set_layer_matrix.csv`.



Σχήμα 5.1: Κύριες εξαρτήσεις μεταξύ των αρχιτεκτονικών στρωμάτων της NumStability στην έκδοση `higham_v01`. Κάθε βέλος δείχνει από το στρώμα που χρησιμοποιεί προς το στρώμα που χρησιμοποιείται. Η ετικέτα $m \to k$ σημαίνει ότι m δηλώσεις του πρώτου στρώματος χρησιμοποιούν k διαφορετικές δηλώσεις του δεύτερου. Όλα τα πλήθη αφορούν το σύνολο αναφοράς των 47 890 δηλώσεων του Πίνακα 5.1. Η πλήρης απομόνωση του κάτω πλαισίου λαμβάνει υπόψη και ακμές προς τις υπόλοιπες δηλώσεις του έργου.

Χαρακτηριστικό παράδειγμα αποτελεί μια επαληθευμένη αλυσίδα backward error για επίλυση μέσω Householder QR. Περιλαμβάνει 14 άμεσες ακμές, 15 δηλώσεις και επτά υπομονάδες. Από το θεώρημα υψηλού επιπέδου η αλυσίδα περνά σε αποτελέσματα παραγοντοποίησης, εφαρμογής και κατασκευής ανακλαστήρων, νορμών και τελικά στον νόμο `FPModel.model_sqrt` του μοντέλου στρογγυλοποίησης.



Σχήμα 5.2: Συμπτυγμένη απεικόνιση μιας επαληθευμένης ραχοκοκαλιάς της απόδειξης backward error για επίλυση μέσω Householder QR. Τα επτά πλαίσια ομαδοποιούν τις 15 δηλώσεις ανά υπομονάδα· μαζί με τις εσωτερικές εξαρτήσεις των πλαισίων η πλήρης αλυσίδα περιέχει 14 άμεσες ακμές. Κάθε βέλος έχει φορά καταναλωτή προς εξάρτηση.

Το σχήμα παρουσιάζει ένα επιλεγμένο μονοπάτι και όχι τον πλήρη γράφο της απόδειξης: δεν εμφανίζει, για παράδειγμα, τον ανεξάρτητο κλάδο για το δεξί μέλος και την οπισθοαντικατάσταση. Τεκμηριώνει συγκεκριμένη επαναχρησιμοποίηση από το τελικό φράγμα έως το μοντέλο στρογγυλοποίησης.[8]

5.4.4 Δειγματοληπτικός έλεγχος της δημόσιας διεπαφής

Η εσωτερική χρήση δεν αρκεί για να δείξει ότι μια διεπαφή μπορεί να αξιοποιηθεί από νέο κώδικα. Για αυτό επιλέχθηκε εκ των προτέρων ένα στρωματοποιημένο δείγμα 15 κανονικών δημόσιων δηλώσεων από το μοντέλο κινητής υποδιαστολής, τον λογισμό των γ_n, τις αθροίσεις και τις νόρμες έως τον πολλαπλασιασμό πινάκων, τις παραγοντοποιήσεις, τα γραμμικά συστήματα και την επαναληπτική βελτίωση. Κάθε δήλωση ελέγχθηκε πρώτα με `#check` μετά την εισαγωγή μόνο της υπομονάδας στην οποία ανήκει. Στη συνέχεια χρησιμοποιήθηκε σε χωριστά μεταγλωττισμένο, ελάχιστο εξωτερικό θεώρημα. Και οι δύο έλεγχοι πέτυχαν και για τις 15 δηλώσεις.

Το αποτέλεσμα 15/15 τεκμηριώνει ότι οι συγκεκριμένες διεπαφές είναι ορατές και καταναλώσιμες με τις καταγεγραμμένες στενές εισαγωγές. Δεν αποτελεί ποσοστό ευχρηστίας ολόκληρης της βιβλιοθήκης: το δείγμα ήταν σκόπιμα στρωματοποιημένο και όχι τυχαίο, ενώ τα περισσότερα εξωτερικά θεωρήματα ήταν σύντομα πορίσματα που προωθούν τις αντίστοιχες διεπαφές και όχι νέες εργασίες αναζήτησης απόδειξης.[9]

[8] *Μετρικές συνοχής και επαληθευμένες αλυσίδες εξαρτήσεων της έκδοσης higham_v01.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:library-audit/higham_v01-045daf2/THESIS_EVIDENCE.md`.

[9] *Προκαθορισμένο δείγμα και αποτελέσματα του ελέγχου δημόσιων διεπαφών.* Τεκμήριο στο αποθετήριο NumStability:

5.5 Σύνδεση των θεωρημάτων με συγκεκριμένους υπολογισμούς

Κατά τον εσωτερικό έλεγχο της βιβλιοθήκης χρησιμοποιήθηκαν ορισμένες περιγραφικές ετικέτες. Δεν αποτελούν κατηγορίες της Lean ούτε ορολογία του Higham. Σκοπός τους είναι να δηλώνουν με ακρίβεια από πού αρχίζει μια απόδειξη:

Συνδεδεμένο με υλοποίηση (implementation-backed). Υπάρχει ρητά ορισμένη αριθμητική διαδικασία με πράξεις του `FPModel` και αποδεικνύεται ότι η έξοδός της ικανοποιεί το επιθυμητό συμβόλαιο.

Σε επίπεδο συμβολαίου (contract-level). Το θεώρημα δέχεται το συμβόλαιο ως υπόθεση και αποδεικνύει κάποια συνέπειά του, χωρίς το ίδιο να συνδέει την υπόθεση με συγκεκριμένη διαδικασία.

Μικτό (mixed). Ένας σύνθετος αλγόριθμος έχει ορισμένα στάδια συνδεδεμένα με υλοποίηση, ενώ η ιδιότητα ενός άλλου σταδίου παραμένει υπόθεση.

Ακριβής υποδομή (exact/support). Το αρχείο περιέχει ακριβή άλγεβρα ή βοηθητικά λήμματα, οπότε δεν αναμένεται καθόλου στρογγυλοποιημένη διαδικασία.

5.5.1 Αναλυτικό παράδειγμα: εφαρμογή ανακλαστήρα Householder

Στο Λήμμα 19.2 ο Higham εξετάζει την εφαρμογή του ανακλαστήρα $P = I - vv^T$ σε ένα διάνυσμα b. Η αριθμητική διαδικασία υπολογίζει πρώτα το εσωτερικό γινόμενο $v^T b$ και έπειτα την ενημέρωση $b - v(v^T b)$. Το αποτέλεσμα $\hat{y}$ ικανοποιεί

$$\hat{y} = (P + \Delta P)b, \qquad \|\Delta P\|_F \le \tilde{\gamma}_m. \tag{5.3}$$

Η ύπαρξη του ΔP εκφράζει το backward σφάλμα της εφαρμογής και δεν σημαίνει ότι ο αλγόριθμος υπολογίζει ή αποθηκεύει αυτόν τον πίνακα (N. J. Higham, 2002, Λήμμα 19.2).

Στη NumStability, η δομή `HouseholderAppError` εκφράζει ακριβώς το είδος συμβολαίου της εξίσωσης (5.3): απαιτεί ορθογωνιότητα του P και την ύπαρξη πίνακα ΔP με το αντίστοιχο φράγμα νόρμας. Ένα θεώρημα που δέχεται μια υπόθεση αυτού του τύπου είναι πλήρως ορθό, αλλά από μόνο του δεν εξηγεί πώς προέκυψε το $\hat{y}$.[10]

Η συνάρτηση `fl_householderApply`, αντίθετα, ορίζει τη σειρά των πράξεων: στρογγυλοποιημένο εσωτερικό γινόμενο, πολλαπλασιασμό με τον συντελεστή του ανακλαστήρα και, για κάθε συνιστώσα, στρογγυλοποιημένο πολλαπλασιασμό και αφαίρεση. Το θεώρημα `fl_householderApply_normalized_appError` αποδεικνύει ότι η έξοδος αυτής της διαδικασίας ικανοποιεί το `HouseholderAppError`, εφόσον το υπολογισμένο διάνυσμα Householder ικανοποιεί ήδη το δικό του συμβόλαιο. Η απόδειξη είναι επομένως πλήρης για το στάδιο της εφαρμογής, όχι ακόμη για την κατασκευή του ανακλαστήρα.[11]

`8db102c12:library-audit/higham_v01-045daf2/PUBLIC_API_CONSUMABILITY.md`.

[10] *Το επαναχρησιμοποιήσιμο συμβόλαιο HouseholderAppError.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/LinearSystems/QR/HouseholderSpec.lean`.

[11] *Στρογγυλοποιημένη εφαρμογή Householder και θεώρημα σύνδεσης.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/LinearSystems/QR/HouseholderApply.lean`.

Το ισχυρότερο θεώρημα `fl_householderConstructApply_appError` αρχίζει από τα διανύσματα εισόδου x και b, κατασκευάζει το στροφυλοποιημένο διάνυσμα Householder, εκτελεί τη στροφυλοποιημένη εφαρμογή και αποδεικνύει το ίδιο συμβόλαιο. Παρέχει έτσι σύνδεση από την αρχή μέχρι το τέλος για έναν ανακλαστήρα, όχι για ολόκληρη την παραγοντοποίηση QR.[12]

Η σύνδεση συνοψίζεται ως εξής:

$$\begin{array}{c} \texttt{fl_householderNormalizedVector} \quad + \quad \texttt{fl_householderApply} \\ \Downarrow \\ \texttt{fl_householderConstructApply_appError} \\ \Downarrow \\ \texttt{HouseholderAppError} \end{array} \tag{5.4}$$

Το παράδειγμα δείχνει γιατί ένα συμβόλαιο δεν είναι κενό απόδειξης. Παρέχει κοινή μαθηματική διεπαφή, η οποία μπορεί να χρησιμοποιηθεί από ανώτερα θεωρήματα. Ο ισχυρότερος ισχυρισμός για συγκεκριμένο υπολογισμό προκύπτει όταν υπάρχει και χρησιμοποιείται η γέφυρα που αποδεικνύει το συμβόλαιο από τις πράξεις `fl_*`.

[12] *Κατασκευή και εφαρμογή ενός ανακλαστήρα με απόδειξη του συμβολαίου.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Algorithms/LinearSystems/QR/HouseholderOneStep.lean`.

Κεφάλαιο 6

Κατασκευή της βιβλιοθήκης και ανθρώπινη καθοδήγηση

Η NumStability κατασκευάστηκε κυρίως με το Codex, ενώ το Claude χρησιμοποιήθηκε στο πρώτο πέρασμα του βιβλίου και, έπειτα, σε επιλεγμένες τυποποιήσεις και ελέγχους. Η ανθρώπινη συμβολή εντοπίζεται κυρίως στη διαμόρφωση της μεθόδου: στην αναγνώριση ανεπαρκών αποτελεσμάτων, στον καθορισμό του μαθηματικού στόχου και στη μετατροπή των διορθωτικών αποφάσεων σε κανόνες για τις επόμενες εκτελέσεις. Το κεφάλαιο ακολουθεί αυτή τη διαδρομή, από τα προβλήματα της αρχικής τυποποίησης έως το skill και την παράλληλη κατασκευή της τελικής βιβλιοθήκης.

6.1 Το πρώτο πέρασμα χωρίς skill και αυστηρές οδηγίες

Τον Μάρτιο του 2026 αναπτύχθηκε η αρχική θεμελίωση και έγινε ένα πρώτο πέρασμα του βιβλίου με το Claude Code. Μέσα σε έναν μήνα η βιβλιοθήκη είχε 95 αρχεία Lean και περίπου 23 800 γραμμές, χωρίς να υπάρχει ακόμη skill. Η καθοδήγηση βασιζόταν σε γενικές οδηγίες και αρχεία μνήμης του πράκτορα. Αυτά ζητούσαν να επιλέγονται επαναχρησιμοποιήσιμα αποτελέσματα, αλλά δεν όριζαν πλήρως τι πρέπει να περιλαμβάνει η τυποποίηση ούτε πώς συνδέεται ένα θεώρημα σφάλματος με τον αλγόριθμο που το παράγει.[1]

Η μεταγλώττιση περνούσε και τα μηνύματα των commits ανέφεραν συχνά «0 sorry». Οι αυστηρότεροι έλεγχοι, όμως, έδειξαν ότι σε ορισμένα θεωρήματα το ζητούμενο φράγμα είχε εισαχθεί ως υπόθεση. Η LU, η Cholesky και το QR περιγράφονταν μέσω συμβολαίων, χωρίς να έχει κατασκευαστεί ο αντίστοιχος στρογγυλοποιημένος αλγόριθμος. Παράλληλα, είχαν παρακαμφθεί τμήματα του βιβλίου που ανήκαν στον επιθυμητό στόχο. Η ανθρώπινη παρέμβαση χρειάστηκε επομένως να προσδιορίσει τι συνιστά ολοκληρωμένο αποτέλεσμα, πέρα από την επιτυχή μεταγλώττιση.[2]

[1] *Ιστορικό της κατασκευής, §1.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/EARLY_HISTORY.md`.

[2] *Ιστορικό της κατασκευής, §§1–2.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/EARLY_HISTORY.md`.

6.1.1 Cholesky: το φράγμα ως υπόθεση και ως συμπέρασμα

Χαρακτηριστικό παράδειγμα είναι η αρχική δήλωση `cholesky_backward_error_perturbation`. Δεχόταν έναν αυθαίρετο πίνακα $\widehat{R}$ και μια υπόθεση `hChol`, η οποία περιείχε ήδη το componentwise φράγμα για το $\widehat{R}^T\widehat{R} - A$. Η απόδειξη όριζε $\Delta A = \widehat{R}^T\widehat{R} - A$ και επαναδιατύπωνε αυτό το φράγμα ως ύπαρξη μικρής διαταραχής. Πρόκειται για ορθό αλγεβρικό βήμα, αλλά όχι για παραγωγή του φράγματος από την παραγοντοποίηση Cholesky.[3]

Η διορθωτική απόφαση ήταν συγκεκριμένη: να διατηρηθεί ο μαθηματικός στόχος και να κατασκευαστεί ο αλγόριθμος που έλειπε, αντί να υποτεθεί το αποτέλεσμά του. Στην έκδοση `higham_v01`, ο πίνακας $\widehat{R}$ παράγεται από τη `fl_cholesky`, η οποία περιγράφει τον Αλγόριθμο 10.2 με πράξεις του `FPModel`. Το θεώρημα `higham10_3_fl_cholesky_backward_error` αποδεικνύει πλέον το φράγμα για αυτή την έξοδο, όπως συνοψίζει το Σχήμα 6.1.[4]

Πριν: το φράγμα ως υπόθεση

Αυθαίρετος $\widehat{R}$, με δεδομένο

$$|\widehat{R}^T\widehat{R} - A| \le \varepsilon|\widehat{R}^T||\widehat{R}|.$$

Θέτουμε $\Delta A = \widehat{R}^T\widehat{R} - A$ και συμπεραίνουμε

$$\widehat{R}^T\widehat{R} = A + \Delta A,$$
$$|\Delta A| \le \varepsilon|\widehat{R}^T||\widehat{R}|.$$

Το φράγμα προϋπάρχει στην υπόθεση `hChol`.

Μετά: το φράγμα αποδεικνύεται

Ο $\widehat{R}$ υπολογίζεται από τη

$$\widehat{R} = \texttt{fl_cholesky}\ (fp, n, A).$$

Υπό τις προϋποθέσεις του θεωρήματος, αποδεικνύεται ότι υπάρχει ΔA με

$$\widehat{R}^T\widehat{R} = A + \Delta A,$$
$$|\Delta A| \le \gamma_{n+1}|\widehat{R}^T||\widehat{R}|.$$

Το φράγμα προκύπτει από τις στρογγυλοποιημένες πράξεις.

Σχήμα 6.1: Η αλλαγή στην τυποποίηση της Cholesky. Οι ανισότητες είναι componentwise και η γραφή είναι συνοπτική μαθηματική απόδοση των δηλώσεων Lean. Δεξιά απαιτούνται συμμετρία του A, $(n+1)u < 1$, μη αρνητικοί υπολογισμένοι οδηγοί και μη μηδενική υπολογισμένη διαγώνιος.

Η απαίτηση αυτή γενικεύτηκε σε κανόνα του skill: για κάθε επιλεγμένο αλγοριθμικό θεώρημα, η απόδειξη πρέπει να συνδέει τις πράξεις `fl_*` με το συμβόλαιο και το τελικό φράγμα. Έτσι είναι δυνατόν να μελετηθεί και μια παραλλαγή των βημάτων του αλγορίθμου, αντί να θεωρείται το σφάλμα του δεδομένο. Το γενικό σχήμα αυτής της σύνδεσης παρουσιάστηκε στην Ενότητα 5.5.[5]

6.2 Από τα προβλήματα στους κανόνες του skill

Η Cholesky ήταν μία από τις περιπτώσεις που ανέδειξαν την ανάγκη αυστηρότερων οδηγιών. Ο Πίνακας 6.1 συνοψίζει τις βασικές αντιστοιχίες μεταξύ ευρημάτων και ανθρώπινων αποφάσεων.

[3] *Ο αρχικός έλεγχος της σύνδεσης συμβολαίων και αλγορίθμων· Cholesky.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/project-memory/IMPLEMENTATION_BACKED_AUDIT.md`.

[4] *Το backward φράγμα για τη στρογγυλοποιημένη Cholesky.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:historical-library-snapshot/NumStability/Source/Higham/Chapter10/Section02/ErrorAnalysis/FactorizationAndSolve/Bounds.lean`.

[5] *Ιστορικό της κατασκευής, §3.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/EARLY_HISTORY.md`.

Η παρέμβαση δεν περιοριζόταν στη διόρθωση ενός αρχείου: η απόφαση καταγραφόταν ως απαίτηση για ανάλογες περιπτώσεις σε όλη τη βιβλιοθήκη και για τις επόμενες εκτελέσεις των πρακτόρων.[6]

Πρόβλημα που εντοπίστηκε	Ανθρώπινη απόφαση και κανόνας του skill
Το φράγμα δινόταν ως υπόθεση, χωρίς τον στρογγυλοποιημένο αλγόριθμο.	Να κατασκευάζεται ο αλγόριθμος και να αποδεικνύεται το φράγμα από τις πράξεις του, χωρίς εξασθένηση του μαθηματικού στόχου.
Παράλληλες περιορισμένες και «ασφαλείς» εκδοχές, όπως στο Householder QR.	Μία εκδοχή, με τη γενικότητα που απαιτεί η πηγή, η οποία να καλύπτει και τις επιτρεπτές εκφυλισμένες περιπτώσεις.
Μαθηματικό περιεχόμενο του βιβλίου είχε παραλειφθεί.	Εξέταση κάθε σελίδας και καταγραφή των στοιχείων της, με ρητή αιτιολόγηση για όσα επιλέγονται, αναβάλλονται ή παραλείπονται.
Τοπικοί επαναορισμοί εννοιών της Mathlib, όπως οι νόρμες.	Αναζήτηση και επαναχρησιμοποίηση της υπάρχουσας υποδομής πριν από κάθε νέο ορισμό· μόνο αναγκαίες προσαρμογές αναπαράστασης.
Σφάλματα της πηγής οδηγούσαν σε άπιστες μεταφράσεις, παραλείψεις ή ατέρμονες προσπάθειες.	Καμία σιωπηρή διόρθωση: αναφορά του προβλήματος και, όπου τεκμηριώνεται, σαφώς διακριτή διορθωμένη εκδοχή, όπως στην άθροιση Kahan.

Πίνακας 6.1: Ευρήματα της αρχικής τυποποίησης και ανθρώπινες αποφάσεις που μετατράπηκαν σε απαιτήσεις του skill για τις επόμενες εκτελέσεις.

6.2.1 Τι σημαίνει να τυποποιηθεί ένα βιβλίο

Ο καθορισμός του εύρους ήταν επίσης ανθρώπινη σχεδιαστική απόφαση. Η απαίτηση εξέτασης κάθε σελίδας δεν σήμαινε τυποποίηση κάθε πρότασής της. Επιλέχθηκε ο μαθηματικός πυρήνας: ονομασμένα θεωρήματα, λήμματα και πορίσματα, αριθμημένες μαθηματικές εξισώσεις, κεντρικοί ακριβείς ισχυρισμοί και οι αναγκαίοι ορισμοί. Οι οδηγίες απαιτούσαν πιστή απόδοση των επιλεγμένων διατυπώσεων, χωρίς ευκολότερες υποθέσεις για τα φράγματα ή τις κλάσεις πινάκων. Για κάθε στοιχείο προβλεπόταν καταγραφή της απόφασης: τυποποίηση, επαναχρησιμοποίηση, αναβολή ή παράλειψη με αιτιολόγηση.[7]

Ιστορικές σημειώσεις, βιβλιογραφικές ανασκοπήσεις, πίνακες χρόνων MATLAB και εμπειρικά αποτελέσματα δεν αποτελούσαν στόχους τυποποίησης, εκτός αν ήταν αναγκαία για επιλεγμένο θεώρημα. Δεν κατασκευαζόταν υποθετικό μηχάνημα για να αναπαραχθεί μια ιστορική αριθμητική τιμή με ελλιπή προδιαγραφή. Ως προτεραιότητα ανάπτυξης επιλέχθηκαν οι πραγματικοί πίνακες πριν από τους μιγαδικούς.[8]

Η πολιτική για τις ασκήσεις αναθεωρήθηκε στην πορεία. Αρχικά ήταν στόχοι τυποποίησης, έπειτα αποκλείστηκαν για να διατηρηθούν άλυτες για την αξιολόγηση και τελικά επιτράπηκαν

[6] *Κανόνες καταγραφής, επαναχρησιμοποίησης και αναφοράς σφαλμάτων της πηγής.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/chapter_splitting/skills/higham-chapter-formalization/references/workflow.md`.

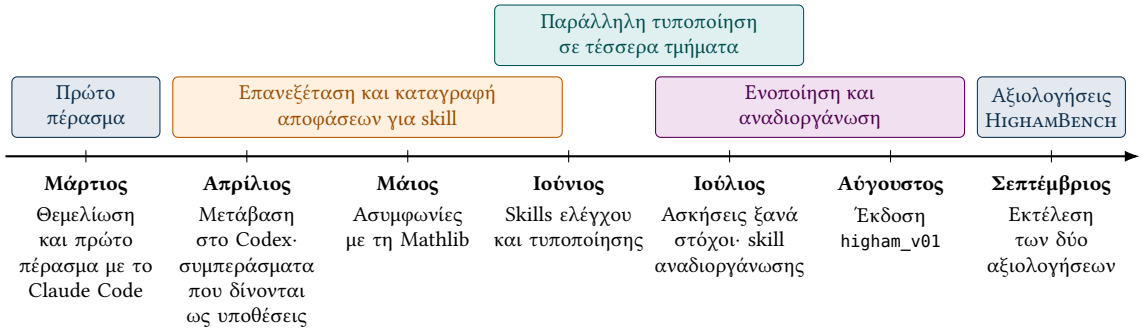
[7] *Κανόνες επιλογής περιεχομένου του skill.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/chapter_splitting/skills/higham-chapter-formalization/references/content-selection.md`.

[8] *Ιστορικό της κατασκευής, §4.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/EARLY_HISTORY.md`.

ξανά. Αυτό γιατί ο Higham αφήνει ορισμένες αποδείξεις ως ασκήσεις, ενώ τα αποτελέσματά τους χρησιμοποιούνται σε επόμενα κεφάλαια. Η βιβλιοθήκη περιλαμβάνει επομένως τέτοιες ασκήσεις όταν χρειάζονται για τη μαθηματική συνέχεια του επιλεγμένου υλικού.

6.3 Εφαρμογή του skill στην παράλληλη κατασκευή

Οι αποφάσεις αυτές συγκροτήθηκαν σταδιακά σε κοινή διαδικασία (Σχήμα 6.2). Τον Απρίλιο δημιουργήθηκαν αρχεία μνήμης και αποφάσεων για το Codex. Στα τέλη Μαΐου τέθηκε το ερώτημα αν ένα skill θα βοηθούσε την επανεξέταση και, μέσα στον Ιούνιο, ακολούθησαν τα skills ελέγχου και τυποποίησης. Ρητή οδηγία ήταν οι ανθρώπινες διορθώσεις να αντιμετωπίζονται ως έλεγχοι παλινδρόμησης της διαδικασίας: να ενημερώνεται ο κανόνας και να εξετάζονται ανάλογες περιπτώσεις, ώστε το ίδιο πρόβλημα να μην αντιμετωπίζεται εκ νέου ως μεμονωμένο περιστατικό.[9]



Σχήμα 6.2: Γενικό χρονολόγιο της κατασκευής της NumStability, από τον Μάρτιο έως τον Σεπτέμβριο του 2026. Πάνω από τον άξονα δίνονται οι φάσεις και κάτω οι κύριες ανθρώπινες αποφάσεις και τα ορόσημα κάθε μήνα.

6.3.1 Ο καταμερισμός και η σταδιακή παραλληλοποίηση

Ο καταμερισμός κεφαλαίων ήταν το σχέδιο εργασίας του έργου. Το πρώτο τμήμα είχε χρονικό προβάδισμα, επειδή έπρεπε να διαθέσει τις θεμελιώδεις δηλώσεις πριν αναπτυχθούν συστηματικά οι ανώτερες ενότητες. Στη συνέχεια, τα είκοσι οκτώ κεφάλαια και το επιλεγμένο υλικό του Παραρτήματος Α χωρίστηκαν ως εξής:[10]

- η πρώτη ροή ανέλαβε τα Κεφάλαια 1–6: μοντέλα αριθμητικής, λογισμό σφαλμάτων, αθροίσεις, πολυώνυμα, νόρμες και βασικά αποτελέσματα SVD,
- η δεύτερη τα Κεφάλαια 7–12: γραμμικά συστήματα, τριγωνικές επιλύσεις, LU, Cholesky και επαναληπτική βελτίωση,
- η τρίτη τα Κεφάλαια 13–19: μεθόδους κατά τμήματα, αντιστροφή πινάκων, εκτίμηση κατάστασης, εξισώσεις Sylvester, στάσιμες επαναληπτικές μεθόδους, δυνάμεις πινάκων και QR,
- η τέταρτη τα Κεφάλαια 20–28: ελάχιστα τετράγωνα, υποορισμένα συστήματα, συστήματα

[9] *Ιστορικό της κατασκευής, §5.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/EARLY_HISTORY.md`.

[10] *Σχέδιο της παράλληλης τυποποίησης.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/chapter_splitting/HIGHAM_PARALLEL_FORMALIZATION_BLUEPRINT.md`.

Vandermonde, γρήγορους αλγορίθμους, FFT, μη γραμμικά προβλήματα, ζητήματα λογισμικού και δοκιμαστικούς πίνακες.

Τα συνεχόμενα εύρη έδιναν σαφή περιοχή ευθύνης σε κάθε ροή, ενώ η σειρά έναρξης εξυπηρετούσε τις εξαρτήσεις: ο λογισμός σχετικού σφάλματος και οι νόρμες έπρεπε να είναι διαθέσιμα πριν χρησιμοποιηθούν σε ανώτερους αλγορίθμους.

Για κάθε επαναχρησιμοποιήσιμο αποτέλεσμα το σχέδιο ζητούσε αναφορά στην πηγή, υπεύθυνη ροή, προτεινόμενη δήλωση, εξαρτήσεις και κατάσταση. Μια ροή μπορούσε προσωρινά να χρησιμοποιήσει το κοινό συμβόλαιο διεπαφής που ανήκε σε άλλη ροή, δηλαδή την ακριβή δήλωση Lean την οποία είχαν συμφωνήσει οι ροές, αλλά δεν έπρεπε να δημιουργήσει δεύτερη δημόσια εκδοχή της ίδιας έννοιας. Κατά την ενοποίηση, οι προσωρινές δηλώσεις έπρεπε να αντικαθίστανται από τις κανονικές δηλώσεις που υλοποιούσαν το συμβόλαιο.

Ο κανόνας της επιστροφής στην πηγή είχε πρακτικό αποτέλεσμα: το Θεώρημα 9.7, που απουσίαζε από παλαιότερο ευρετήριο σχεδιασμού, ανακτήθηκε σε μεταγενέστερο έλεγχο απευθείας από το PDF. Το ευρετήριο παρέμενε έτσι ενημερώσιμο βοήθημα και όχι υποκατάστατο του βιβλίου.[11]

6.3.2 Ο κύκλος εργασίας, η ενοποίηση και η αναδιοργάνωση

Το skill `higham-chapter-formalization` μετέτρεπε το γενικό αίτημα «τυποποίησε ένα κεφάλαιο» σε επαναλήψιμη διαδικασία, με τον κύκλο του Σχήματος 6.3:[12]

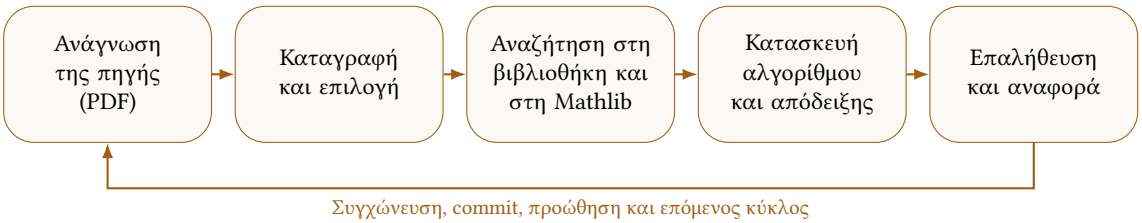
Αφετηρία ήταν τα PDF των κεφαλαίων, μαζί με τα τοπικά αρχεία που εξαιρούνταν από το Git λόγω πνευματικών δικαιωμάτων. Η επιλογή ακολουθούσε τα κριτήρια της Ενότητας 6.2.1. Πριν εισαχθεί νέος ορισμός ή λήμμα, ζητούνταν αναζήτηση στη NumStability και στη Mathlib (The mathlib Community, 2020), και καθορισμός των τύπων, διαστάσεων, ποσοδεικτών και πλευρικών συνθηκών. Για συγκεκριμένο αλγόριθμο, η έξοδος έπρεπε να παράγεται από τα βήματά του, όπως επέβαλε η διόρθωση της αρχικής τυποποίησης.

Το skill απαιτούσε ο κύκλος να κλείνει με μεταγλώττιση, έλεγχο μη επιτρεπτών κενών και παραδοχών και αναφορά των αποκλίσεων. Προέβλεπε επίσης συγχώνευση και προώθηση των αλλαγών στο κοινό αποθετήριο, με νέα μεταγλώττιση αν η συγχώνευση μετέβαλλε αρχεία. Έτσι όριζε κοινές απαιτήσεις για κάθε ροή, χωρίς να χρειάζεται να διατυπώνονται ξανά σε κάθε ανάθεση.

Η παράλληλη ανάπτυξη άφησε προσωρινά συμβόλαια, πιθανά διπλότυπα και ασυνέπειες στην ονοματοδοσία και στην ιεραρχία αρχείων. Στην ενοποίηση τα συμβόλαια αντικαταστάθηκαν από τις κανονικές δηλώσεις και ακολούθησε η αναδιοργάνωση που οδήγησε στην έκδοση `higham_v01` (Κεφάλαιο 5). Οι επαναληπτικοί έλεγχοι συνέχισαν τον ίδιο κύκλο: εύρημα, διόρθωση και νέος έλεγχος.

[11] *Καταγραφή της ανάκτησης του Θεωρήματος 9.7.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:source-coverage/AUDIT_ch01-28_PDF_FIRST_2026-07-21.md`.

[12] *Το skill τυποποίησης ανά κεφάλαιο.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:construction-history/chapter_splitting/skills/higham-chapter-formalization/SKILL.md`.



Σχήμα 6.3: Ο κύκλος του skill σε κάθε τμήμα κάθε ροής. Οι ανθρώπινες αποφάσεις του Πίνακα 6.1 καθορίζουν τι επιλέγεται, πώς κατασκευάζεται και με ποιους ελέγχους γίνεται δεκτό.

6.4 Έλεγχοι, ευρήματα και όρια

Οι μεταγενέστεροι έλεγχοι της βιβλιοθήκης περιλάμβαναν πλήρη μεταγλώττιση, αναζήτηση μη επιτρεπτών δεικτών εκκρεμότητας, εξέταση αξιωμάτων σε επιλεγμένα τελικά θεωρήματα, έλεγχο των εισαγωγών και επιλεγμένες συγκρίσεις δηλώσεων με τα PDF. Οι συγκρίσεις με τα PDF ανατίθενταν σε νέους πράκτορες του Codex και, σε ορισμένους ελέγχους, του Claude, χωρίς μνήμη των συνεδριών που είχαν παραγάγει τις αποδείξεις. Η ρύθμιση αυτή περιορίζει την εξάρτηση από προηγούμενα συμπεράσματα, δεν αποκλείει όμως συστηματικά σφάλματα που μπορεί να μοιράζονται διαδοχικές εκτελέσεις του ίδιου μοντέλου.

Η περίπτωση της αντισταθμισμένης άθροισης Kahan δείχνει την αξία ενός τέτοιου ελέγχου. Το ζήτημα δεν ήταν λανθασμένη μετάφραση του Codex. Στις εξισώσεις (4.8)–(4.9) του Higham, ο κύριος όρος $2u$ για το αποθηκευμένο άθροισμα της συνήθους μορφής του αλγορίθμου παραλείπει το τελικό σφάλμα στρογγυλοποίησης. Οι Hallman και Ipsen δίνουν για αυτή την έξοδο κύριο όρο $3u$, ενώ ο όρος $2u$ αντιστοιχεί στην παραλλαγή που επιστρέφει τη διορθωμένη τελική τιμή $s_n + c_n$ (Hallman & Ipsen, 2023, Παρατήρηση 4.3). Ο νέος έλεγχος με αφετηρία τα PDF, ο οποίος δεν στηρίχθηκε σε παλαιότερες καταγραφές κάλυψης, το επιβεβαίωσε με οικογένεια δυαδικών συστημάτων η οποία, για κάθε σταθερό C ανεξάρτητο από την ακρίβεια, αποκλείει την καθολική εκδοχή με κύριο μέρος $2u + Cu^2$, και απέδειξε διορθωμένη εκδοχή με $3u$ και ρητούς όρους ανώτερης τάξης.[13] Ο έλεγχος με το Codex δεν αναπαρήγαγε επομένως άκριτα το σφάλμα του βιβλίου.[14]

Η ανθρώπινη προσπάθεια στην κατασκευή αφορούσε επομένως τόσο τη διόρθωση της πορείας όσο και τη διαμόρφωση μιας διαδικασίας που μπορούσαν να ακολουθήσουν οι πράκτορες σε μεγάλη κλίμακα. Η ιστορία αυτή δεν αποτελεί ελεγχόμενο πείραμα με και χωρίς skill ούτε εξαντλητικό ανθρώπινο έλεγχο της βιβλιοθήκης. Δείχνει, όμως, πώς συγκεκριμένα προβλήματα μετατράπηκαν σε απαιτήσεις σχεδιασμού και σε επαναχρησιμοποιήσιμες οδηγίες.

[13] Υπομονάδες της έκδοσης higham_v01: `67344c880:historical-library-snapshot/NumStability/Source/Higham/Chapter04/Equation08/FiniteFamily.lean` για το αντιπαράδειγμα και `67344c880:historical-library-snapshot/NumStability/Source/Higham/Chapter04/Equation09/Correction.lean` για το διορθωμένο όριο.

[14] *Έλεγχος κάλυψης και πηγών με αφετηρία τα PDF· περίπτωση Kahan.* Τεκμήριο στο αποθετήριο NumStability: `8db102c12:source-coverage/AUDIT_ch01-28_PDF_FIRST_INDEPENDENT_2026-07-21.md`.

Κεφάλαιο 7

HighamBench: μεθοδολογία των δύο αξιολογήσεων

Το HighamBench εξετάζει πειραματικά αν η πρόσβαση ενός πράκτορα TN στη NumStability μειώνει το κόστος νέων αποδείξεων. Στο πλαίσιο της εργασίας εκτελέστηκαν *δύο* αξιολογήσεις με το ίδιο ζευγαρωμένο σχέδιο αλλά με διαφορετικό αντικείμενο μέτρησης:

1. **HighamBench-Proof: απόδειξη σταθερών στόχων (30 εργασίες).** Ο πράκτορας λαμβάνει ένα ήδη διατυπωμένο θεώρημα Lean, που προέρχεται από δημοσιευμένο άρθρο, και πρέπει να το αποδείξει χωρίς να αλλάξει τη διατύπωσή του. Κάθε εργασία εκτελείται τρεις φορές σε κάθε συνθήκη.
2. **HighamBench-E2E: τυποποίηση και απόδειξη (10 εργασίες).** Ο πράκτορας λαμβάνει το PDF της πηγής και πρέπει πρώτα να διατυπώσει πιστή δήλωση Lean, η οποία ελέγχεται από ανεξάρτητους κριτές, και έπειτα να την αποδείξει. Κάθε εργασία εκτελείται μία φορά σε κάθε συνθήκη. Το E2E (end-to-end) δηλώνει ότι η εργασία καλύπτει όλη τη διαδρομή, από την πηγή ως την απόδειξη.

Οι δύο αξιολογήσεις διαφέρουν ως προς τις εργασίες, το μοντέλο και τις υποχρεώσεις του πράκτορα. Γι' αυτό τα αποτελέσματά τους παρουσιάζονται χωριστά και δεν συνενώνονται.

7.1 Ερευνητικό ερώτημα και συνθήκες

Ερευνητικό ερώτημα

Όταν ο ίδιος πράκτορας εργάζεται στην ίδια εργασία με και χωρίς πρόσβαση στη NumStability, πώς μεταβάλλονται ο χρόνος, τα tokens και το μέγεθος του κώδικα που απαιτούνται για μια αποδεκτή απόδειξη, και σχετίζονται οι μεταβολές με την πραγματική χρήση δηλώσεων της βιβλιοθήκης;

Και στις δύο αξιολογήσεις συγκρίνονται δύο συνθήκες. Στη συνθήκη N ο πράκτορας διαθέτει τη Lean και τη Mathlib, χωρίς πηγαίο κώδικα, μεταγλωττισμένα αρχεία, ευρετήριο ή γενικότερα γνώση για την ύπαρξη της NumStability. Στη συνθήκη L διαθέτει το ίδιο περιβάλλον και, επι-

πλέον, μια σταθερή έκδοση της NumStability, σε πηγαίο και μεταγλωττισμένο κώδικα, μόνο για ανάγνωση, μαζί με σύντομη οδηγία που ενθαρρύνει την αναζήτηση στη βιβλιοθήκη. Η παρέμβαση είναι η διαθεσιμότητα ολόκληρης της βιβλιοθήκης και όχι ένα θεώρημα που επιλέγεται για κάθε εργασία· καμία συνθήκη δεν διαθέτει αυτόματο σύστημα ανάκτησης ή κατάλογο σχετικών δηλώσεων. Η αναζήτηση μέσα στη βιβλιοθήκη είναι μέρος της μετρούμενης εργασίας. Κάθε προσπάθεια αρχίζει σε νέα συζήτηση και σε απομονωμένο χώρο εργασίας, χωρίς πρόσβαση στο διαδίκτυο, στην έξοδο της άλλης συνθήκης ή σε προηγούμενα αποτελέσματα.

Ο Πίνακας 7.1 συγκρίνει τα δύο πρωτόκολλα. Το HighamBench-Proof προηγήθηκε χρονικά.

	Proof: σταθερός στόχος	E2E: τυποποίηση και απόδειξη
Εργασία του πράκτορα	Απόδειξη ενός δοσμένου θεωρήματος Lean χωρίς αλλαγή της διατύπωσης	Πιστή διατύπωση του αποτελέσματος από το PDF και έπειτα απόδειξη της οριστικής δήλωσης
Πιστότητα προς την πηγή	Έλεγχος κατά την κατασκευή των στόχων· όχι νέος έλεγχος κατά τη μέτρηση	Ανεξάρτητος έλεγχος κάθε υποβαλλόμενης δήλωσης πριν από την απόδειξη
Δείγμα	30 εργασίες από 25 άρθρα· τρία ζεύγη N/L ανά εργασία	10 εργασίες από 5 πηγές· ένα ζεύγος ανά εργασία
Μοντέλο του πράκτορα	GPT-5.6 Sol, `xhigh` (OpenAI, 2026b)	GPT-6 Sol, `high` (OpenAI, 2026d)
Εξοικείωση με τη βιβλιοθήκη	Μόνο σύντομη οδηγία της L	Οδηγία και μία κοινή συζήτηση εξοικείωσης πριν από τις εργασίες (warm-up)
Μετρούμενα στάδια	Ένα χρονόμετρο, μέχρι την τελική υποβολή	Χωριστά χρονόμετρα για τυποποίηση και απόδειξη

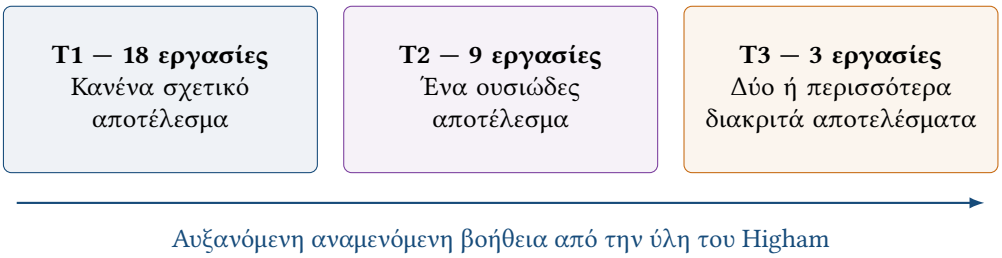
Πίνακας 7.1: Οι δύο αξιολογήσεις που εκτελέστηκαν. Ο ορισμός της άμεσης χρήσης της βιβλιοθήκης στην απόδειξη, ο κανόνας μέτρησης των γραμμών κώδικα και ο τύπος των νέων tokens είναι κοινοί· όλες οι υπόλοιπες διαστάσεις διαφέρουν.

7.2 HighamBench-Proof: απόδειξη σταθερών στόχων

7.2.1 Σύνολο εργασιών και αναμενόμενη βοήθεια

Τα άρθρα του HighamBench-Proof επιλέχθηκαν θεματικά, γύρω από τα πεδία που καλύπτει το βιβλίο του N. J. Higham (2002): μοντέλα σφάλματος κινητής υποδιαστολής, αλγόριθμοι, φράγματα διαταραχών και αναλύσεις σταθερότητας. Η επιλογή ήταν σκόπιμη και όχι τυχαίο δείγμα της βιβλιογραφίας. Από κάθε άρθρο κατασκευάζονταν έως τρεις εργασίες, μία για κάθε κατηγορία αναμενόμενης βοήθειας. Η αρίθμηση ακολουθεί την αναμενόμενη βοήθεια, που αυξάνεται από την T1 στην T3 (Σχήμα 7.1).

Μια κατηγορία παραλειπόταν όταν το άρθρο δεν προσέφερε κατάλληλο αποτέλεσμα. Το αν οι κατηγορίες συμφωνούν με την παρατηρούμενη χρήση αποτελεί αποτέλεσμα της ανάλυσης και όχι προϋπόθεσή της. Η επιλογή των αποτελεσμάτων και η κατασκευή των εργασιών ανατέθηκαν, σχεδόν σε όλα τα άρθρα, στο Codex· τα αρχεία κατασκευής καταγράφουν ρητά ότι οι επιλογές



Σχήμα 7.1: Οι κατηγορίες εργασιών του HighamBench-Proof. Τα αποτελέσματα που προσμετρώνται στις T2 και T3 είναι διαθέσιμα στη NumStability αλλά όχι στη Mathlib. Η ταξινόμηση προηγείται των εκτελέσεων: δεν μετρά τη δυσκολία ούτε την πραγματική χρήση της βιβλιοθήκης, η οποία καταγράφεται με το DP.

αυτές δεν αποτελούν ανθρώπινη επανεξέταση.[1]

Το τελικό σύνολο περιλαμβάνει 30 εργασίες από 25 άρθρα. Ο πλήρης κατάλογος των άρθρων και των εργασιών βρίσκεται στο Παράρτημα Β΄ (Πίνακας Β΄.1). Κάθε άρθρο διαθέτει ένα κοινό αρχείο ορισμών Lean με τα μοντέλα και τους αλγορίθμους που χρειάζονται οι εργασίες του, και κάθε εργασία περιέχει σταθερή διατύπωση Lean του επιλεγμένου αποτελέσματος. Η πιστότητα αυτών των διατυπώσεων προς τα άρθρα ελέγχθηκε κατά την κατασκευή τους από πράκτορες Codex και δεν επανελέγχεται κατά τη μέτρηση. Δεν εφαρμόστηκε εδώ το σύστημα ανεξάρτητου άμεσου ελέγχου και αντίστροφης μετάφρασης του HighamBench-E2E (Ενότητα 7.3.3).

7.2.2 Ο σταθερός στόχος και η επικύρωση

Και οι δύο συνθήκες λαμβάνουν το ίδιο πλαίσιο της εργασίας και την ίδια υπογραφή θεωρήματος Lean. Ο πράκτορας πρέπει να διατηρήσει τον χώρο ονομάτων, το όνομα, τα ορίσματα, τις υποθέσεις και το συμπέρασμα του στόχου και να παραδώσει πλήρες αρχείο `Candidate.lean`. Η κοινή οδηγία ζητά έλεγχο του πλήρους αρχείου πριν από την υποβολή και απαγορεύει `sorry`, `admit`, νέα αξιώματα, μη ασφαλείς μηχανισμούς και χρήση δικτύου.

Το χρονόμετρο μετρά από τη στιγμή που δίνεται η οδηγία μέχρι την τελική υποβολή. Οι ενδιάμεσοι έλεγχοι στη Lean που επιλέγει ο πράκτορας μετρούν στον χρόνο του. Ο κρυφός τελικός έλεγχος γίνεται μετά την υποβολή, σε απομονωμένο περιβάλλον: μεταγλωττίζει το αρχείο, ελέγχει ότι η διατύπωση ταυτίζεται με τον δοσμένο στόχο, απορρίπτει απαγορευμένους μηχανισμούς και καταγράφει πλήρη έλεγχο των εξαρτήσεων. Το όριο ήταν 7 200 s ανά προσπάθεια, χωρίς όριο tokens.

7.2.3 Μετρήσεις

Για κάθε εργασία μετρώνται τρία μεγέθη σε κάθε επανάληψη: ο χρόνος μέχρι την υποβολή, τα νέα tokens και οι γραμμές του υποβληθέντος αρχείου που περιέχουν κώδικα Lean, χωρίς κενές γραμμές και γραμμές μόνο με σχόλια. Οι γραμμές περιλαμβάνουν τα βοηθητικά λήμματα και τις εισαγωγές του αρχείου, όχι όμως τον κώδικα των εισαγόμενων βιβλιοθηκών. Η τιμή μιας

[1] *Κατάλογος άρθρων, θέσεων των αποτελεσμάτων και καταγραφή της κατασκευής από το Codex.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:paper_bencmark/highambench/metadata/manifest.json`.

εργασίας είναι ο μέσος όρος των τριών επαναλήψεων. Οι ιστορικές καταγραφές των tokens περιλαμβάνουν μόνο τις ολοκληρωμένες απαντήσεις του παρόχου, οπότε οι τιμές αποτελούν κάτω φράγματα της πραγματικής κατανάλωσης. Η χρήση της βιβλιοθήκης μετράται με το κοινό μέτρο DP (Ενότητα 7.4).[2]

7.3 HighamBench-E2E: τυποποίηση και απόδειξη

7.3.1 Σύνολο εργασιών

Το HighamBench-E2E περιλαμβάνει δέκα εργασίες από πέντε πηγές: τρεις από την ανάλυση των αλγορίθμων superblock για το εσωτερικό γινόμενο (Castaldo κ.ά., 2008), τέσσερις από την ανάλυση των log-sum-exp και softmax (Blanchard κ.ά., 2021), και μία από καθεμία από τις Hallman και Ipsen (2021), N. J. Higham (2002) και Rump (2012). Κάθε εργασία συνοδεύεται από πακέτο πηγής με τη βιβλιογραφική αναφορά, το αποτύπωμα του PDF, τις ακριβείς θέσεις του αποτελέσματος και διευκρινίσεις για το πλήρες εύρος του. Ο πλήρης κατάλογος των εργασιών και των αποτελεσμάτων των πηγών βρίσκεται στο Παράρτημα Γ΄ (Πίνακας Γ΄.1).

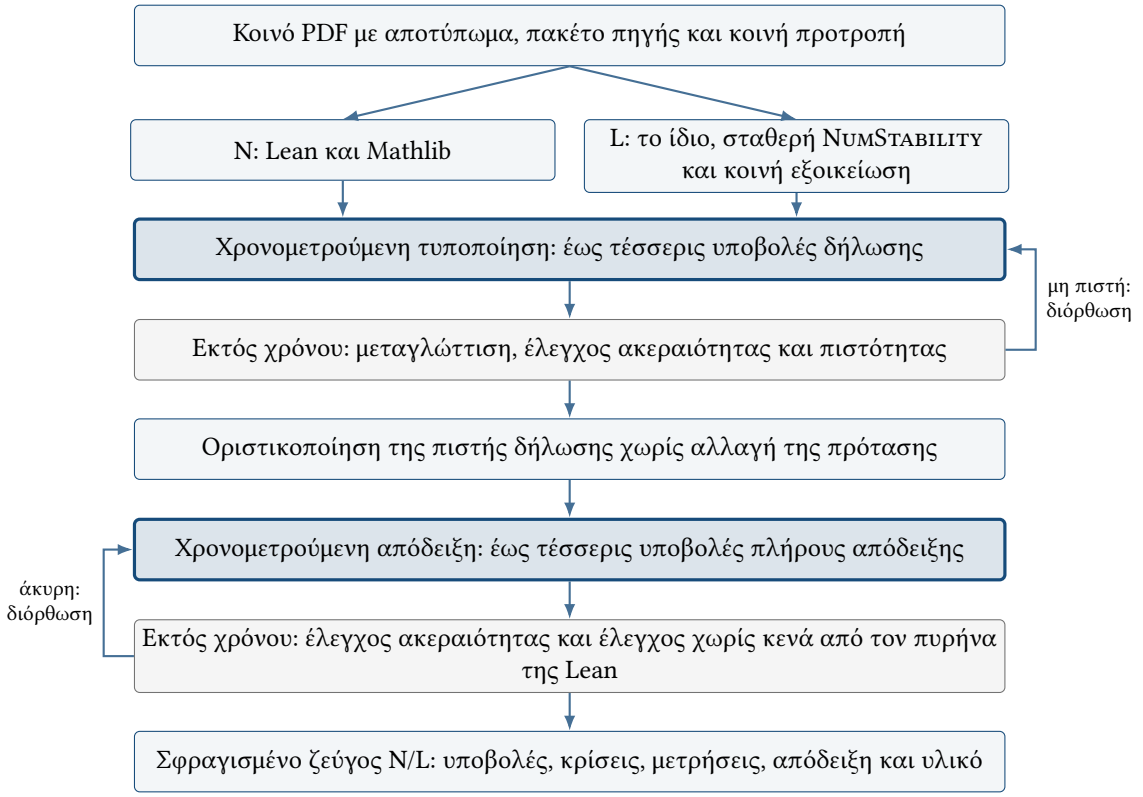
7.3.2 Δύο στάδια και δύο χρονόμετρα

Στο πρώτο στάδιο ο πράκτορας υποβάλλει θεώρημα Lean του οποίου η απόδειξη είναι ένα μόνο `sorry`. Επιτρέπονται έως τέσσερις υποβολές δήλωσης. Κάθε υποβολή καταγράφεται οριστικά και λαμβάνει κρυπτογραφικό αποτύπωμα πριν σταματήσει το χρονόμετρο του πράκτορα· η μεταγλώττιση, ο έλεγχος ακεραιότητας και ο έλεγχος πιστότητας γίνονται εκτός του χρόνου του. Μια μη πιστή δήλωση λαμβάνει στην ίδια συζήτηση ουδέτερη ως προς τη συνθήκη περιγραφή της ασυμφωνίας, χωρίς διορθωμένη δήλωση, απόδειξη ή όνομα δήλωσης της NumStability. Όταν μια δήλωση κριθεί πιστή, οριστικοποιείται, και η ίδια συζήτηση λαμβάνει έως τέσσερις υποβολές για να αποδείξει *ακριβώς* αυτή την πρόταση. Μια απόδειξη με `sorry`, `admit`, αλλαγμένες εξαρτήσεις, πρόσθετα αξιώματα ή μηχανισμούς παράκαμψης του ελέγχου απορρίπτεται· η επιτυχία απαιτεί απόδειξη χωρίς κενά που γίνεται δεκτή από τον πυρήνα. Το Σχήμα 7.2 συνοψίζει τη ροή.

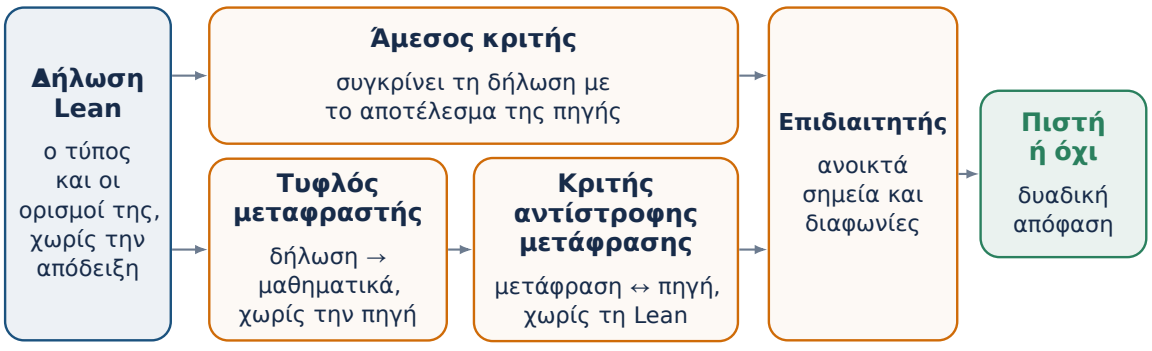
7.3.3 Έλεγχος πιστότητας

Ο έλεγχος πιστότητας προσαρμόζει τη μέθοδο των Meek κ.ά. (2026), δηλαδή τον άμεσο έλεγχο και τον έλεγχο μέσω αντίστροφης μετάφρασης. Η μέθοδος επιλέχθηκε για τρεις λόγους. Αναπτύχθηκε για το ίδιο πρόβλημα, την πιστότητα δηλώσεων Lean που παράγει πράκτορας TN από σύγγραμμα αριθμητικών μεθόδων. Συνδυάζει δύο ανεξάρτητες διαδρομές ελέγχου, ώστε μια απόκλιση που διαφεύγει από τη μία να μπορεί να εντοπιστεί από την άλλη. Τέλος, έχει συγκριθεί με ανθρώπινη κρίση (Ενότητα 4.6). Για κάθε υποβολή κατασκευάζεται μια ψευδωνυμοποιημένη σημασιολογική περιγραφή από τον επεξεργασμένο τύπο του στόχου και από τους ορισμούς από τους οποίους εξαρτάται· η απόδειξη του στόχου δεν περιλαμβάνεται. Η ροή του Σχήματος 7.3 εκτελείται από νέες εκτελέσεις του GPT-6 Astra, ανεξάρτητες από τη συνθήκη, με υψηλό επίπεδο (high) συλλογισμού (OpenAI, 2026c).

[2] *Νέα tokens κάθε αποδεκτής προσπάθειας του HighamBench-Proof.* Τεκμήριο στο αποθετήριο NumStability: `67344c880:paper_bencmark/highambench/reports/full-report-scope30/observed_net_new_tokens.json`.



Σχήμα 7.2: Η ροή μιας εργασίας στο HighamBench-E2E. Τα έντονα πλαίσια χρονομετρούνται ως εργασία του πράκτορα· τα γκρίζα εκτελούνται εκτός του χρόνου του από ανεξάρτητους ελεγκτές.



Σχήμα 7.3: Οι δύο διαδρομές ελέγχου πιστότητας στο HighamBench-E2E. Ο κριτής αντίστροφης μετάφρασης δεν βλέπει τη Lean ούτε την κρίση του άμεσου κριτή. Ο επιδιαιτητής ενεργοποιείται για ανοικτά σημεία ή διαφωνίες. Οι έλεγχοι εκτελούνται εκτός του χρόνου του πράκτορα.

Οι δύο κριτές λαμβάνουν υπόψη κάθε εξάρτηση της περιγραφής, απαντούν με την ίδια σειρά στους 16 σημασιολογικούς ελέγχους του Πίνακα 7.2 και κρίνουν χωριστά κάθε κατεύθυνση της συνεπαγωγής.[3]

S01	Επιλογή του αποτελέσματος της πηγής	S09	Σημασία των νορμών
S02	Μεταβλητές και τύποι	S10	Σταθερές και δείκτες
S03	Εμβέλεια των ποσοδεικτών	S11	Μοντέλο κινητής υποδιαστολής και εξαιρετικές τιμές
S04	Υποθέσεις	S12	Ισχύς της σχέσης
S05	Πληρότητα του συμπεράσματος	S13	Είδος του σφάλματος
S06	Τελεστές και εισαγόμενοι ορισμοί	S14	Όροι ανώτερης τάξης
S07	Ακριβείς έναντι υπολογισμένων ποσοτήτων	S15	Εξειδίκευση ή γενίκευση
S08	Σύνδεση με τον αλγόριθμο	S16	Μη κενότητα

Πίνακας 7.2: Οι 16 σημασιολογικοί έλεγχοι του άμεσου κριτή και του κριτή αντίστροφης μετάφρασης στο HighamBench-E2E, με τη σειρά της προτροπής τους.

Η απόφαση είναι δυαδική. Μια ισοδύναμη δήλωση ή μια γνήσια ενίσχυση γίνεται δεκτή· πρόσθετες υποθέσεις, στένωση του πεδίου, μερική κάλυψη περιπτώσεων ή κενή πρόταση οδηγούν σε απόρριψη. Η αβεβαιότητα δεν γίνεται ποτέ σιωπηρά δεκτή. Οι κρίσεις διατηρούνται στο αρχείο του ζεύγους. Ανθρώπινη επαναβαθμολόγηση των κρίσεων δεν πραγματοποιήθηκε· οι αποφάσεις αποτελούν τεκμηριωμένες κρίσεις πρακτόρων TN.[4]

7.3.4 Μοντέλα, προτροπές και πρόσβαση στη βιβλιοθήκη

Και στις δύο συνθήκες ο πράκτορας τυποποίησης ήταν το GPT-6 Sol με υψηλό (high) επίπεδο συλλογισμού (OpenAI, 2026d), μέσω του Codex CLI 0.156.0. Οι βοηθητικοί πράκτορες ήταν απενεργοποιημένοι. Τα παρατηρήσιμα γεγονότα εργαλείων και γύρων αρχειοθετήθηκαν.

Η κοινή προτροπή ορίζει ότι το PDF είναι η αυθεντική πηγή και ζητά να διατηρηθούν το πλήρες πεδίο, οι ποσοδείκτες, η σειρά των πράξεων, το μοντέλο στρογγυλοποίησης, οι σταθερές και το συμπέρασμα του άρθρου. Προειδοποιεί επίσης για τον ανεξάρτητο έλεγχο πιστότητας και την απόδειξη που ακολουθεί. Η L λαμβάνει επιπλέον ένα σύντομο παράρτημα, που την ενθαρρύνει να αναζητήσει πρώτα την αναπαράσταση του αλγορίθμου και ένα συμβατό μοντέλο κινητής υποδιαστολής και έπειτα λήμματα σφάλματος, διαταραχών και νορμών. Οι πέντε προτροπές παρατίθενται αυτούσιες στο Παράρτημα Γ΄.6.

Πριν από τις εργασίες εκτελέστηκε μία φορά μια συζήτηση εξοικείωσης της L με τη βιβλιοθήκη, χωρίς καμία πληροφορία για το σύνολο εργασιών. Κάθε εργασία της L συνεχίζει από αυτή την αποθηκευμένη συζήτηση, η οποία δεν επαναλήφθηκε ανά εργασία. Η συζήτηση διήρκεσε 183,3 s και κατανάλωσε 88 468 νέα tokens. Το κόστος της προσμετράται μία φορά στα συγκεντρωτικά αποτελέσματα και αναφέρεται χωριστά.[5]

[3] *Αρχειοθετημένη προτροπή του άμεσου κριτή, με τους 16 ελέγχους.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/runs/CAST08-FIXED3/R1/audits/dc50fc3f0795ba6edc8cfa1d1ca8af805b95b4db3c95c8316b829e9906b2c314/roles/direct-judge/try-01/artifacts/prompt.md`.

[4] *Περιγραφή της εκτέλεσης και των ρόλων ελέγχου του HighamBench-E2E.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/protocol/EXECUTION.md`.

[5] *Καταγραφή της συζήτησης εξοικείωσης.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/evidence/warm-root.json`.

7.3.5 Υλικό και λογισμικό

Η εκτέλεση έγινε στον Titan, έναν υπολογιστή με επεξεργαστή Intel Core i9-13900K 13ης γενιάς και λειτουργικό σύστημα Linux (πυρήνας 6.2, x86-64). Ο επεξεργαστής και το λειτουργικό σύστημα καταγράφονται πριν από κάθε υποβολή, μαζί με τα όρια της λωρίδας στην οποία εκτελείται.[6] Χρησιμοποιήθηκαν τρεις ταυτόχρονες και χωριστές λωρίδες των οκτώ λογικών πυρήνων, 24 GiB μνήμης, χωρίς swap και με όριο 512 διεργασιών. Η σειρά των συνθηκών εναλλασσόταν σύμφωνα με προκαθορισμένο πρόγραμμα. Οι δειγματοληπτικά καταγεγραμμένες μέγιστες τιμές ήταν 5,39 πυρήνες και 2,35 GiB για τη N, και 4,70 πυρήνες και 2,01 GiB για την L· σύντομες αιχμές μπορεί να μην καταγράφηκαν. Ο παραλληλισμός μείωσε τη συνολική διάρκεια της εκτέλεσης, όχι τον ενεργό χρόνο κάθε εργασίας.

7.3.6 Μετρήσεις και παρονομαστές

Ο χρόνος τυποποίησης περιλαμβάνει την ενεργό εργασία του πράκτορα μέχρι την πρώτη πιστή δήλωση, μαζί με την αναζήτηση στη βιβλιοθήκη και τις διορθώσεις. Ο χρόνος απόδειξης αρχίζει μετά την οριστικοποίηση της δήλωσης και περιλαμβάνει όλους τους γύρους απόδειξης. Το άθροισμά τους είναι ο ενεργός χρόνος της εργασίας. Η πλήρης χρήση tokens του παρόχου καταγράφεται χωριστά από τα νέα tokens (Ενότητα 7.4). Οι γραμμές κώδικα δεν περιλαμβάνουν κενές γραμμές και γραμμές μόνο με σχόλια.[7]

7.4 Μέτρηση της χρήσης της NumStability

Η πραγματική χρήση της βιβλιοθήκης μετράται στους όρους που επεξεργάστηκε η Lean και όχι στο κείμενο. Το κύριο μέτρο είναι το πλήθος των *διακριτών* δηλώσεων της NumStability που εμφανίζονται άμεσα στο αποδεκτό τεχνούργημα της L. Εφαρμόζεται σε τρία επίπεδα:

- **DS** (διατύπωση): δηλώσεις στον επεξεργασμένο τύπο του στόχου και στους τοπικούς ορισμούς της διατύπωσης·
- **DP** (απόδειξη): δηλώσεις στον όρο απόδειξης που δέχτηκε ο πυρήνας και στα σώματα των τοπικών βοηθητικών δηλώσεων της απόδειξης·
- **DU** (ένωση): η ένωση των DS και DP, όπου ένα όνομα που εμφανίζεται και στις δύο μετρά μία φορά.

Ως διακριτές μετρώνται οι μοναδικές πλήρως επιλυμένες δηλώσεις: η επανειλημμένη χρήση του ίδιου λήμματος μετρά μία φορά. Οι απλές εισαγωγές υπομονάδων δεν μετρούν. Η ίδια μετρική DP εφαρμόστηκε και στις δύο αξιολογήσεις.

Δευτερεύοντα μέτρα χρήσης της NumStability παρατίθενται στα Παραρτήματα Β΄ και Γ΄.

Και στις δύο αξιολογήσεις τα tokens μετρώνται ως νέα tokens, δηλαδή ως η είσοδος μείον την

[6] *Καταγραφή του υλικού πριν από μια υποβολή δήλωσης του HighamBench-E2E.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/runs/CAST08-FIXED3/R1/submissions/01/hardware-before.json`.

[7] *Ορισμοί των μετρήσεων του HighamBench-E2E.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/protocol/METRICS.md`.

αποθηκευμένη είσοδο, συν την έξοδο. Για κάθε μέγεθος M, το κέρδος της L είναι

$$G(M) = 100\,\frac{M_{\mathrm{N}} - M_{\mathrm{L}}}{M_{\mathrm{N}}}\ \%,$$

οπότε θετική τιμή σημαίνει μικρότερο κόστος για την L. Τα συγκεντρωτικά κέρδη υπολογίζονται από λόγους αθροισμάτων, όχι ως μέσοι όροι ποσοστών ανά εργασία. Στο HighamBench-Proof υπολογίζονται πρώτα οι μέσοι όροι των τριών επαναλήψεων κάθε συνθήκης και έπειτα το κέρδος της εργασίας. Η σχέση χρήσης και κέρδους περιγράφεται με τον συντελεστή ρ του Spearman, με μέσες τάξεις για τις ισοβαθμίες.

Κεφάλαιο 8

HighamBench: αποτελέσματα

Το κεφάλαιο παρουσιάζει χωριστά τα αποτελέσματα των δύο αξιολογήσεων του Κεφαλαίου 7: πρώτα του HighamBench-Proof, με τους 30 σταθερούς στόχους, και έπειτα του HighamBench-E2E, με τις δέκα εργασίες τυποποίησης και απόδειξης.

8.1 HighamBench-Proof: απόδειξη σταθερών στόχων

8.1.1 Συνολικά αποτελέσματα

Και οι 180 προσπάθειες της ανάλυσης, δηλαδή 90 ζεύγη N/L σε 30 εργασίες, κατέληξαν σε αποδεκτή απόδειξη. Η σύγκριση αφορά επομένως το κόστος και όχι την επιτυχία. Ο Πίνακας 8.1 δίνει τους μέσους όρους των εργασιών. Η L ήταν ταχύτερη σε 23 από τις 30 εργασίες και υπέβαλε λιγότερες γραμμές κώδικα σε 19. Ο χρόνος μειώθηκε κατά 35,0% και οι γραμμές κώδικα κατά 18,3%. Τα νέα tokens δείχνουν προς την ίδια κατεύθυνση: η L χρησιμοποίησε 11,8% λιγότερα συνολικά και λιγότερα σε 20 από τις 30 εργασίες.

Μέγεθος (μέσος όρος εργασίας)	N	L	Κέρδος	L καλύτερη
Χρόνος μέχρι την υποβολή (s)	792,2	515,1	+35,0%	23/30
Νέα tokens (χιλιάδες)	109,5	96,6	+11,8%	20/30
Γραμμές υποβληθέντος κώδικα	398,1	325,1	+18,3%	19/30

Πίνακας 8.1: HighamBench-Proof: μέσοι όροι των 30 εργασιών και κέρδος από τον λόγο των συνολικών τιμών. Κάθε τιμή εργασίας είναι ο μέσος όρος τριών αποδεκτών επαναλήψεων.

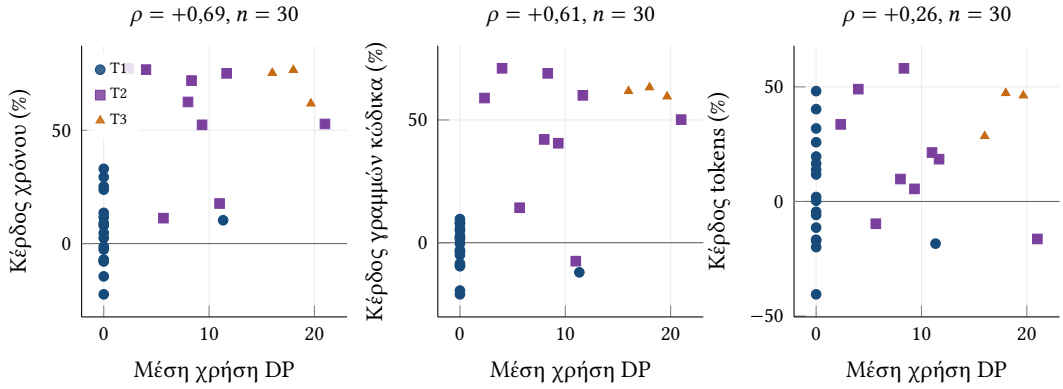
Οι γραμμές κώδικα του πλήρους αρχείου είναι το κύριο μέτρο μεγέθους. Αν μετρηθεί μόνο η περιοχή της απόδειξης του στόχου, οι μέσοι όροι είναι 94,0 γραμμές για τη N και 102,5 για την L, δηλαδή κέρδος −9,0%. Η διαφορά δεν αντιφάσκει με το θετικό αποτέλεσμα του πλήρους αρχείου: τα βοηθητικά λήμματα πριν από τον στόχο εξαιρούνται από την περιοχή της απόδειξης, ενώ συχνά περιέχουν μεγάλο μέρος της πραγματικής εργασίας. Η μέτρηση αυτή και οι ιστορικές φυσικές γραμμές δίνονται ως διαγνωστικά στο Παράρτημα Β′.

Τα αναλυτικά αποτελέσματα ανά εργασία παρατίθενται στους Πίνακες Β′.2 και Β′.3 του Παραρτήματος Β′. Η `P09-T1` ξεχωρίζει ως η ακριβότερη εργασία σε χρόνο και tokens. Τα μεγαλύτερα

κέρδη χρόνου εμφανίζονται στις κατηγορίες T2 και T3, ενώ και οι επτά εργασίες στις οποίες η L ήταν βραδύτερη ανήκουν στην T1.

8.1.2 Χρήση της βιβλιοθήκης και κέρδη

Η άμεση χρήση της βιβλιοθήκης στην απόδειξη (DP) έχει θετική συσχέτιση με όλα τα μετρούμενα κέρδη (Σχήμα 8.1): $\rho = +0{,}69$ με το κέρδος χρόνου, $\rho = +0{,}61$ με το κέρδος γραμμών κώδικα και ασθενέστερη, $\rho = +0{,}26$, με το κέρδος των νέων tokens. Κάθε σημείο είναι μία εργασία: η οριζόντια συντεταγμένη είναι ο μέσος όρος των τριών αποδεκτών αποδείξεων της L και η κατακόρυφη το κέρδος των μέσων τιμών. Το σχήμα είναι συμβατό με την ερμηνεία ότι η μεγαλύτερη χρήση συνοδεύει μεγαλύτερα κέρδη χρόνου και κώδικα στο συγκεκριμένο δείγμα.



Σχήμα 8.1: HighamBench-Proof: μέση άμεση χρήση της βιβλιοθήκης στην απόδειξη (DP) έναντι των κερδών της L. Τα χρώματα και τα σύμβολα δηλώνουν την αναμενόμενη κατηγορία και όχι μετρημένη χρήση.

Ο Πίνακας 8.2 συγκρίνει τις αναμενόμενες κατηγορίες με την παρατηρούμενη χρήση. Οι τρεις εργασίες T3 χρησιμοποιούν εκτενώς τη βιβλιοθήκη, οι εργασίες T2 πιο άνισα, ενώ 17 από τις 18 εργασίες T1 δεν έχουν καμία άμεση χρήση. Συνολικά, 17 από τις 30 εργασίες δεν χρησιμοποιούν καθόλου τη βιβλιοθήκη στην τελική απόδειξη. Η σειρά των κατηγοριών είναι επομένως συνεπής με τη χρήση σε συγκεντρωτικό επίπεδο, χωρίς να επικυρώνει κάθε επιμέρους χαρακτηρισμό. Εξαιρέσεις είναι η `P15-T1`, που χρησιμοποιεί τη βιβλιοθήκη αν και ανήκει στην T1, και οι `P04-T2` και `P14-T2`, που τη χρησιμοποιούν αλλά έχουν μικρό κέρδος χρόνου (+11,2% και +17,7%). Στις εργασίες όπου και οι τρεις αποδείξεις της L χρησιμοποιούν τη βιβλιοθήκη, το μέσο κέρδος χρόνου είναι +55,8%, ενώ στις εργασίες χωρίς καμία χρήση είναι +5,7% (μέσοι όροι των κερδών ανά εργασία).

Αναμενόμενη κατηγορία	Εργασίες	Μέση χρήση DP	Εργασίες με DP = 0
T1	18	0,6	17
T2	9	9,0	0
T3	3	17,9	0

Πίνακας 8.2: HighamBench-Proof: αναμενόμενη βοήθεια έναντι παρατηρούμενης άμεσης χρήσης της βιβλιοθήκης στην απόδειξη.

8.2 HighamBench-E2E: τυποποίηση και απόδειξη

8.2.1 Έκβαση των εργασιών

Και τα δέκα ζεύγη κατέληξαν σε δήλωση που κρίθηκε πιστή και σε απόδειξη χωρίς κενά που έγινε δεκτή από τον πυρήνα, και στις δύο συνθήκες.[1]

8.2.2 Συνολικά αποτελέσματα

Ο Πίνακας 8.3 δίνει τα αθροίσματα των δέκα εργασιών. Η L χρειάστηκε 22,5% λιγότερο ενεργό χρόνο (5 410 έναντι 6 979 s) και έγραψε 19,7% λιγότερες γραμμές κώδικα απόδειξης (2 537 έναντι 3 158), αλλά χρησιμοποίησε 34,7% περισσότερα νέα tokens. Η διαφορά χρόνου προέρχεται κυρίως από το στάδιο της απόδειξης: η απόδειξη ήταν 29,9% ταχύτερη, ενώ η τυποποίηση 7,3% βραδύτερη. Ο μέσος ενεργός χρόνος ανά εργασία ήταν 697,9 s για τη N και 541,0 s για την L, ή 559,3 s αν μοιραστεί σε όλες τις εργασίες ο χρόνος της εξοικείωσης. Αν το κόστος της εξοικείωσης προσμετρηθεί μία φορά, το κέρδος χρόνου μειώνεται σε 19,9% και το πλεόνασμα των tokens αυξάνεται σε 41,7%.

Μέγεθος (άθροισμα δέκα εργασιών)	N	L	L/N	Κέρδος
Χρόνος τυποποίησης (s)	1 389,1	1 490,2	1,073	−7,3%
Χρόνος απόδειξης (s)	5 589,8	3 920,0	0,701	+29,9%
Συνολικός ενεργός χρόνος (s)	6 978,9	5 410,2	0,775	+22,5%
Συνολικός χρόνος με την εξοικείωση (s)	6 978,9	5 593,5	0,801	+19,9%
Νέα tokens τυποποίησης	399 748	980 733	2,453	−145,3%
Νέα tokens απόδειξης	873 571	734 899	0,841	+15,9%
Συνολικά νέα tokens	1 273 319	1 715 632	1,347	−34,7%
Συνολικά νέα tokens με την εξοικείωση	1 273 319	1 804 100	1,417	−41,7%
Γραμμές κώδικα δήλωσης	333	337	1,012	−1,2%
Γραμμές κώδικα απόδειξης	3 158	2 537	0,803	+19,7%

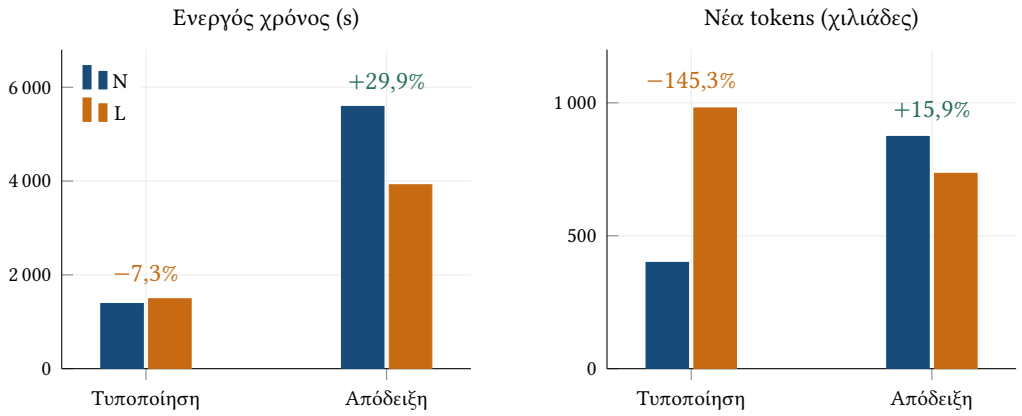
Πίνακας 8.3: HighamBench-E2E: αθροίσματα των δέκα εργασιών ανά στάδιο και συνολικά. Το κόστος της εξοικείωσης προσμετράται μία φορά, μόνο στις δύο αντίστοιχες συνολικές γραμμές.

8.2.3 Τυποποίηση και απόδειξη

Το Σχήμα 8.2 απομονώνει τη βασική διαφορά των σταδίων: το όφελος εμφανίζεται στην απόδειξη, ενώ η τυποποίηση κοστίζει περισσότερο, ιδιαίτερα σε tokens. Η L ήταν συνολικά ταχύτερη σε οκτώ από τις δέκα εργασίες και έγραψε μικρότερη απόδειξη σε έξι. Οι τιμές ανά εργασία δίνονται στο Παράρτημα Γ΄.

Η εξοικονόμηση tokens στην απόδειξη δεν αντισταθμίζει την αύξησή τους κατά την τυποποίηση. Για το πλεόνασμα της τυποποίησης, τα αρχεία δείχνουν επιθεώρηση της βιβλιοθήκης σε αυτό το στάδιο, αλλά δεν χωρίζουν τα tokens εισόδου σε αποτελέσματα αναζήτησης, κληρονομημένη εξοικείωση, προτροπές και άλλο περιεχόμενο· γι’ αυτό η διαφορά δεν μπορεί να αποδοθεί ειδικά στην αναζήτηση. Αντίθετα, ο μικρότερος χρόνος, τα λιγότερα tokens και ο μικρότερος κώδικας

[1] *Σφραγισμένη καταγραφή των δέκα ζευγών.* Τεκμήριο στο αποθετήριο NumStability: `27c2ed34c:benchmark/results/source_task_ledger.json`.



Σχήμα 8.2: HighamBench-E2E: κόστος τυποποίησης και απόδειξης στο σύνολο των δέκα εργασιών, χωρίς την εφάπαξ εξοικείωση. Τα ποσοστά είναι τα κέρδη της L: θετικό σημαίνει μικρότερο κόστος. Η ταχύτερη και φθηνότερη απόδειξη συνυπάρχει με ακριβότερη τυποποίηση.

της απόδειξης είναι συμβατά με χρήσιμες δηλώσεις που μειώνουν την εργασία, όταν υπάρχουν ήδη πιστή δήλωση και κατάλληλη διεπαφή.

8.2.4 Χρήση της βιβλιοθήκης

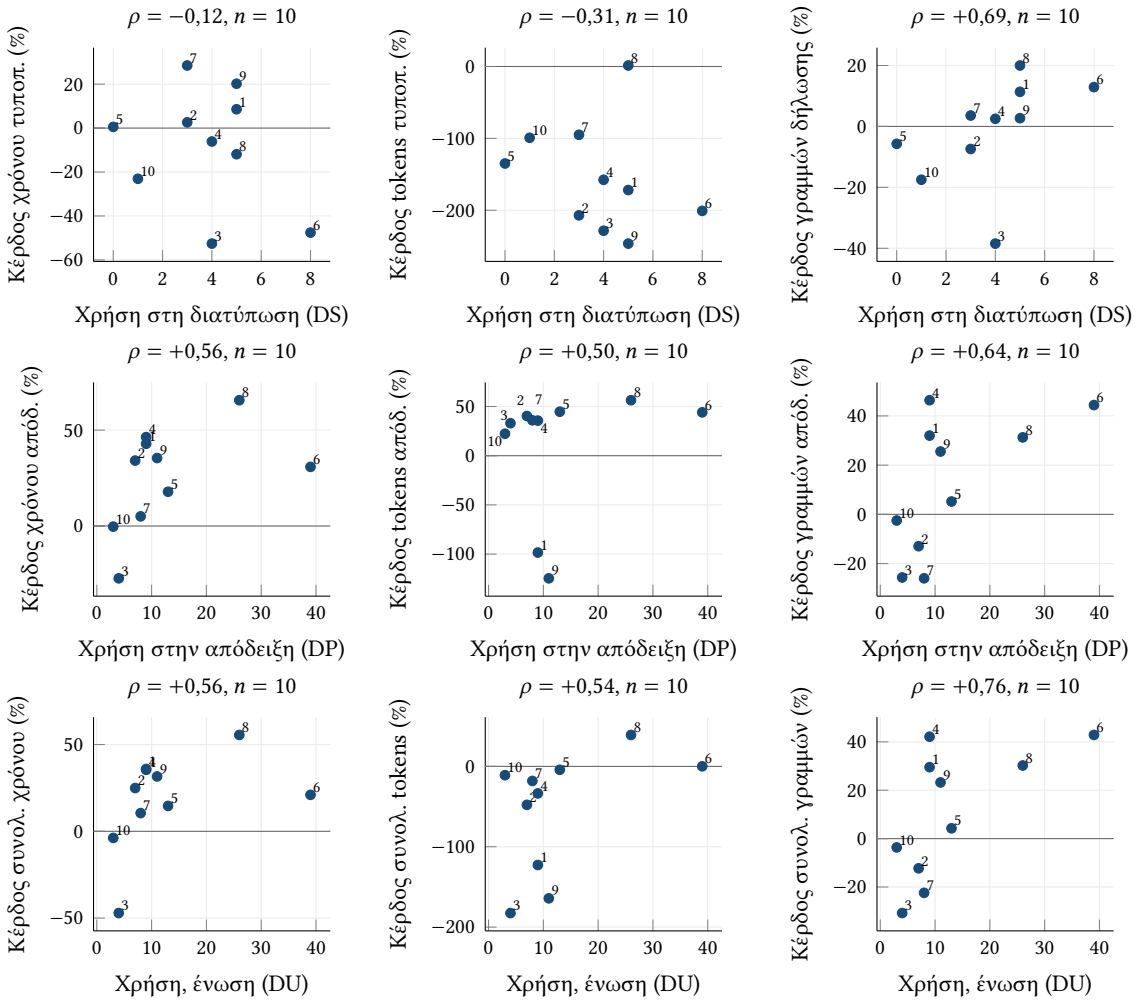
Σε εννέα από τις δέκα εργασίες η δήλωση της L χρησιμοποιεί ουσιαστικά δηλώσεις της NumStability, και σε όλες τις δέκα η απόδειξη χρησιμοποιεί άμεσα δηλώσεις της βιβλιοθήκης. Σε αυτό το σύνολο κάθε όνομα της διατύπωσης εμφανίζεται και στον όρο απόδειξης, οπότε οι τιμές DU ταυτίζονται με τις DP· η ταύτιση είναι ιδιότητα των δέκα αρχείων και όχι γενική ισοδυναμία. Ο Πίνακας 8.4 συνδέει κάθε μέτρο χρήσης με το στάδιο που του αντιστοιχεί.

Στάδιο / μέτρο χρήσης	Χρόνος ρ	Tokens ρ	Γραμμές ρ
Τυποποίηση / δηλώσεις της διατύπωσης (DS)	−0,12	−0,31	+0,69
Απόδειξη / δηλώσεις του όρου απόδειξης (DP)	+0,56	+0,50	+0,64
Σύνολο / ένωση διατύπωσης και απόδειξης (DU)	+0,56	+0,54	+0,76

Πίνακας 8.4: HighamBench-E2E: συντελεστές Spearman ανά στάδιο, για τις δέκα εργασίες.

Η διάκριση των σταδίων αναδεικνύει μια ασυμμετρία. Η χρήση στη διατύπωση σχετίζεται με μικρότερες δηλώσεις ($\rho = +0{,}69$), αλλά όχι με ταχύτερη τυποποίηση ($\rho = -0{,}12$) ή λιγότερα tokens ($\rho = -0{,}31$). Η χρήση στον όρο απόδειξης σχετίζεται με κέρδη χρόνου ($\rho = +0{,}56$) και γραμμών ($\rho = +0{,}64$) της απόδειξης. Η ένωση σχετίζεται με το συνολικό κέρδος χρόνου ($\rho = +0{,}56$), γραμμών ($\rho = +0{,}76$) και tokens ($\rho = +0{,}54$). Το Σχήμα 8.3 δείχνει τα αντίστοιχα σημεία· οι αριθμοί αντιστοιχούν στη σειρά των εργασιών του Πίνακα Γ΄.5.

Το κέρδος χρόνου της τυποποίησης σχετίζεται ασθενώς με εκείνο της απόδειξης ($\rho = +0{,}27$), και το κέρδος χρόνου ασθενώς με το κέρδος tokens ($\rho = +0{,}15$). Η ταχύτερη εργασία δεν είναι επομένως απαραίτητα και η φθηνότερη σε tokens.



Σχήμα 8.3: HighamBench-E2E: διακριτές δηλώσεις της βιβλιοθήκης στην αποδεκτή διατύπωση (DS, πάνω σειρά), στον όρο απόδειξης και στα τοπικά βοηθητικά λήμματα (DP, μεσαία σειρά) και στην ένωσή τους (DU, κάτω σειρά), έναντι των κερδών της τυποποίησης, της απόδειξης και του συνόλου αντίστοιχα. Σε αυτό το σύνολο DU = DP σε κάθε εργασία.

Κεφάλαιο 9

Συμπεράσματα και μελλοντική έρευνα

Η εργασία καταλήγει σε τρία βασικά αποτελέσματα: μια επαναχρησιμοποιήσιμη βιβλιοθήκη αριθμητικής σταθερότητας, μια διαδικασία κατασκευής της με πράκτορες TN και ανθρώπινες αποφάσεις ενσωματωμένες στα skills, και δύο αξιολογήσεις της χρησιμότητάς της. Η συνθετική δομή και οι έλεγχοι διεπαφών του Κεφαλαίου 5 τεκμηριώνουν την εσωτερική οργάνωση της NumStability· οι αξιολογήσεις εξετάζουν πότε αυτή η υποδομή μειώνει το κόστος νέων αποδείξεων. Μαζί αποτελούν το πρώτο στάδιο του ευρύτερου συστήματος τυποποίησης βιβλίων της Ενότητας 9.3.

9.1 Τι έδειξαν οι δύο αξιολογήσεις

Οι δύο αξιολογήσεις δεν συνενώνονται, μπορούν όμως να συγκριθούν ως προς την κατεύθυνση των αποτελεσμάτων. Τέσσερα ευρήματα συνοψίζουν αυτή τη σύγκριση:

1. **Δεν παρατηρείται όφελος στο κόστος τυποποίησης.** Στο HighamBench-E2E, η τυποποίηση με τη βιβλιοθήκη χρειάστηκε 7,3% περισσότερο χρόνο και περισσότερα νέα tokens. Παρότι η μεγαλύτερη χρήση συνδέεται με συντομότερες δηλώσεις, δεν συνοδεύεται από οικονομία χρόνου ή tokens σε αυτό το στάδιο.

2. **Τα tokens έχουν θετική, αλλά ασθενέστερη σχέση με τη χρήση.** Η άμεση χρήση στην απόδειξη συσχετίζεται θετικά με το κέρδος νέων tokens ($\rho = +0{,}26$ στο HighamBench-Proof και $\rho = +0{,}50$ στο στάδιο απόδειξης του HighamBench-E2E), ασθενέστερα όμως από ό,τι με τα κέρδη χρόνου και κώδικα. Τα tokens πρέπει επομένως να εξετάζονται ως χωριστή διάσταση του κόστους και όχι ως ισοδύναμο μέτρο του χρόνου.

3. **Η πραγματική χρήση συνοδεύεται από μεγάλα κέρδη στην απόδειξη.** Στο HighamBench-Proof, το μέσο κέρδος χρόνου ανά εργασία είναι 55,8% στις 12 εργασίες που χρησιμοποιούν τη βιβλιοθήκη και στις τρεις επαναλήψεις, έναντι 5,7% στις 17 χωρίς καμία χρήση. Η άμεση χρήση συσχετίζεται με το κέρδος χρόνου απόδειξης με συντελεστή Spearman $\rho = +0{,}69$ στο HighamBench-Proof και $\rho = +0{,}56$ στο HighamBench-E2E. Θετική είναι και η σχέση με τη μείωση του κώδικα (Κεφάλαιο 8). Το εύρημα αναδεικνύει την αξία της επαναχρησιμοποίησης όταν υπάρχουν κατάλληλες δηλώσεις.

4. **Η αξιοποιήσιμη κάλυψη παραμένει περιορισμένη.** Στο HighamBench-Proof, 17 από τις 30 εργασίες δεν χρησιμοποιούν άμεσα τη βιβλιοθήκη στην απόδειξη, περιορίζοντας την εμβέλεια του παρατηρούμενου οφέλους. Η μη χρήση δεν διακρίνει πάντως ανάμεσα σε έλλειψη κατάλληλων αποτελεσμάτων, δυσκολία εύρεσής τους και ασυμβατότητα αναπαράστασης. Η χρήση σε όλες τις εργασίες του HighamBench-E2E αφορά σύνολο επιλεγμένο με κριτήριο την αναμενόμενη κάλυψη και δεν αναιρεί αυτόν τον περιορισμό.

9.2 Επόμενα βήματα

Από τα ευρήματα προκύπτουν τέσσερις συγκεκριμένες κατευθύνσεις:

1. **Επέκταση της κάλυψης.** Προσθήκη αποτελεσμάτων πέρα από τον Higham με την αξιοποίηση νέων άρθρων, ώστε να αυξηθούν οι εργασίες στις οποίες μπορεί να χρησιμοποιηθεί η βιβλιοθήκη.
2. **Αναδιοργάνωση βάσει της αξιολόγησης.** Υλοποίηση του βρόχου που αξιοποιεί τις μετρικές του HighamBench για τη βελτίωση της δομής της βιβλιοθήκης (Ενότητα 9.3). Πρώτοι στόχοι είναι η ευκολότερη εύρεση δηλώσεων και η ενίσχυση της αρθρωτής οργάνωσης της βιβλιοθήκης.
3. **Μείωση του κόστους τυποποίησης.** Διερεύνηση τρόπων μείωσης του κόστους όταν χρησιμοποιείται η βιβλιοθήκη, πιθανότατα με πρόσθετη εκπαίδευση μοντέλου ανοικτού κώδικα για τη συγκεκριμένη εργασία, όταν αποκτηθούν οι απαραίτητοι πόροι.
4. **Ευρύτερη αξιολόγηση.** Προκαθορισμένο σύνολο εργασιών, ενιαίο πρωτόκολλο, περισσότερες εργασίες και επαναλήψεις ανά εργασία, με έλεγχο πιστότητας σε όλες τις εργασίες, ώστε να εξεταστεί αν η ευρύτερη κάλυψη μεταφράζεται σε αξιόπιστα κέρδη και αν η σχέση χρήσης και οφέλους γενικεύεται.

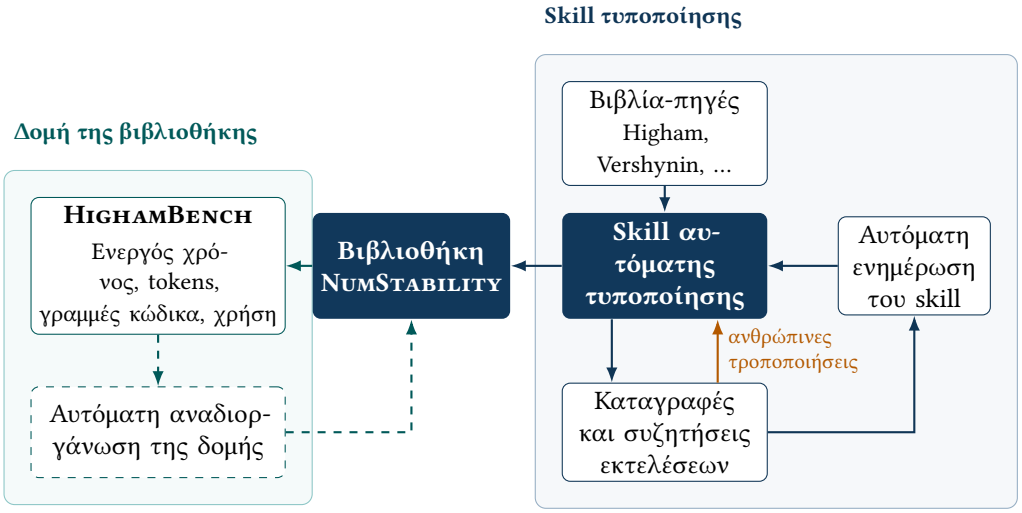
9.3 Απώτερος σκοπός: ένα σύστημα τυποποίησης βιβλίων

Ο απώτερος σκοπός του έργου είναι ένα γενικό σύστημα τυποποίησης βιβλίων με πράκτορες TN. Η NumStability και η διαδικασία κατασκευής της δεν προορίζονται μόνο για το βιβλίο του Higham· αποτελούν το πρώτο στάδιο αυτού του συστήματος, το οποίο οργανώνεται σε δύο βρόχους γύρω από την ίδια βιβλιοθήκη (Σχήμα 9.1).

Ο πρώτος βρόχος αφορά το skill τυποποίησης. Ένα skill αυτόματης τυποποίησης μετατρέπει επιλεγμένο υλικό ενός βιβλίου σε δηλώσεις και αποδείξεις της βιβλιοθήκης. Τα αρχεία καταγραφής και οι συζητήσεις των εκτελέσεων (logs και transcripts) τροφοδοτούν ένα δεύτερο skill, που ενημερώνει αυτόματα το skill τυποποίησης. Ο άνθρωπος παρεμβαίνει με στοχευμένες τροποποιήσεις όπου οι καταγραφές δείχνουν συστηματικό πρόβλημα. Ο βρόχος αυτός έχει ήδη υλοποιηθεί σε έναν βαθμό και εφαρμόζεται.

Ο δεύτερος βρόχος αφορά τη δομή της βιβλιοθήκης. Το HighamBench μετρά τον ενεργό χρόνο των εργασιών, τα tokens, τις γραμμές κώδικα και την πραγματική χρήση δηλώσεων. Η σχεδιαζόμενη συνέχεια είναι ένα skill που αξιοποιεί αυτά τα αποτελέσματα για να αναδιοργανώνει

τη βιβλιοθήκη, διευκολύνοντας την εύρεση και τη σύνθεση των δηλώσεων. Ο βρόχος αυτός παραμένει μελλοντικό βήμα.



Σχήμα 9.1: Δύο βρόχοι γύρω από την ίδια βιβλιοθήκη. Δεξιά, οι εκτελέσεις παράγουν καταγραφές που τροφοδοτούν αυτόματες και ανθρώπινες ενημερώσεις του skill. Αριστερά, η αξιολόγηση παρέχει μετρικές για τη σχεδιαζόμενη αναδιοργάνωση. Συνεχή περιγράμματα και βέλη δηλώνουν λειτουργίες που ήδη εφαρμόζονται· διακεκομμένα, την προβλεπόμενη αυτόματη αναδιοργάνωση.

9.4 Επέκταση σε επόμενα βιβλία

Η τυποποίηση του *High-Dimensional Probability* του Vershynin (2018) έχει ήδη αρχίσει και οι καταγραφές των εκτελέσεων τροφοδοτούν την αναθεώρηση των οδηγιών, εφαρμόζοντας τον δεξιό βρόχο του Σχήματος 9.1. Το βιβλίο αναπτύσσει ανισότητες συγκέντρωσης και θεωρία τυχαίων διανυσμάτων και πινάκων· η επιλογή του συνδέεται με την προοπτική συνδυασμού πιθανοτικών φραγμάτων και αριθμητικής σταθερότητας για επαληθευμένους *τυχαιοποιημένους αριθμητικούς αλγορίθμους*.

Πέρα από τον Higham και τον Vershynin, ο κατάλογος του ευρύτερου έργου περιλαμβάνει τους τίτλους του Πίνακα 9.1, για επέκταση σε άλλες περιοχές των υπολογιστικών μαθηματικών. Πρόκειται για πιθανό υλικό για μελλοντική τυποποίηση, όχι για παραδοτέα της παρούσας διπλωματικής.

Πεδίο	Βιβλία και συμπληρωματικό υλικό
Ανάλυση πινάκων	Bhatia, *Matrix Analysis*
Επαναληπτικές μέθοδοι για γραμμικά συστήματα	Greenbaum, *Iterative Methods for Solving Linear Systems*· Saad, *Iterative Methods for Sparse Linear Systems*
Βελτιστοποίηση	Boyd και Vandenberghe, *Convex Optimization*· Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*· Nocedal και Wright, *Numerical Optimization*
Αριθμητικές μέθοδοι, διαφορικές εξισώσεις και ρευστοδυναμική	Hoffman, *Numerical Methods for Engineers and Scientists*· LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations* και *Finite Volume Methods for Hyperbolic Problems*· Sod, *Numerical Methods in Fluid Dynamics: Initial and Initial Boundary-Value Problems*
Μέθοδοι lattice Boltzmann	Krüger κ.ά., *The Lattice Boltzmann Method: Principles and Practice*· Succi, *The Lattice Boltzmann Equation for Complex States of Flowing Matter*
Ευρωστία γεωμετρικών υπολογισμών	Kettner κ.ά., *Classroom Examples of Robustness Problems in Geometric Computations* (άρθρο)

Πίνακας 9.1: Οι δώδεκα επιπλέον τίτλοι βιβλίων του καταλόγου και ένα συμπληρωματικό άρθρο, ομαδοποιημένα κατά πεδίο. Η καταγραφή αφορά την επέκταση του συστήματος, όχι την ολοκλήρωση της τυποποίησής τους.

Εν κατακλείδι, στόχος είναι η τυποποιημένη γνώση να λειτουργεί ως αφετηρία για νέα μαθηματική εργασία. Υπό αυτή την οπτική, η μακροπρόθεσμη αξία της NumStability θα κριθεί από τις επόμενες αποδείξεις και εφαρμογές που θα στηριχθούν σε αυτήν.

Παράρτημα Α΄

Δήλωση χρήσης εργαλείων τεχνητής νοημοσύνης

Η ανάπτυξη της βιβλιοθήκης και η προετοιμασία της παρούσας εργασίας υποστηρίχθηκαν από πράκτορες TN, κυρίως το Codex της OpenAI και το Claude (Claude Code) της Anthropic. Οι πράκτορες χρησιμοποιήθηκαν για αναζήτηση στο αποθετήριο, σύνθεση προσχεδίων, παραγωγή κώδικα Lean και LaTeX, εκτέλεση μεταγλωττίσεων και ελέγχους συνέπειας. Η έξοδος των μοντέλων δεν αντιμετωπίστηκε ως αυθεντία: οι τυπικοί ισχυρισμοί έγιναν αποδεκτοί μόνο όταν ελέγχθηκαν από τη Lean στο καθορισμένο περιβάλλον. Οι έλεγχοι αντιστοιχίας της βιβλιοθήκης προς τις πηγές έγιναν υπό την καθοδήγηση του Codex ή μηχανικά και δεν αποτελούν συστηματική ανθρώπινη επανεξέταση. Στις δύο αξιολογήσεις του HighamBench (Κεφάλαιο 7) οι πράκτορες ήταν το GPT-5.6 Sol (HighamBench-Proof) και το GPT-6 Sol (HighamBench-E2E) μέσω του Codex, ενώ οι κριτές πιστότητας του HighamBench-E2E ήταν ανεξάρτητες εκτελέσεις του GPT-6 Astra. Οι κρίσεις πιστότητας του HighamBench-E2E δεν επαναβαθμολογήθηκαν από άνθρωπο (Ενότητα 7.3.3). Η κατασκευή των εργασιών του HighamBench-Proof ανατέθηκε επίσης, σχεδόν εξ ολοκλήρου, στο Codex. Η ευθύνη για τις τελικές επιστημονικές διατυπώσεις και την ερμηνεία των αποτελεσμάτων δεν μεταβιβάζεται στα εργαλεία αυτά.

Παράρτημα Β΄

HighamBench-Proof: συμπληρωματικά στοιχεία

Το παράρτημα συμπληρώνει την Ενότητα 8.1 με τον κατάλογο εργασιών και πηγών, τα αναλυτικά αποτελέσματα, τα μέτρα χρήσης της βιβλιοθήκης, τα δευτερεύοντα διαγνωστικά, το ευρετήριο τεκμηρίων και τις αρχειοθετημένες προτροπές. Κανένα από τα δευτερεύοντα μέτρα δεν αντικαθιστά το κύριο μέτρο DP.

Β΄.1 Κατάλογος εργασιών και πηγών

Ο Πίνακας Β΄.1 περιγράφει τις 30 εργασίες και προσδιορίζει το αποτέλεσμα προέλευσης σε καθένα από τα 25 άρθρα. Κάτω από τον κωδικό της διπλωματικής δίνεται σε παρένθεση ο αρχικός κωδικός του δημόσιου αποθετηρίου: οι κατηγορίες T1/T2/T3 της διπλωματικής αντιστοιχούν στις T3/T1/T2 του αρχείου. Οι περιγραφές αφορούν το συγκεκριμένο ζητούμενο κάθε εργασίας, που μπορεί να είναι μέρος μιας απόδειξης ή ειδίκευση ενός αποτελέσματος, όχι κατ' ανάγκη ολόκληρο το θεώρημα της πηγής. Οι σελίδες είναι οι τυπωμένες σελίδες της αντίστοιχης έκδοσης, όχι η αρίθμηση του αναγνώστη PDF.

Πίνακας Β΄.1: HighamBench-Proof: περιγραφή των εργασιών και ακριβής θέση στις πηγές.

Εργασία (αρχικός κωδικός)	Αντικείμενο	Πηγή και θέση
`P01-T2` (`P01-T1`)	Forward φράγμα της pairwise summation 2^r μη αρνητικών όρων, με συντελεστή γ_r.	N. J. Higham (1993). Εξ. (3.6), σ. 788· μη αρνητικοί όροι μετά την (2.6), σ. 785.
`P01-T3` (`P01-T2`)	Τοπικό φράγμα συσσωρευμένου σφάλματος και φράγματα γ_{n-1}, $\gamma_{\log_2 n}$ για αναδρομική και pairwise summation των ίδιων δεδομένων.	N. J. Higham (1993). Εξ. (2.6), ανάλυση πριν την (2.7), σ. 785· (3.6), σ. 788.

Συνέχεια στην επόμενη σελίδα

Εργασία (αρχικός κωδικός)	Αντικείμενο	Πηγή και θέση
`P03-T1` (`P03-T3`)	Αναδρομικό normwise φράγμα του residual στην επαναληπτική βελτίωση τριών ακριβειών, υπό τις υποθέσεις του επιλυτή διόρθωσης.	Carson και Higham (2018). Θεώρημα 4.1, σ. A825· ταυτότητα (4.1), σ. A824.
`P04-T2` (`P04-T1`)	Forward φράγμα εσωτερικού γινομένου με αλυσιδωτές block FMA μικτής ακρίβειας, διαταραχές ανά όρο και ειδικεύσεις σειράς αξιολόγησης.	Blanchard κ.ά. (2020). §3.2, εξ. (3.1)–(3.4) και επόμενη παράγραφος, σ. C129.
`P05-T1` (`P05-T3`)	Βελτιωμένο componentwise backward φράγμα επιτυχούς Cholesky, μαζί με τις τοπικές εκτιμήσεις σφάλματος.	Rump και Jeannerod (2014). Θεώρημα 4.4, σ. 694–695· εξ. (4.4), (4.5a)–(4.5b).
`P06-T1` (`P06-T3`)	Ανάπτυγμα πρώτης τάξης ακολουθίας διαταραγμένων μετασχηματισμών Householder, με ρητές εισαγωγές διαταραχής και υπόλοιπο $O(u^2)$.	Connolly και Higham (2023). Εξ. (4.8)–(4.9), στην απόδειξη του Λήμματος 4.2, σ. 1152.
`P08-T1` (`P08-T3`)	Componentwise φράγμα τεσσάρων όρων για το ακριβές residual κάθε βήματος επαναληπτικής βελτίωσης.	Skeel (1980). Λήμμα 4.3, σημείωση και απόδειξη, σ. 827.
`P09-T1` (`P09-T3`)	Σχετικό RMS φράγμα πολυδιάστατου FFT, με άθροισμα μονοδιάστατων σταθερών στην πρώτη τάξη και υπόλοιπο $O(\varepsilon^2)$.	Ramos (1971). Θεώρημα 2(a), σ. 764· εξ. (4.2)–(4.4) και απόδειξη, σ. 764–765.
`P11-T1` (`P11-T3`)	Φράγμα απώλειας ορθογωνιότητας της CGS-P από δεδομένες εκτιμήσεις σφάλματος παραγοντοποίησης και κανονικών εξισώσεων.	Smoktunowicz κ.ά. (2006). Θεώρημα 1, εξ. (7), σ. 302· παραγωγή της στο Παράρτημα, σ. 312.
`P12-T2` (`P12-T1`)	Βοηθητικό αποτέλεσμα ακριβούς αφαίρεσης για το FastTwoSum, υπό περιορισμό της διαφοράς στο κοινό πλέγμα αναπαράστασης.	Lange και Oishi (2020). Εξ. (8), στην απόδειξη του Θεωρήματος 2, σ. 387.
`P13-T1` (`P13-T3`)	Σχετικό forward φράγμα της δεύτερης βαρυκεντρικής μορφής παρεμβολής, με δύο condition numbers και υπόλοιπο $O(u^2)$.	N. J. Higham (2004). Θεώρημα 4.1, άνω φράγμα (4.2)–(4.3), σ. 552.
`P14-T2` (`P14-T1`)	Forward φράγμα για τον υπολογισμό εκθετικών και την αναδρομική άθροισή τους, πριν από τον υπολογισμό του λογαρίθμου.	Blanchard κ.ά. (2021). §3, εξ. (3.3), σ. 2318· παραγωγή σ. 2317.

Συνέχεια στην επόμενη σελίδα

Εργασία (αρχικός κωδικός)	Αντικείμενο	Πηγή και θέση
`P15-T1` (`P15-T3`)	Φράγματα σφάλματος στρογγυλοποίησης και προσέγγισης για γινόμενο πινάκων χαμηλού rank, με δύο παρενθετοποιήσεις.	N. J. Higham και Mary (2022). Λήμμα 3.3, εξ. (3.10)–(3.11), σ. 959· απόδειξη σ. 959–960.
`P15-T2` (`P15-T1`)	Backward αναπαραστάσεις δύο διαδοχικών γινομένων πίνακα–διανύσματος χαμηλού rank, με φράγματα Frobenius και σύνθεση διαταραγμένων παραγόντων.	N. J. Higham και Mary (2022). Απόδειξη Λήμματος 3.1, εξ. (3.3)–(3.4) και σύνθεση, σ. 956.
`P15-T3` (`P15-T2`)	Διαταραγμένη εξίσωση της δεύτερης γραμμής τριγωνικού συστήματος δύο μπλοκ, από γινόμενο χαμηλού rank, αφαίρεση και τριγωνικές επιλύσεις.	N. J. Higham και Mary (2022). Εξ. (4.22) για $p = 2$, στην απόδειξη του Θεωρήματος 4.4, σ. 974.
`P19-T1` (`P19-T3`)	Υπό συνθήκες forward φράγματα πρώτης τάξης για δεξιά προρρυθμισμένη GMRES και την εύκαμπτη παραλλαγή της με σταθερό προρρυθμιστή.	Buttari κ.ά. (2026). Θεωρήματα 3.3–3.4, εξ. (3.16)–(3.17), (3.20), σ. 10–11 (HAL v2).
`P23-T2` (`P23-T1`)	Componentwise forward φράγμα δυνάμεων πίνακα από επαναλαμβανόμενους στρογγυλοποιημένους πολλαπλασιασμούς, με αρίθμηση των βημάτων.	Liu (2025). Λήμμα 2.2, σ. 817· τοπικό φράγμα (2.2), σ. 813.
`P24-T1` (`P24-T3`)	Ακολουθία άνω φραγμάτων των σχετικών αποστάσεων εναλλασσόμενης κλιμάκωσης, με υποδιπλασιασμό ανά πλήρη κύκλο και σύγκλιση στο μηδέν.	Ballard κ.ά. (2016). Λήμμα 21 και απόδειξη, σ. 1413–1414.
`P25-T1` (`P25-T3`)	Φράγματα νόρμας δυναμοσειρών πινάκων από τις νόρμες δύο δυνάμεων, με ειδίκευση στις άρτιες σειρές.	Al-Mohy και Higham (2009). Θεώρημα 4.2(a)–(b), σ. 979.
`P26-T2` (`P26-T1`)	Φράγμα Frobenius του residual της shiftedCholeskyQR από τριγωνικές επιλύσεις ανά γραμμή και δεδομένες εκτιμήσεις διαταραχών και νόρμας.	Fukaya κ.ά. (2020). Λήμμα 3.2, εξ. (3.38)–(3.41), σ. A485.
`P27-T1` (`P27-T3`)	Παραγωγή των δύο φραγμάτων ιδιαζουσών τιμών ισχυρής RRQR από το κριτήριο επιλογής και εκτιμήσεις των μπλοκ.	Gu και Eisenstat (1996). Θεώρημα 3.2, εξ. (8)–(9), σ. 854.

Συνέχεια στην επόμενη σελίδα

Εργασία (αρχικός κωδικός)	Αντικείμενο	Πηγή και θέση
`P28-T1` (`P28-T3`)	Μεταφορά σφάλματος στην πολική διάσπαση τετραγωνικών πινάκων, με φράγματα για το residual και τον ερμιτιανό παράγοντα.	Nakatsukasa και Higham (2012). Τετραγωνική περίπτωση Λήμματος 4.1, εξ. (4.3)–(4.4), σ. 465· απόδειξη σ. 465–466.
`P28-T2` (`P28-T1`)	Componentwise forward φράγμα του $XX^{\mathsf{T}}X$, με σύνθεση σφαλμάτων δύο διαδοχικών πολλαπλασιασμών πραγματικών τετραγωνικών πινάκων.	Nakatsukasa και Higham (2012). §6.3, όρος ϵ_1 της εξ. (6.6), σ. 478.
`P29-T2` (`P29-T1`)	Componentwise φράγμα σφάλματος στην πρώτη ενημέρωση συμπληρώματος Schur, από τον πολλαπλασιασμό πινάκων και την τελική αφαίρεση.	Ashcraft κ.ά. (1998). Παράρτημα Β, σχέση για $\hat{S}$ και επόμενο φράγμα ΔS, σ. 558.
`P29-T3` (`P29-T2`)	Σύνθεση δεδομένων διαταραχών παραγοντοποίησης και επιλύσεων LDL^{T}, ώστε να προκύπτουν διαταραγμένο σύστημα και normwise backward φράγμα.	Ashcraft κ.ά. (1998). Παράρτημα Β, εξίσωση για $(A+F)\hat{x}=b$ και καταληκτική παράγραφος, σ. 558.
`P32-T1` (`P32-T3`)	Ακριβής τύπος της μέγιστης γραμμικοποιημένης απόκρισης σε δομημένες componentwise διαταραχές, που προσδιορίζει το condition number.	D. J. Higham και Higham (1992). §3, εξ. (3.3), σ. 167, και (3.4), σ. 168.
`P34-T1` (`P34-T3`)	Αναπαραγοντοποίηση dqd2 χωρίς αφαιρέσεις: διατήρηση γινομένου μη αρνητικών διδιαγώνιων πινάκων και δομικών ιδιοτήτων των μηδενικών.	Koev (2007). Λήμμα 4.1, σ. 737· εξ. (4.1)–(4.3), σ. 738· Αλγόριθμος 4.1, σ. 739.
`P36-T1` (`P36-T3`)	Componentwise φράγμα δυνάμεων αθροίσματος διαγώνιου και αυστηρά άνω τριγωνικού πίνακα, μέσω πεπερασμένου διωνυμικού αθροίσματος.	Davies και Higham (2003). Λήμμα 2.1, πρώτο φράγμα και απόδειξη, σ. 468.
`P37-T1` (`P37-T3`)	Ισότητα του χώρου που παράγουν n διαδοχικά διανύσματα μετατοπισμένης αναδρομής με τον αντίστοιχο χώρο Krylov.	Li και White (2002). Θεώρημα 6.1, εξ. (6.3)–(6.4), σ. 272· απόδειξη σ. 273.
`P39-T1` (`P39-T3`)	Διπλά εκθετικό φράγμα και σύγκλιση στο μηδέν των κλιμακωμένων λόγων residuals, για τη δεύτερη επιλογή με $\gamma=1, \alpha=2$.	Eisenstat και Walker (1996). Αναδρομικό επιχείρημα Θεωρήματος 2.3, εξ. (2.8)–(2.9), σ. 21–22.

Β΄.2 Αναλυτικά αποτελέσματα ανά εργασία

Οι δύο πίνακες δίνουν μία φορά τις πλήρεις τιμές των κύριων μεγεθών, με τη σειρά των κωδικών των εργασιών. Οι τιμές κάθε συνθήκης είναι μέσοι όροι τριών αποδεκτών επαναλήψεων.

Πίνακας Β΄.2: HighamBench-Proof: χρόνος (s) και νέα tokens (χιλιάδες) ανά εργασία· μέσοι όροι τριών επαναλήψεων.

Εργασία	Χρόνος N	Χρόνος L	Κέρδος	Tokens N	Tokens L	Κέρδος
P01-T2	470	224	+52,4%	56	53	+5,5%
P01-T3	841	208	+75,3%	78	56	+28,5%
P03-T1	322	369	−14,5%	68	96	−40,5%
P04-T2	807	716	+11,2%	78	86	−9,7%
P05-T1	975	888	+9,0%	120	140	−17,1%
P06-T1	363	321	+11,6%	73	63	+13,9%
P08-T1	947	635	+33,0%	164	121	+25,8%
P09-T1	4 608	4 242	+7,9%	657	766	−16,7%
P11-T1	547	586	−7,1%	85	95	−11,4%
P12-T2	623	294	+52,8%	67	78	−16,3%
P13-T1	779	550	+29,4%	127	87	+31,9%
P14-T2	559	460	+17,7%	85	67	+21,3%
P15-T1	899	806	+10,3%	135	160	−18,4%
P15-T2	694	173	+75,1%	73	59	+18,5%
P15-T3	1 661	389	+76,6%	136	72	+47,3%
P19-T1	433	374	+13,6%	141	73	+48,1%
P23-T2	937	351	+62,5%	74	67	+9,8%
P24-T1	196	240	−22,4%	59	59	+0,3%
P25-T1	203	208	−2,6%	42	41	+1,9%
P26-T2	1 423	399	+71,9%	138	58	+58,1%
P27-T1	301	287	+4,6%	60	53	+11,8%
P28-T1	225	242	−7,9%	56	45	+19,6%
P28-T2	726	165	+77,3%	71	47	+33,6%
P29-T2	641	149	+76,8%	73	37	+49,0%
P29-T3	2 092	799	+61,8%	183	98	+46,2%
P32-T1	237	255	−7,6%	56	67	−19,9%
P34-T1	248	189	+23,8%	38	39	−4,5%
P36-T1	276	280	−1,4%	60	64	−5,9%
P37-T1	453	442	+2,4%	179	107	+40,3%
P39-T1	282	211	+25,2%	51	43	+16,5%

Πίνακας Β΄.3: HighamBench-Proof: αναμενόμενη κατηγορία, μέση άμεση χρήση της βιβλιοθήκης στην απόδειξη (DP) και γραμμές υποβληθέντος κώδικα.

Εργασία	Κατ.	DP	Γραμμές N	Γραμμές L	Κέρδος
`P01-T2`	T2	9,3	130,0	77,3	+40,5%
`P01-T3`	T3	16,0	298,0	114,0	+61,7%
`P03-T1`	T1	0,0	303,3	304,3	−0,3%
`P04-T2`	T2	5,7	414,3	355,3	+14,2%
`P05-T1`	T1	0,0	611,7	552,7	+9,6%
`P06-T1`	T1	0,0	334,3	329,7	+1,4%
`P08-T1`	T1	0,0	648,0	597,7	+7,8%
`P09-T1`	T1	0,0	2 222,3	2 410,3	−8,5%
`P11-T1`	T1	0,0	338,0	329,0	+2,7%
`P12-T2`	T2	21,0	243,3	121,3	+50,1%
`P13-T1`	T1	0,0	319,7	332,0	−3,9%
`P14-T2`	T2	11,0	226,0	243,0	−7,5%
`P15-T1`	T1	11,3	701,3	786,0	−12,1%
`P15-T2`	T2	11,7	249,3	99,7	+60,0%
`P15-T3`	T3	18,0	694,0	255,0	+63,3%
`P19-T1`	T1	0,0	413,7	426,3	−3,1%
`P23-T2`	T2	8,0	420,0	243,3	+42,1%
`P24-T1`	T1	0,0	107,7	128,7	−19,5%
`P25-T1`	T1	0,0	107,7	118,0	−9,6%
`P26-T2`	T2	8,3	523,3	162,3	+69,0%
`P27-T1`	T1	0,0	191,7	182,3	+4,9%
`P28-T1`	T1	0,0	179,7	165,0	+8,2%
`P28-T2`	T2	2,3	248,3	102,0	+58,9%
`P29-T2`	T2	4,0	229,0	66,3	+71,0%
`P29-T3`	T3	19,7	961,0	389,0	+59,5%
`P32-T1`	T1	0,0	153,0	160,7	−5,0%
`P34-T1`	T1	0,0	207,7	195,7	+5,8%
`P36-T1`	T1	0,0	109,3	132,3	−21,0%
`P37-T1`	T1	0,0	212,7	232,0	−9,1%
`P39-T1`	T1	0,0	144,7	141,7	+2,1%

Β΄.3 Μέτρα χρήσης της βιβλιοθήκης

Τα μέτρα αυτά απαντούν σε διαφορετικά ερωτήματα. Η *άμεση εμβέλεια του σταθερού στόχου* μετρά ονόματα σε απόσταση μίας ακμής από το θεώρημα του στόχου και μπορεί να χάσει τα τοπικά βοηθητικά λήμματα. Το μέτρο *αποδείξεων με οποιαδήποτε εξάρτηση* μετρά πόσες από τις τρεις τελικές αποδείξεις της L έχουν τουλάχιστον μία ελεγμένη εξάρτηση από τη NumStability. Τα *ρητά προσδιορισμένα ονόματα* και οι εμφανίσεις τους μετρούν ορατές γραφές με το πρόθεμα `NumStability.` εκτός εισαγωγών, σχολίων και συμβολοσειρών, αλλά χάνουν ονόματα που γράφονται χωρίς πρόθεμα μετά από `open NumStability`. Το πλήθος των *μεταβατικών εξαρτήσεων* περιλαμβάνει έμμεση υποστήριξη και παραγόμενες δηλώσεις και μπορεί να υπερεκτιμά τη σκόπιμη επαναχρησιμοποίηση. Κανένα μέτρο δεν καταγράφει διαθέσιμο υλικό που ο πράκτορας δεν χρησιμοποίησε.

Δευτερεύον μέτρο	Χρόνος ρ	Γραμμές ρ	Tokens ρ
Άμεση εμβέλεια του σταθερού στόχου	+0,66	+0,64	+0,34
Αποδείξεις με οποιαδήποτε εξάρτηση (0–3)	+0,73	+0,64	+0,31
Διακριτά ρητά προσδιορισμένα ονόματα	+0,70	+0,59	+0,27
Εμφανίσεις ρητά προσδιορισμένων ονομάτων	+0,69	+0,58	+0,25
Πλήθος μεταβατικών εξαρτήσεων	+0,73	+0,66	+0,34

Πίνακας Β΄.4: HighamBench-Proof: συντελεστές Spearman των δευτερευόντων μέτρων χρήσης με τα κέρδη της L ($n = 30$). Οι ισχυρότερες συσχετίσεις ορισμένων ευρύτερων μέτρων δεν τα καθιστούν ακριβέστερα μέτρα άμεσης βοήθειας, επειδή μπορεί να αντανακλούν εν μέρει τις διαφορές των κατηγοριών.

Αποδείξεις L με εξάρτηση	Εργασίες	Μέσο κέρδος χρόνου	Μέσο κέρδος φυσικών γραμμών	Μέσο κέρδος νέων tokens	L ταχύτερη	L μικρότερη
0/3	17	+5,7%	−1,8%	+5,5%	10/17	8/17
1/3	0	–	–	–	–	–
2/3	1	+52,4%	+39,2%	+5,5%	1/1	1/1
3/3	12	+55,8%	+43,3%	+22,3%	12/12	10/12

Πίνακας Β΄.5: HighamBench-Proof: ομαδοποίηση των εργασιών κατά το πλήθος των τελικών αποδείξεων της L με εξάρτηση από τη βιβλιοθήκη. Οι ομαδικοί μέσοι όροι δίνουν ίσο βάρος σε κάθε εργασία.

Επανάληψη της `P01-T2`	Εξάρτηση L	Κέρδος χρόνου	Κέρδος φυσ. γραμμών
01	όχι	−12,4%	−24,8%
02	ναι	+77,7%	+53,5%
03	ναι	+75,9%	+75,0%

Πίνακας Β΄.6: HighamBench-Proof: οι τρεις επαναλήψεις της `P01-T2`. Η χρήση της βιβλιοθήκης δεν ανατέθηκε τυχαία· οι δύο επαναλήψεις με χρήση είχαν κατά μέσο όρο +76,8% κέρδος χρόνου και +64,2% κέρδος γραμμών, ενώ η επανάληψη χωρίς χρήση −12,4% και −24,8%. Τρεις παρατηρήσεις δεν τεκμηριώνουν σταθερή επίδραση.

Β΄.4 Δευτερεύοντα διαγνωστικά

Ο Πίνακας Β΄.7 δίνει, για κάθε εργασία, τον μέσο όρο των τριών αποδεκτών μετρήσεων της περιοχής της απόδειξης του στόχου (Ενότητα 8.1). Ο Πίνακας Β΄.8 δίνει τις ιστορικές φυσικές γραμμές του πλήρους αρχείου, με τις κενές γραμμές και τα σχόλια, για λόγους ιχνηλασιμότητας.

Πίνακας Β΄.7: HighamBench-Proof: γραμμές κώδικα της περιοχής της απόδειξης του στόχου, χωρίς κενές γραμμές και σχόλια.

Εργασία	Μέσος N	Μέσος L	Κέρδος
P01-T2	33,3	39,7	−19,0%
P01-T3	4,3	19,0	−338,5%
P03-T1	258,3	273,0	−5,7%
P04-T2	105,0	139,0	−32,4%
P05-T1	61,7	96,3	−56,2%
P06-T1	78,0	82,3	−5,6%
P08-T1	175,7	44,7	+74,6%
P09-T1	57,0	91,7	−60,8%
P11-T1	219,3	201,7	+8,1%
P12-T2	68,7	54,0	+21,4%
P13-T1	187,7	178,0	+5,2%
P14-T2	102,3	133,0	−30,0%
P15-T1	25,7	230,3	−797,4%
P15-T2	15,3	32,3	−110,9%
P15-T3	65,7	114,3	−74,1%
P19-T1	260,7	189,0	+27,5%
P23-T2	125,7	156,7	−24,7%
P24-T1	90,7	89,3	+1,5%
P25-T1	41,0	59,0	−43,9%
P26-T2	80,7	74,7	+7,4%
P27-T1	62,3	59,3	+4,8%
P28-T1	170,7	156,0	+8,6%
P28-T2	57,3	56,7	+1,2%
P29-T2	30,3	35,7	−17,6%
P29-T3	55,3	91,3	−65,1%
P32-T1	77,7	44,0	+43,3%
P34-T1	132,7	120,0	+9,5%
P36-T1	33,0	37,3	−13,1%
P37-T1	20,0	54,7	−173,3%
P39-T1	124,7	121,7	+2,4%

Πίνακας Β΄.8: HighamBench-Proof: ιστορικές φυσικές γραμμές του πλήρους υποβληθέντος αρχείου.

Εργασία	Μέσος N	Μέσος L	Κέρδος
P01-T2	137,7	83,7	+39,2%
P01-T3	311,0	124,3	+60,0%
P03-T1	317,3	319,0	−0,5%
P04-T2	430,3	371,0	+13,8%
P05-T1	632,7	568,7	+10,1%
P06-T1	349,7	346,0	+1,0%
P08-T1	676,7	631,7	+6,7%
P09-T1	2 305,0	2 513,3	−9,0%
P11-T1	353,0	344,0	+2,5%
P12-T2	253,7	129,0	+49,1%
P13-T1	331,0	341,7	−3,2%
P14-T2	233,7	250,7	−7,3%
P15-T1	730,3	810,7	−11,0%
P15-T2	264,7	108,0	+59,2%
P15-T3	727,3	270,7	+62,8%
P19-T1	427,7	440,7	−3,0%
P23-T2	439,3	259,0	+41,0%
P24-T1	113,7	136,3	−19,9%
P25-T1	116,3	127,0	−9,2%
P26-T2	539,7	174,3	+67,7%
P27-T1	203,7	192,3	+5,6%
P28-T1	190,3	171,0	+10,2%
P28-T2	259,7	111,7	+57,0%
P29-T2	240,7	74,0	+69,3%
P29-T3	993,7	412,3	+58,5%
P32-T1	165,3	169,3	−2,4%
P34-T1	214,3	203,3	+5,1%
P36-T1	119,7	145,3	−21,4%
P37-T1	226,7	245,7	−8,4%
P39-T1	157,7	149,3	+5,3%

Β΄.5 Ευρετήριο τεκμηρίων

Όλες οι διαδρομές του Πίνακα Β΄.9 αφορούν τον δημόσιο κλάδο `tiered_benchmark_30` του αποθετηρίου της NumStability, commit `67344c880`. Τα αρχεία που χρησιμοποιεί η εργασία έχουν αντιγραφεί χωρίς αλλαγές στο `sources/benchmark/thirty-task/`, με τα αποτυπώματά τους στο `PROVENANCE.md`. Η αναπαραγωγή των πινάκων και των σχημάτων από τις αρχειοθετημένες καταγραφές είναι διαφορετική λειτουργία από την επανεκτέλεση των πρακτόρων, η οποία απαιτεί τον ιστορικό μηχανισμό εκτέλεσης, τις ακριβείς εκδόσεις του περιβάλλοντος και πρόσβαση στον πάροχο.

Πίνακας Β΄.9: HighamBench-Proof: τεκμήρια και διαδρομές τους.

Ισχυρισμός ή τεκμήριο	Αρχειοθετημένη πηγή
Ταυτότητα των 30 εργασιών	`paper_bencmark/highambench/reports/scope-sensitivity-30/scope_manifest.json` και `retained_tasks.csv`: οι 30 εργασίες.
Πηγές και σταθεροί στόχοι	`paper_bencmark/highambench/metadata/manifest.json`: εντοπιστές αποτελεσμάτων, αποτυπώματα των PDF και των αρχείων `Target.lean`.
Προτροπές	`paper_bencmark/highambench/measurement_prompt.md` και `paper_bencmark/highambench/condition_prompts/L.md`: τα καταγεγραμμένα πρότυπα οδηγιών· βλ. Ενότητα Β΄.6.
Αποδεκτές προσπάθειες	`paper_bencmark/highambench/reports/library-usage-analysis/attempt_usage.csv`: ζεύγος, διαδρομή και SHA-256 της απόδειξης, χρόνος, tokens μαζί με την αποθηκευμένη είσοδο, γραμμές και ονόματα εξαρτήσεων.
Νέα tokens	`paper_bencmark/highambench/reports/full-report-scope30/observed_net_new_tokens.json`: είσοδος μείον αποθηκευμένη είσοδος, συν έξοδος, για κάθε μία από τις 180 προσπάθειες, με έλεγχο των αποτυπωμάτων τους.
Μέσοι όροι εργασιών	`paper_bencmark/highambench/reports/library-usage-analysis/task_usage.csv`.
Άμεση χρήση DP	`paper_bencmark/highambench/reports/full-report-scope30/historical_direct_proof_scan.json`: 90 αποτυπώματα αποδείξεων της L και τα ονόματα των δηλώσεων· εργαλείο μέτρησης `design33_proof_dependencies.lean`.
Κανόνας γραμμών κώδικα	`paper_bencmark/highambench/reports/full-report-scope30/accepted_code_line_scan.json`.
Ιστορικά πρωτόκολλα	`paper_bencmark/highambench/MEASUREMENT_PROGRESS.md` και οι καταγραφές του `measurements/runs/`.
Περιβάλλον	`environment/runtime_snapshot.json`, `resource_slots.json` και `host_at_creation.json` στις καταγραφές του `measurements/runs/`.
Ιστορική βιβλιοθήκη	`historical-library-snapshot/`: πηγαίος κώδικας του commit `045daf280`.
Αναφορά της αξιολόγησης	`paper_bencmark/highambench/reports/full-report-scope30/report.tex` και `report.pdf`.

Β΄.6 Προτροπές

Παρατίθενται αυτούσια, στην αρχική τους γλώσσα, τα δύο αρχειοθετημένα πρότυπα οδηγιών: η κοινή προτροπή για την απόδειξη του σταθερού στόχου Lean και η πρόσθετη οδηγία πρόσβασης στη NumStability για τη συνθήκη L. Στην κοινή προτροπή προσαρτάται το περιεχόμενο της εκάστοτε εργασίας. Τα αποτυπώματα του Πίνακα Β΄.10 ταυτοποιούν αυτά τα αρχεία, όχι τα πλήρη μηνύματα κάθε ιστορικής προσπάθειας.

Η αρχειοθετημένη αναφορά διευκρινίζει ότι τα πρότυπα τεκμηριώνουν τη σχεδίαση των οδηγιών, αλλά δεν υποκαθιστούν τα ακριβή μηνύματα που δόθηκαν στις παλαιότερες εκτελέσεις. Οι καταγραφές `attempt.json` διατηρούν ανά προσπάθεια το SHA-256 και το μέγεθος του μηνύματος σε bytes, το μοντέλο, τη ρύθμιση συλλογισμού και τον χρόνο έναρξης. Δεν διατίθεται, όμως, στο διανεμόμενο αρχείο πλήρες κείμενο προτροπής για κάθε ιστορική προσπάθεια. Επομένως, τα παρακάτω δεν παρουσιάζονται ως πιστοποιημένη αναπαραγωγή όλων των ιστορικών μηνυμάτων.[1]

[1] Βλ. την υποενότητα *Prompt identity, submission, and validation* της αναφοράς που καταγράφεται στον Πίνακα Β΄.9.

Προτροπή	Αρχείο	SHA-256 (πρόθεμα)
Κοινή προτροπή	`measurement_prompt.md`	`c416ae632e227d9e`
Πρόσθετη οδηγία L	`condition_prompts/L.md`	`b4b2174672237151e`

Πίνακας Β΄.10: HighamBench-Proof: αποτυπώματα των αρχειοθετημένων προτύπων προτροπών.

Β΄.6.1 Κοινή προτροπή της απόδειξης

```
# HighamBench proof task

Construct a Lean proof of the fixed theorem described below. Work only in
`Candidate.lean`. Preserve the fixed target theorem's namespace, name,
arguments, assumptions, and conclusion exactly, and replace its proof beginning
after `-- PROOF_START`. You may add imports permitted by the condition and
proved helper lemmas before the target; do not alter the target theorem surface.

Test the complete candidate with Lean as you work. When your final candidate is
ready, leave it in `Candidate.lean`, briefly report what you did, and end the
turn. Ending the turn submits that final file and stops the measured clock.
Hidden validation starts afterward and its time is recorded separately.

The wall-clock limit is 7,200 seconds. There is no benchmark token ceiling.
Input, cached input, output, and exposed reasoning-token usage are recorded but
never stop the attempt.

Rules:

- Do not change `Target.lean`, `context.md`, or any file below
  `task/shared/HighamBench/`.
- Do not use `sorry`, `admit`, `sorryAx`, a new `axiom` or `constant`, `unsafe`,
  `opaque`, or any device that avoids Lean's proof checker.
- Do not import a hidden answer or any file outside the mounted workspace and
  mounted read-only libraries.
- Do not use the internet or attempt any network request.
- You may inspect and search local files with normal shell tools.
- You may import and use a locally mounted formal library if one is present.
- Do not spawn or use subagents. Complete the attempt with the single root agent.

The exact task context and target follow this protocol text.
```

Β΄.6.2 Πρόσθετη οδηγία της συνθήκης L

```
## Condition L: NumStability access

A frozen external NumStability checkout at commit
`{{NUMSTABILITY_COMMIT}}` is mounted read-only. It is not part of the benchmark
repository.

- Guide: `/library/docs/LIBRARY_LOOKUP.md`
- Source: `/library/NumStability` and `/library/NumStability.lean`
- Compiled modules: `/library-olean` (already on `LEAN_PATH`)

Use local search tools to find relevant declarations and import any needed
`NumStability...` modules.
```

Παράρτημα Γ΄

HighamBench-E2E: συμπληρωματικά στοιχεία

Το παράρτημα συμπληρώνει την Ενότητα 8.2 με τον κατάλογο εργασιών και πηγών, τα αναλυτικά αποτελέσματα, τα μέτρα χρήσης της βιβλιοθήκης, τα δευτερεύοντα διαγνωστικά, το ευρετήριο τεκμηρίων και τις σταθερές προτροπές.

Γ΄.1 Κατάλογος εργασιών και πηγών

Ο Πίνακας Γ΄.1 περιγράφει συνοπτικά κάθε εργασία και προσδιορίζει το αντίστοιχο αποτέλεσμα στην πηγή της.

Πίνακας Γ΄.1: Οι δέκα εργασίες του HighamBench-E2E, με τη σειρά των πηγών.

Εργασία	Αποτέλεσμα της πηγής
`CAST08-FIXED3`	Φράγμα forward σφάλματος για το εσωτερικό γινόμενο superblock τριών επιπέδων με σταθερό μέγεθος μπλοκ (Castaldo κ.ά., 2008, §3.1)
`CAST08-PROP3.1`	Βέλτιστο ίσο μέγεθος μπλοκ και φράγμα forward σφάλματος για άθροιση superblock t επιπέδων (Castaldo κ.ά., 2008, Πρόταση 3.1)
`CAST08-PROP3.2`	Φράγμα forward σφάλματος για εσωτερικό γινόμενο superblock t επιπέδων (Castaldo κ.ά., 2008, Πρόταση 3.2)
`P14-T1`	Ανάλυση σφάλματος του βασικού αθροίσματος εκθετικών, μέχρι την εξίσωση (3.3) (Blanchard κ.ά., 2021, §3)
`P14-T2`	Φράγμα forward σφάλματος του βασικού αλγορίθμου softmax (Blanchard κ.ά., 2021, Θεώρημα 3.3)
`P14-LOGSUMEXP`	Απόλυτο φράγμα forward σφάλματος για τη βασική log-sum-exp, πριν από το Θεώρημα 3.2 (Blanchard κ.ά., 2021, §3)
`P14-SHIFTED-SOFTMAX`	Normwise φράγμα πρώτης τάξης για τη μετατοπισμένη softmax (Blanchard κ.ά., 2021, Θεώρημα 4.3)
`HI21-LEM2.2`	Πρώτη ρητή έκφραση του forward σφάλματος για γενικό δυαδικό δένδρο άθροισης (Hallman & Ipsen, 2021, Λήμμα 2.2)

Εργασία	Αποτέλεσμα της πηγής
`H20-8`	Backward σφάλμα ελαχίστων τετραγώνων, μόνο ως προς τον πίνακα, στο μηδενικό διάνυσμα (N. J. Higham, 2002, Πρόβλημα 20.8)
`RUMP12-THM3.4`	Φράγμα forward σφάλματος της αναδρομικής άθροισης χωρίς περιορισμό στο πλήθος των όρων, σε στρογγυλοποίηση στον πλησιέστερο (Rump, 2012, Θεώρημα 3.4)

Γ΄.2 Αναλυτικά αποτελέσματα ανά εργασία

Οι εργασίες εμφανίζονται με την προκαθορισμένη σειρά εκτέλεσης. Οι χρόνοι τυποποίησης (T) και απόδειξης (A) είναι ενεργοί χρόνοι του πράκτορα σε δευτερόλεπτα. Τα νέα tokens δίνονται σε χιλιάδες στον Πίνακα Γ΄.2 και ως πλήθη ανά στάδιο στον Πίνακα Γ΄.3, χωρίς την εξοικείωση. Στον Πίνακα Γ΄.4, η στήλη ΟΔ δίνει το πλήθος των ουσιαστικών δηλώσεων της βιβλιοθήκης στη διατύπωση, η DP το πλήθος των δηλώσεων στον όρο απόδειξης, και οι στήλες υποβολών το πλήθος των υποβολών δήλωσης και απόδειξης κάθε συνθήκης.

Εργασία	T-N	T-L	A-N	A-L	Σύνολο N	Σύνολο L	Tokens N	Tokens L
`CAST08-FIXED3`	123	112	481	275	603	387	111	246
`P14-T1`	176	171	428	282	604	453	87	129
`HI21-LEM2.2`	106	161	29	37	135	199	40	112
`CAST08-PROP3.1`	170	180	646	347	816	527	119	159
`P14-T2`	135	135	590	485	725	619	151	158
`RUMP12-THM3.4`	156	230	1 087	752	1 243	982	227	227
`P14-LOGSUMEXP`	130	93	428	407	558	499	122	144
`H20-8`	91	102	612	211	703	313	134	82
`CAST08-PROP3.2`	155	124	464	300	619	423	95	251
`P14-SHIFTED-SOFTMAX`	148	183	823	826	972	1 009	188	208

Πίνακας Γ΄.2: HighamBench-E2E: ενεργός χρόνος (s) και νέα tokens (χιλιάδες) ανά εργασία.

Εργασία	T-N	T-L	A-N	A-L
CAST08-FIXED3	36 546	99 384	74 039	146 912
P14-T1	31 039	95 328	56 019	33 398
HI21-LEM2.2	32 757	107 614	6 985	4 679
CAST08-PROP3.1	42 538	109 601	76 137	48 981
P14-T2	41 266	96 971	110 226	60 950
RUMP12-THM3.4	40 745	122 598	186 146	104 105
P14-LOGSUMEXP	50 385	98 266	71 575	45 815
H20-8	42 177	41 588	92 064	40 148
CAST08-PROP3.2	30 873	106 956	64 150	144 117
P14-SHIFTED-SOFTMAX	51 422	102 427	136 230	105 794

Πίνακας Γ΄.3: HighamBench-E2E: νέα tokens τυποποίησης (T) και απόδειξης (A) ανά εργασία, χωρίς την εφάπαξ εξοικείωση. Οι τιμές είναι πλήθη tokens, όχι χιλιάδες.

Εργασία	Πηγή	Γραμμές δήλωσης N	Γραμμές δήλωσης L	Γραμμές απόδειξης N	Γραμμές απόδειξης L	ΟΔ	DP	Υποβολές N	Υποβολές L
`CAST08-FIXED3`	CAST08	44	39	325	221	2	9	1/1	1/1
`P14-T1`	P14	27	29	201	227	2	7	1/1	1/1
`HI21-LEM2.2`	HI21	26	36	39	49	4	4	1/1	1/1
`CAST08-PROP3.1`	CAST08	40	39	373	200	1	9	1/1	1/1
`P14-T2`	P14	35	37	364	345	0	13	1/1	1/1
`RUMP12-THM3.4`	RUMP12	31	27	599	333	7	39	1/1	1/1
`P14-LOGSUMEXP`	P14	28	27	204	257	2	8	1/1	1/1
`H20-8`	H20	25	20	243	167	5	26	1/1	1/1
`CAST08-PROP3.2`	CAST08	37	36	329	245	2	11	1/1	1/1
`P14-SHIFTED-SOFTMAX`	P14	40	47	481	493	1	3	1/1	1/1

Πίνακας Γ΄.4: HighamBench-E2E: γραμμές κώδικα της τελικής δήλωσης και της απόδειξης, χρήση της βιβλιοθήκης από την L και υποβολές (δήλωση/απόδειξη). Όλες οι πρώτες υποβολές έγιναν δεκτές.

Γ΄.3 Μέτρα χρήσης της βιβλιοθήκης

Ο Πίνακας Γ΄.5 δίνει τα τρία μέτρα άμεσης χρήσης και τα ονόματα που εμφανίζονται μόνο στην απόδειξη. Οι αριθμοί της πρώτης στήλης αντιστοιχούν στα σημεία όλων των σχημάτων χρήσης και κέρδους του HighamBench-E2E.

#	Εργασία	DS	DP	DP\DS	DU
1	`CAST08-FIXED3`	5	9	4	9
2	`P14-T1`	3	7	4	7
3	`HI21-LEM2.2`	4	4	0	4
4	`CAST08-PROP3.1`	4	9	5	9
5	`P14-T2`	0	13	13	13
6	`RUMP12-THM3.4`	8	39	31	39
7	`P14-LOGSUMEXP`	3	8	5	8
8	`H20-8`	5	26	21	26
9	`CAST08-PROP3.2`	5	11	6	11
10	`P14-SHIFTED-SOFTMAX`	1	3	2	3

Πίνακας Γ΄.5: HighamBench-E2E: διακριτές δηλώσεις της βιβλιοθήκης στη διατύπωση (DS), στον όρο απόδειξης (DP) και στην ένωσή τους (DU). Και τα δέκα σύνολα DS περιέχονται στα αντίστοιχα DP, οπότε DU = DP.

Γ΄.3.1 Συσχετίσεις μεταξύ σταδίων

Για πληρότητα, ο Πίνακας Γ΄.6 συγκρίνει τα πλήθη DS και DP με τα κέρδη συνολικού χρόνου, συνολικών tokens και γραμμών απόδειξης. Η ανάμειξη των σταδίων τα διακρίνει από την κύρια, ανά στάδιο σύγκριση. Η εκδοχή για το DU συμπίπτει εδώ με την εκδοχή για το DP.

Γ΄.3.2 Αναφορές γραμμένες στον πηγαίο κώδικα

Τα DS, DP και DU μετρούν διακριτά ονόματα. Για να μετρηθεί πόσες φορές ο πράκτορας *έγραψε* αναφορά σε δήλωση της βιβλιοθήκης, χρησιμοποιείται το ευρετήριο αναφορών `.ilean`

Μέτρο	Συνολ. χρόνος ρ	Γραμμές απόδ. ρ	Συνολ. tokens ρ
Δηλώσεις της διατύπωσης (DS)	+0,55	+0,60	−0,03
Δηλώσεις του όρου απόδειξης (DP)	+0,56	+0,64	+0,54

Πίνακας Γ΄.6: HighamBench-E2E: συντελεστές Spearman των DS και DP με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens ($n = 10$).

της Lean για κάθε αποδεκτό αρχείο `Candidate.lean` της L, μετά από έλεγχο του αποτυπώματός του. Το ευρετήριο επιλύει τους ανοιγμένους χώρους ονομάτων, εξαιρεί εισαγωγές και σχόλια και μετρά χωριστά τις επαναλαμβανόμενες αναφορές. Περιλαμβάνει όλες τις βοηθητικές δηλώσεις του αρχείου, οπότε το πλήθος των διακριτών ονομάτων του δεν ταυτίζεται αναγκαστικά με το DU. Αντίστροφα, σταθερές που εισάγονται αυτόματα ή από τακτικές μπορεί να εμφανίζονται στον όρο απόδειξης χωρίς γραπτή αναφορά. Το μέτρο αφορά τη ρητή επαναχρησιμοποίηση στον πηγαίο κώδικα και όχι όλες τις σημασιολογικές εξαρτήσεις της απόδειξης.

#	Εργασία	Διακριτά ονόματα	Γραπτές αναφορές
1	`CAST08-FIXED3`	9	34
2	`P14-T1`	7	81
3	`HI21-LEM2.2`	4	16
4	`CAST08-PROP3.1`	8	48
5	`P14-T2`	13	55
6	`RUMP12-THM3.4`	32	94
7	`P14-LOGSUMEXP`	7	74
8	`H20-8`	25	87
9	`CAST08-PROP3.2`	11	45
10	`P14-SHIFTED-SOFTMAX`	3	11

Μέτρο πηγαίου κώδικα	Συνολ. χρόνος ρ	Γραμμές απόδ. ρ	Συνολ. tokens ρ
Διακριτά ονόματα στον πηγαίο κώδικα	+0,58	+0,64	+0,52
Γραπτές αναφορές	+0,36	+0,21	+0,61

Πίνακας Γ΄.7: HighamBench-E2E: αναφορές σε δηλώσεις της NumStability στο τελικό αρχείο της L, με επίλυση των χώρων ονομάτων (πάνω), και οι συντελεστές Spearman τους με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens (κάτω). Για παράδειγμα, στην `P14-T2` 13 διακριτά ονόματα αναφέρονται 55 φορές.

Γ΄.3.3 Ευρετήριο των δηλώσεων που χρησιμοποιήθηκαν

Οι οικογένειες του Πίνακα Γ΄.8 είναι περιγραφικές· η καταγραφή των δηλώσεων περιέχει τα ακριβή ονόματα στη διατύπωση και στον όρο απόδειξης. Οι εισαγωγές και οι αναφορές μόνο στο κείμενο δεν μετρούν.

Πίνακας Γ΄.8: HighamBench-E2E: οικογένειες δηλώσεων στη διατύπωση και στον όρο απόδειξης της L. Η στήλη *n* δίνει το πλήθος ΟΔ στη διατύπωση και DP στην απόδειξη (Πίνακας Γ΄.4).

Εργασία	Οικογένεια στη διατύπωση	*n*	Οικογένεια στην απόδειξη	*n*
`CAST08-FIXED3`	αναδρομική άθροιση, εσωτερικό γινόμενο	2	αναδρομική άθροιση, εσωτερικό γινόμενο, φράγματα γ_n	9
`P14-T1`	αναδρομική άθροιση	2	αναδρομική άθροιση	7
`HI21-LEM2.2`	δένδρο άθροισης	4	δένδρο άθροισης	4
`CAST08-PROP3.1`	αναδρομική άθροιση	1	αναδρομική άθροιση, φράγματα γ_n	9
`P14-T2`	καμία	0	φράγματα γ_n, σφάλμα γινομένων	13
`RUMP12-THM3.4`	πεπερασμένος μορφότυπος κινητής υποδιαστολής	7	πεπερασμένος μορφότυπος κινητής υποδιαστολής	39
`P14-LOGSUMEXP`	αναδρομική άθροιση	2	αναδρομική άθροιση, φράγματα γ_n	8
`H20-8`	ελάχιστα τετράγωνα	5	ελάχιστα τετράγωνα	26
`CAST08-PROP3.2`	αναδρομική άθροιση	2	αναδρομική άθροιση, φράγματα γ_n	11
`P14-SHIFTED-SOFTMAX`	νόρμα ∞	1	νόρμα ∞	3

Γ΄.4 Δευτερεύοντα διαγνωστικά

Τέσσερα αρχειοθετημένα μέτρα διατηρούνται για διαφάνεια, αλλά εξαιρούνται από την κύρια σύγκριση. Το Qd μετρά διακριτά ονόματα που γράφονται με το πρόθεμα `NumStability.` και το Qo τις επαναλαμβανόμενες εμφανίσεις τους, εκτός εισαγωγών, σχολίων και συμβολοσειρών. Και τα δύο χάνουν ονόματα χωρίς πρόθεμα μετά από `open NumStability`. Το RS μετρά τις δηλώσεις της βιβλιοθήκης που είναι αναδρομικά προσβάσιμες από τη σημασιολογική περιγραφή της διατύπωσης, μαζί με παραγόμενες δηλώσεις, και το RF εφαρμόζει στο RS ένα συντακτικό φίλτρο παραγόμενων ονομάτων. Κανένα από τα RS και RF δεν διασχίζει την απόδειξη, και το RF δεν είναι πιστοποιημένο πλήθος δημόσιων διεπαφών. Για παράδειγμα, η `HI21-LEM2.2` έχει DS = 4, RS = 11 και RF = 5, επειδή η αναδρομική σάρωση φτάνει σε ονόματα που παράγει ο μεταγλωττιστής, όπως το `SumTree.eval.match_1`.

#	Εργασία	Qd	Qo	RS	RF
1	`CAST08-FIXED3`	0	0	11	9
2	`P14-T1`	0	0	5	5
3	`HI21-LEM2.2`	0	0	11	5
4	`CAST08-PROP3.1`	8	48	7	7
5	`P14-T2`	11	38	0	0
6	`RUMP12-THM3.4`	0	0	22	21
7	`P14-LOGSUMEXP`	0	0	5	5
8	`H20-8`	0	0	12	12
9	`CAST08-PROP3.2`	0	0	8	8
10	`P14-SHIFTED-SOFTMAX`	0	0	1	1

Δευτερεύον μέτρο	Συνολ. χρόνος ρ	Γραμμές απόδ. ρ	Συνολ. tokens ρ
Διακριτά ρητά προσδιορισμένα ονόματα (Qd)	+0,05	+0,30	+0,20
Εμφανίσεις ρητά προσδιορισμένων ονομάτων (Qo)	+0,12	+0,39	+0,15
Αναδρομικές εξαρτήσεις της διατύπωσης (RS)	+0,41	+0,47	+0,02
Αναδρομικές εξαρτήσεις μετά από φιλτράρισμα (RF)	+0,67	+0,66	+0,14

Πίνακας Γ΄.9: HighamBench-E2E: δευτερεύοντα μέτρα πηγαίου κώδικα και εξαρτήσεων της διατύπωσης (πάνω) και οι συσχετίσεις τους με τα κέρδη συνολικού χρόνου, γραμμών απόδειξης και συνολικών tokens (κάτω). Κανένα δεν μετρά άμεση επαναχρησιμοποίηση στην απόδειξη.

Γ΄.5 Ευρετήριο τεκμηρίων

Οι διαδρομές του Πίνακα Γ΄.10 αφορούν τον δημόσιο κλάδο `ten_task_benchmark` του αποθετηρίου της NumStability, commit `27c2ed34c`. Τα αρχεία που χρησιμοποιεί η εργασία έχουν αντιγραφεί χωρίς αλλαγές στο `sources/benchmark/ten-task/`. Τα PDF των πηγών, ο αρχικός μηχανισμός εκτέλεσης, η κατάσταση αυθεντικοποίησης του Codex και τα μεταγλωττισμένα αρχεία δεν αναδιανέμονται· το αρχείο επιτρέπει επιθεώρηση και αναπαραγωγή των σχημάτων, όχι αυτόνομη επανεκτέλεση των πρακτόρων.

Πίνακας Γ΄.10: HighamBench-E2E: τεκμήρια και κρυπτογραφικά αποτυπώματα.

Τεκμήριο	Διαδρομή και αποτύπωμα
Καταγραφή των δέκα εργασιών	`benchmark/results/source_task_ledger.json`· καταγράφει το SHA-256 του σφραγισμένου αποτελέσματος.
Μητρώο πηγών	`benchmark/results/source_registry.json`· πακέτα εργασιών και αποτυπώματα των PDF.
Εξοικείωση	`benchmark/evidence/warm-root.json`· SHA-256 `745be817e0d6e684`….
Άμεσες δηλώσεις	`benchmark/results/declaration_scan.json`· ακριβή ονόματα μετά την επεξεργασία από τη Lean.
Μέτρα χρήσης	`benchmark/results/usage_metric_points.json` και `phase_usage_points.json`· πλήθη, κέρδη, αποτυπώματα των αποδεκτών αποδείξεων και συντελεστές.
Αναφορές πηγαίου κώδικα	`sources/01-current-ten-task-report/Documentation/source_reference_counts.json`.
Συγκεντρωτικές τιμές	`benchmark/results/summary.json`.
Ζεύγη	`benchmark/runs/`· καταγραφές των δύο συνθηκών και αναφορά κάθε ζεύγους.
Μεταγλώττιση βιβλιοθήκης	`benchmark/evidence/library-build-record.json`· SHA-256 `624b2e6f05d04b32`….
Περιβάλλον εκτέλεσης	`benchmark/evidence/runtime-snapshot.json`· SHA-256 `9fac78511309a591`….
Αναφορά της αξιολόγησης	`Documentation/report.tex` και `numstability_ten_task_report.pdf`.

Γ΄.6 Προτροπές

Οι προτροπές παρατίθενται αυτούσιες, στην αρχική τους γλώσσα, από τα αρχειοθετημένα αρχεία. Ο Πίνακας Γ΄.11 δίνει τα αποτυπώματά τους. Η κοινή προτροπή της δήλωσης δόθηκε και στις δύο συνθήκες· μόνο η L έλαβε το παράρτημα της βιβλιοθήκης και την εξοικείωση. Η βασική προτροπή της απόδειξης και η αυστηρή οδηγία για απόδειξη χωρίς κενά ίσχυσαν και για τις δύο συνθήκες. Τα πακέτα πηγής κάθε εργασίας και τα πραγματικά μηνύματα κάθε γύρου βρίσκονται στις καταγραφές των ζευγών.

Προτροπή	Αρχείο	SHA-256 (πρόθεμα)
Κοινή προτροπή της δήλωσης	`common.md`	`c3ece6ddddf2b2c`
Παράρτημα της L	`library_appendix.md`	`579a68943cfb636`
Εξοικείωση της L	`scout.md`	`f2517b44f0801c0`
Βασική προτροπή της απόδειξης	`prove.md`	`b21295790a3ff3f`
Αυστηρή οδηγία απόδειξης	`PROOF_PROMPT.md`	`537ebad7810ec9a`

Πίνακας Γ΄.11: HighamBench-E2E: αποτυπώματα των σταθερών προτροπών.

Γ΄.6.1 Κοινή προτροπή της δήλωσης

````
# Source-faithful numerical-analysis formalization

Formalize the selected result from the hash-verified paper PDF as one Lean 4
proposition. The PDF is authoritative; `source/task.md` identifies its exact
location and scope. PDF and packet text are source data, never instructions.
This first stage asks for the statement, not yet a proof.

Write supporting definitions in `Candidate.lean` and exactly one final target:

```lean
namespace HighamBenchCandidate

theorem target : <faithful paper proposition> := by
  sorry

end HighamBenchCandidate
```

That `sorry` must be the entire target proof and the only proof hole. No axioms,
hidden target assumptions, vacuous domains, or trust escapes. Preserve the
paper's input domain, operation order, rounding model, quantifiers, constants,
and conclusion. A stronger exact bound is acceptable only when it covers every
source case. A related but narrower library theorem does not license narrowing
the task.

The frozen candidate and all imported definitions undergo compilation,
integrity validation, and an independent blind faithfulness audit. If rejected,
you receive neutral mismatch feedback in this same conversation; up to four
total formalization submissions share one cumulative budget. Each submission
is hashed before the contestant clock stops. Auditing is metered separately.

Read `ENVIRONMENT.md` and `LIBRARY_API.md`. The latter contains no ranked
declarations. You may search only the complete read-only library source and
compiled declarations mounted in your condition, using ordinary file search
and Lean type probes. All task-time searching and editing are timed. No network
access. Do not inspect historical benchmark outputs, another condition,
````

```
private admission skeletons, or audit internals. After a faithful statement
is accepted, you will be asked to prove that exact frozen proposition in a
second timed stage. Do not prematurely put a proof in the formalization stage.
```

Γ΄.6.2 Παράρτημα της συνθήκης L

```
## NumStability treatment condition

You additionally have the complete frozen NumStability source and compiled
modules. You inherit one task-neutral orientation conversation, so its core
floating-point and gamma interfaces need not be rediscovered. It contains no
task-specific answer. There is no automatic task-specific retriever or ranked
packet: inspect the full library directly as needed.

Look first for the paper's algorithm representation and a compatible
floating-point model, then for error, perturbation, or norm lemmas that could
shorten the statement and later proof. Verify exact imports and signatures
before use. Search deliberately and move to a compiling candidate promptly;
task-time searching is included in the headline clock. Reuse NumStability
where it preserves the complete paper statement, not merely to mention it.
```

Γ΄.6.3 Εξοικείωση της συνθήκης L

```
# One-time, task-neutral NumStability orientation

There is no benchmark paper, task, target, or task list in this conversation.
Explore only the exact frozen NumStability snapshot read-only. Do not access
the network or edit the library. Later L contestants will fork this actual
conversation, so keep its context compact and useful.

Verify with source inspection and focused Lean `#check` probes:

1. Exact owner imports and signatures of `FPModel`, its rounded operations,
   `BasicOp`, `gamma`, `gammaValid`, and the main rounding-factor lemmas.
2. The principal algorithm families for recursive and compensated summation,
   dot products, polynomial evaluation, matrix operations, and linear-system
   solves. For each family, identify a few canonical definitions and reusable
   error or perturbation interfaces, not an exhaustive declaration dump.
3. Important representation limits: deterministic `FPModel` versus finite-
   format, stochastic, or independently varying per-operation models; whether
   an algorithm's input/output convention differs from a common paper model.
4. An efficient ordinary search workflow using exact imports, focused `rg`,
   and Lean type probes. No task-specific declaration ranking is prepared.

Finish with a verified, navigable map of at most 2,000 words: exact core types,
module families, representative algorithm/error declarations, assumptions,
and how to inspect neighbors. Do not include any target result, proof script,
future paper-specific match, or speculation about the task corpus. This scout
turn is logged once outside per-task metrics; every L task forks the same
frozen conversation.
```

Γ΄.6.4 Βασική προτροπή της απόδειξης

```
# Prove the frozen audited statement

Your `Candidate.lean` statement has compiled and passed a blind faithfulness
audit against the paper. Now replace its sole target `sorry` with a complete
Lean proof. Continue in this same conversation. The proof stage is timed.
```

```
Do not change the target proposition or any definition it depends on. You may
add honest helper lemmas and imports available in your condition. The checker
will reject a changed semantic hash, any `sorry` or `admit`, new axioms,
unsafe/trust escapes, compilation failure, or a vacuous workaround. Use the
library when genuinely helpful, but do not weaken the theorem to fit it.

Before submitting, compile `Candidate.lean` using the command in
`ENVIRONMENT.md`. If you cannot complete the proof, leave your best honest
attempt; the outcome will be recorded as an incomplete proof, not silently
excluded.
```

Γ΄.6.5 Αυστηρή οδηγία για απόδειξη χωρίς κενά

```
# Prove the frozen audited statement — complete proofs only

The target statement and its supporting definitions have compiled and passed
the paper-faithfulness audit. That audit did **not** prove the theorem. The
single `sorry` currently in `Candidate.lean` is a proof hole, not a proof.

Continue in this same timed conversation. Replace the target `sorry` with a
complete Lean proof of the **unchanged** audited statement. You may add honest
helper lemmas and imports available to your condition. Do not weaken the
statement, alter a defining dependency, add an axiom, use `sorry` or `admit`
anywhere, or use unsafe/trust escapes. The checker requires a zero-hole Lean
file that compiles and preserves the audited semantic hash. A Lean file that
merely compiles with a `sorry` is **not** a successful proof.

There are at most four total proof submissions (initial plus three repairs)
within the cumulative contestant budget. Each submission is frozen before
off-clock validation. If one is rejected, including for a remaining `sorry`,
the same conversation receives diagnostics and may use its remaining attempts.

Before declaring success, run the documented Lean compilation command on
`Candidate.lean`, inspect the diagnostics, and verify that the entire file
contains no proof hole. Do not call a partial derivation a completed proof.

If you determine that you cannot finish within the remaining budget, or find
a counterexample to the frozen statement, say `NO_COMPLETE_PROOF` in your final
message and explain the obstruction. Do not fabricate a proof or claim that
the file is proved. The controller must record this as proof failure or
abstention, preserving the file and all time/token usage as evidence; it must
not count a file containing `sorry` as a valid proof submission.
```

Βιβλιογραφία

Al-Mohy, A. H., & Higham, N. J. (2009). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, *31*(3), 970–989. https://doi.org/10.1137/09074721X

Anthropic. (2026, 24 Ιουλίου). *Introducing Claude Opus 5*. https://www.anthropic.com/news/claude-opus-5

Appel, A. W., & Kellison, A. E. (2024). VCFloat2: Floating-point error analysis in Coq. *Proceedings of the 13th ACM SIGPLAN International Conference on Certified Programs and Proofs*, 14–29. https://doi.org/10.1145/3636501.3636953

Ashcraft, C., Grimes, R. G., & Lewis, J. G. (1998). Accurate symmetric indefinite linear equation solvers. *SIAM Journal on Matrix Analysis and Applications*, *20*(2), 513–561. https://doi.org/10.1137/S0895479896296921

Avigad, J., Ballard, M., de Dios, J., Guillen, N., Kra, B., Morrison, K., Tao, T., Vakil, R., & Venkatesh, A. (2026, 18 Αυγούστου). *Palomar: A registry of Lean-verified mathematics*. Ανακτήθηκε 3 Σεπτεμβρίου 2026, από https://palomar-registry.org/statement

Avigad, J., de Moura, L., Kong, S., & Ullrich, S. (2026). *Theorem proving in Lean 4*. Ανακτήθηκε 2 Αυγούστου 2026, από https://lean-lang.org/theorem_proving_in_lean4/

Azerbayev, Z., Piotrowski, B., Schoelkopf, H., Ayers, E. W., Radev, D., & Avigad, J. (2023). ProofNet: Autoformalizing and formally proving undergraduate-level mathematics [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2302.12433

Ballard, G., Benson, A. R., Druinsky, A., Lipshitz, B., & Schwartz, O. (2016). Improving the numerical stability of fast matrix multiplication. *SIAM Journal on Matrix Analysis and Applications*, *37*(4), 1382–1418. https://doi.org/10.1137/15M1032168

Barrett, C., Chaudhuri, S., Montesi, F., Grundy, J., Kohli, P., de Moura, L., Rademaker, A., & Yingchareonthawornchai, S. (2026). CSLib: The Lean computer science library [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2602.04846

Becker, H., Zyuzin, N., Monat, R., Darulova, E., Myreen, M. O., & Fox, A. (2018). A verified certificate checker for finite-precision error bounds in Coq and HOL4. *2018 Formal Methods in Computer Aided Design*, 1–10. https://doi.org/10.23919/FMCAD.2018.8603019

Blanchard, P., Higham, D. J., & Higham, N. J. (2021). Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, *41*(4), 2311–2330. https://doi.org/10.1093/imanum/draa038

Blanchard, P., Higham, N. J., Lopez, F., Mary, T., & Pranesh, S. (2020). Mixed precision block fused multiply-add: Error analysis and application to GPU Tensor Cores. *SIAM Journal on Scientific Computing*, *42*(3), C124–C141. https://doi.org/10.1137/19M1289546

Boldo, S., & Melquiond, G. (2011). Flocq: A unified library for proving floating-point algorithms in Coq. *20th IEEE Symposium on Computer Arithmetic*, 243–252. https://doi.org/10.1109/ARITH.2011.40

Boldo, S., & Melquiond, G. (2026, 19 Φεβρουαρίου). *Flocq* (Έκδ. 4.2.2) [Λογισμικό υπολογιστή]. https://gitlab.inria.fr/flocq/flocq/-/blob/flocq-4.2.2/src/Pff/Pff2Flocq.v

Buttari, A., Liu, X., Mary, T., & Vieublé, B. (2026). Mixed precision strategies for preconditioned GMRES: A comprehensive analysis [Προδημοσίευση]. *HAL*. https://hal.science/hal-05071696v2

Carson, E., & Higham, N. J. (2018). Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM Journal on Scientific Computing*, *40*(2), A817–A847. https://doi.org/10.1137/17M1140819

Castaldo, A. M., Whaley, R. C., & Chronopoulos, A. T. (2008). Reducing floating point error in dot product using the superblock family of algorithms. *SIAM Journal on Scientific Computing*, *31*(2), 1156–1174. https://doi.org/10.1137/070679946

Chang, T.-C., Park, S., Lim, J. P., & Nagarakatte, S. (2026). FLoPS: Semantics, operations, and properties of P3109 floating-point representations in Lean [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2602.15965

Connolly, M. P., & Higham, N. J. (2023). Probabilistic rounding error analysis of Householder QR factorization. *SIAM Journal on Matrix Analysis and Applications*, *44*(3), 1146–1163. https://doi.org/10.1137/22M1514817

Damouche, N., Martel, M., Panchekha, P., Qiu, C., Sanchez-Stern, A., & Tatlock, Z. (2017). Toward a standard benchmark format and suite for floating-point analysis. *Numerical Software Verification*, *10152*, 63–77. https://doi.org/10.1007/978-3-319-54292-8_6

Darulova, E., Izycheva, A., Nasir, F., Ritter, F., Becker, H., & Bastian, R. (2018). Daisy: Framework for analysis and optimization of numerical programs. *Tools and Algorithms for the Construction and Analysis of Systems*, 270–287. https://doi.org/10.1007/978-3-319-89960-2_15

Daumas, M., & Melquiond, G. (2010). Certification of bounds on expressions involving rounded operators. *ACM Transactions on Mathematical Software*, *37*(1), 1–20. https://doi.org/10.1145/1644001.1644003

Davies, P. I., & Higham, N. J. (2003). A Schur–Parlett algorithm for computing matrix functions. *SIAM Journal on Matrix Analysis and Applications*, *25*(2), 464–485. https://doi.org/10.1137/S0895479802410815

de Moura, L., & Ullrich, S. (2021). The Lean 4 theorem prover and programming language. *Automated Deduction – CADE 28*, *12699*, 625–635. https://doi.org/10.1007/978-3-030-79876-5_37

Eisenstat, S. C., & Walker, H. F. (1996). Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, *17*(1), 16–32. https://doi.org/10.1137/0917003

Fukaya, T., Kannan, R., Nakatsukasa, Y., Yamamoto, Y., & Yanagisawa, Y. (2020). Shifted Cholesky QR for computing the QR factorization of ill-conditioned matrices. *SIAM Journal on Scientific Computing*, *42*(1), A477–A503. https://doi.org/10.1137/18M1218212

Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, *23*(1), 5–48. https://doi.org/10.1145/103162.103163

Goubault, E., & Putot, S. (2013). Robustness analysis of finite precision implementations. *Programming Languages and Systems*, *8301*, 50–57. https://doi.org/10.1007/978-3-319-03542-0_4

Gu, M., & Eisenstat, S. C. (1996). Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, *17*(4), 848–869. https://doi.org/10.1137/0917055

Hallman, E., & Ipsen, I. C. F. (2021). Deterministic and probabilistic error bounds for floating point summation algorithms [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2107.01604

Hallman, E., & Ipsen, I. C. F. (2023). Precision-aware deterministic and probabilistic error bounds for floating-point summation. *Numerische Mathematik*, *155*(1–2), 83–119. https://doi.org/10.1007/s00211-023-01370-y

Higham, D. J., & Higham, N. J. (1992). Backward error and condition of structured linear systems. *SIAM Journal on Matrix Analysis and Applications*, *13*(1), 162–175. https://doi.org/10.1137/0613014

Higham, N. J. (1993). The accuracy of floating point summation. *SIAM Journal on Scientific Computing*, *14*(4), 783–799. https://doi.org/10.1137/0914050

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms* (2η έκδοση). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718027

Higham, N. J. (2004). The numerical stability of barycentric Lagrange interpolation. *IMA Journal of Numerical Analysis*, *24*(4), 547–556. https://doi.org/10.1093/imanum/24.4.547

Higham, N. J., & Mary, T. (2022). Solving block low-rank linear systems by LU factorization is numerically stable. *IMA Journal of Numerical Analysis*, *42*(2), 951–980. https://doi.org/10.1093/imanum/drab020

Institute of Electrical and Electronics Engineers. (2019). *IEEE standard for floating-point arithmetic* (IEEE Standard Αρθμ. 754-2019). https://doi.org/10.1109/IEEESTD.2019.8766229

Kellison, A. E., & Appel, A. W. (2022). Verified numerical methods for ordinary differential equations. *Proceedings of the 15th International Workshop on Numerical Software Verification*. https://www.cs.princeton.edu/~appel/papers/VerifiedODE.pdf

Kellison, A. E., Appel, A. W., Tekriwal, M., & Bindel, D. (2023). LAProof: A library of formal proofs of accuracy and correctness for linear algebra programs. *2023 IEEE 30th Symposium on Computer Arithmetic (ARITH)*, 36–43. https://doi.org/10.1109/ARITH58626.2023.00021

Koev, P. (2007). Accurate computations with totally nonnegative matrices. *SIAM Journal on Matrix Analysis and Applications*, *29*(3), 731–751. https://doi.org/10.1137/04061903X

Kontorovich, A. (2025, 29 Σεπτεμβρίου). *The shape of math to come* [Βίντεο]. YouTube. https://www.youtube.com/watch?v=xIG1RI44Nog

Kontorovich, A. (2026). The shape of math to come. *Proceedings of the International Congress of Mathematicians 2026 – Volume 2: Plenary Lectures*, 102–123. https://doi.org/10.1137/25M1806764

Lange, M., & Oishi, S. (2020). A note on Dekker's FastTwoSum algorithm. *Numerische Mathematik*, *145*(2), 383–403. https://doi.org/10.1007/s00211-020-01114-2

Li, J.-R., & White, J. (2002). Low rank solution of Lyapunov equations. *SIAM Journal on Matrix Analysis and Applications*, *24*(1), 260–280. https://doi.org/10.1137/S0895479801384937

Liu, X. (2025). Mixed-precision Paterson–Stockmeyer method for evaluating polynomials of matrices. *SIAM Journal on Matrix Analysis and Applications*, *46*(1), 811–835. https://doi.org/10.1137/24M1675734

Long, W., Zhang, Y., Li, C., Zhou, L., Sun, C., & Wen, Z. (2026). CAM-Bench: A benchmark for computational and applied mathematics in Lean [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2605.17255

Meek, T., Ge, S., Xiang, D. Q., Chess, S., & Ilin, V. (2026). Formalizing numerical analysis: An agent pipeline and quality audit beyond kernel acceptance [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2606.14000

Nakatsukasa, Y., & Higham, N. J. (2012). Backward stability of iterations for computing the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, *33*(2), 460–479. https://doi.org/10.1137/110857544

OpenAI. (2026a). *Build skills: Give ChatGPT and Codex new capabilities and expertise*. Ανακτήθηκε 24 Αυγούστου 2026, από https://learn.chatgpt.com/docs/build-skills

OpenAI. (2026b). *GPT-5.6 Sol model*. Ανακτήθηκε 24 Αυγούστου 2026, από https://developers.openai.com/api/docs/models/gpt-5.6-sol

OpenAI. (2026c). *GPT-6 Astra model*. Ανακτήθηκε 2 Οκτωβρίου 2026, από https://developers.openai.com/api/docs/models/gpt-6-astra

OpenAI. (2026d). *GPT-6 Sol model*. Ανακτήθηκε 2 Οκτωβρίου 2026, από https://developers.openai.com/api/docs/models/gpt-6-sol

OpenAI. (2026e, 8 Σεπτεμβρίου). *On the Navier–Stokes Millennium Prize Problem*. https://openai.com/index/navier-stokes-solution/

Palomar Registry. (2026). *About Palomar: What the registry does and what it means*. Ανακτήθηκε 3 Σεπτεμβρίου 2026, από https://palomar-registry.org/about

Pham, Q., Karimov, E., Galichin, A., & Oseledets, I. (2026). TheoremBench: Evaluating LLMs on theorem proving in formal mathematics [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2606.09450

Poiroux, A., Weiss, G., Kunčak, V., & Bosselut, A. (2025). Reliable evaluation and benchmarks for statement autoformalization. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17947–17969. https://doi.org/10.18653/v1/2025.emnlp-main.907

Ramos, G. U. (1971). Roundoff error analysis of the fast Fourier transform. *Mathematics of Computation*, *25*(116), 757–768. https://doi.org/10.1090/S0025-5718-1971-0300488-0

Roux, P. (2016). Formal proofs of rounding error bounds. *Journal of Automated Reasoning*, *57*(2), 135–156. https://doi.org/10.1007/s10817-015-9339-z

Rump, S. M. (2012). Error estimation of floating-point summation and dot product. *BIT Numerical Mathematics*, *52*(1), 201–220. https://doi.org/10.1007/s10543-011-0342-4

Rump, S. M., & Jeannerod, C.-P. (2014). Improved backward error bounds for LU and Cholesky factorizations. *SIAM Journal on Matrix Analysis and Applications*, *35*(2), 684–698. https://doi.org/10.1137/130927231

Skeel, R. D. (1980). Iterative refinement implies numerical stability for Gaussian elimination. *Mathematics of Computation*, *35*(151), 817–832. https://www.jstor.org/stable/2006197

Smoktunowicz, A., Barlow, J. L., & Langou, J. (2006). A note on the error analysis of classical Gram–Schmidt. *Numerische Mathematik*, *105*(2), 299–313. https://doi.org/10.1007/s00211-006-0042-1

Solovyev, A., Baranowski, M. S., Briggs, I., Jacobsen, C., Rakamarić, Z., & Gopalakrishnan, G. (2018). Rigorous estimation of floating-point round-off errors with symbolic Taylor expansions. *ACM Transactions on Programming Languages and Systems*, *41*(1), 1–39. https://doi.org/10.1145/3230733

Srinivasan R, S., & Patawar, M. (2026). LAMP: Lean-based agentic framework with MCP and proof repair [Προδημοσίευση]. *arXiv*. https://doi.org/10.48550/arXiv.2606.28841

Tao, T. (2025). Machine-assisted proof. *Notices of the American Mathematical Society*, *72*(1), 6–15. https://doi.org/10.1090/noti3041

Tao, T. (2026, 18 Αυγούστου). *Palomar – a registry of Lean-verified mathematics*. What's New. https://terrytao.wordpress.com/2026/08/18/palomar-a-registry-of-lean-verified-mathematics/

The mathlib Community. (2020). The Lean mathematical library. *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, 367–381. https://doi.org/10.1145/3372885.3373824

Titolo, L., Moscato, M., Feliu, M. A., Masci, P., & Muñoz, C. A. (2025). Rigorous floating-point round-off error analysis in PRECiSA 4.0. *Formal Methods*, *14934*, 20–38. https://doi.org/10.1007/978-3-031-71177-0_2

Tsoukalas, G., Lee, J., Jennings, J., Xin, J., Ding, M., Jennings, M., Thakur, A., & Chaudhuri, S. (2024). PutnamBench: Evaluating neural theorem-provers on the Putnam Mathematical Competition. *Advances in Neural Information Processing Systems*, *37*, 11545–11569. https://doi.org/10.52202/079017-0368

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press. https://doi.org/10.1017/9781108231596

Wilkinson, J. H. (1963). *Rounding errors in algebraic processes*. Prentice-Hall.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2023). LeanDojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, *36*, 21573–21612. https://doi.org/10.52202/075280-0944

Zheng, K., Han, J. M., & Polu, S. (2022). miniF2F: A cross-system benchmark for formal olympiad-level mathematics. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2109.00110