

Adaptive CFK-Bäume für mehrere Gaußcluster

G3: Abstand und ungleiche Gewichte bei 100000 Teilchen

Forschungsprotokoll des Projekts CFK

17. September 2026

1 Fragestellung

Untersucht wird, ob die Blattzahl pro Teilchen beim Übergang von einer zu mehreren konzentrierten Punktwolken annähernd konstant bleibt. Anders als in den vorangehenden Studien wird ausschließlich $n = 10^5$ verwendet. Primäre Zielgröße ist L/n ; sekundär werden Konformitätsaufwand, globale Tiefenverteilungen und die Auflösung jeder erzeugenden Komponente betrachtet. Die Arbeitshypothese erwartet größere Veränderungen in der Verteilung der Auflösung als in der gesamten Blattzahl. Es wurde vorab keine quantitative Äquivalenzschwelle für „annähernd konstant“ festgelegt. Eine solche Grenze wird nicht nach Kenntnis der Ergebnisse als bestandener Test ausgegeben.

Wurzeldreieck, Integercodierung, Hypotenusenbisektion und minimaler Konformitätsabschluss sind unverändert:

$$T_1 = \text{conv}\{(0, 0), (2, 0), (0, 2)\}, \quad N = 2L - 1, \quad E = L - 1 - S.$$

S zählt notwendige Teilchentrennungen, E die zusätzlichen Teilungen des Konformitätsabschlusses. Das Netz bedeckt weiterhin die gesamte Wurzel, die Teilchen liegen im Einheitsquadrat.

2 Vorab festgelegtes Design

Die isotrope Komponentenbreite ist $\sigma = 0,025$. Es werden 22 Konfigurationen in zehn unabhängigen gepaarten Blöcken berechnet, insgesamt 220 Produktionsbäume.

Komponenten	Abstand d/σ	Gewichte [%]
1	0	100
2	2, 4, 8, 16	50:50; 90:10; 99:1
4	4, 8, 16	25:25:25:25; 70:10:10:10; 97:1:1:1

Bei vier Komponenten ist d die Seitenlänge des Quadrats ihrer Zentren; diagonale Abstände sind $\sqrt{2}d$. Jede Komponente enthält mindestens 1000 Teilchen.

2.1 Feste Besetzung statt zufälliger Clusterzahlen

Für Gewichte w_j werden exakt $n_j = nw_j$ Teilchen der Komponente j zugewiesen. Innerhalb einer Komponente gelten unabhängige Normaloffsets mit Kovarianz $\sigma^2 I_2$. Die erwartete gepoolte räumliche Dichte lautet

$$f(x) = \sum_{j=0}^{k-1} w_j \frac{\exp(-\|x - \mu_j\|^2 / (2\sigma^2))}{2\pi\sigma^2}.$$

Die Gesamtstichprobe hat feste Komponentenbesetzungen; sie ist nicht durch unabhängige Ziehung zufälliger multinomialer Labels entstanden. Diese zusätzliche Besetzungsstreuung wird bewusst ausgeschlossen.

2.2 Paarung und geometrische Lage

Jeder Block erzeugt eine neue Folge von 100000 Gaußoffsets. Dieselben Offsets werden über alle 22 Konfigurationen nach Teilchenindex wiederverwendet. Aufeinanderfolgende Indexintervalle bilden die Komponenten. Bei Änderung der Gewichte wechseln einige Teilchen die Komponente; bei Änderung des Abstands bleiben die Labels für feste Gewichte erhalten.

Der geometrische Mittelpunkt m wird je Block gleichverteilt in $[0,49; 0,51]^2$, der Winkel θ gleichverteilt in $[0, 2\pi)$ gezogen. Die lokalen Zentren sind

$$(-d/2, 0), (d/2, 0)$$

beziehungsweise

$$(-d/2, -d/2), (d/2, -d/2), (d/2, d/2), (-d/2, d/2).$$

Sie werden mit derselben Rotation und Verschiebung für den ganzen Block transformiert. Bei ungleichen Gewichten ist Komponente 0 die dominierende. Die Referenz liegt am geometrischen Mittelpunkt, nicht am gewichteten Schwerpunkt der Mischung. Gewichtsänderungen verschieben daher auch den Massenschwerpunkt; dies gehört zu diesem Versuchsdesign.

Rotationskoordinaten werden einmal in Gleitkommaarithmetik berechnet und anschließend als exakte rationale Werte festgehalten. Die ebenfalls rational repräsentierten Offsets werden exakt addiert. Die Offsets selbst werden nicht rotiert; ihre Verteilung ist isotrop. Alle Zentren liegen mindestens acht Standardabweichungen vom Rand. Kein Punkt wird abgeschnitten, verworfen oder nachgezogen. Jeder Grenzübertritt hätte die Serie für eine Designprüfung angehalten.

2.3 Seeds und Reihenfolge

Der Seed ist die Big-Endian-Interpretation des SHA-256-Digests von

```
cfk-mixture-g3:20260917:repetition:stream
```

mit den getrennten Streams `offsets`, `pose` und `order`. Verwendet wird `random.Random`; je Teilchen werden nacheinander die beiden Gaußkoordinaten gezogen. Die Konfigurationsreihenfolge wird pro Block reproduzierbar permutiert. Die Metadaten enthalten alle realisierten Lagen und den Quellstand. Die zehn Blöcke, nicht die 220 Bäume, sind die unabhängigen Einheiten für gepaarte Vergleiche.

3 Implementierung und Prüfungen

Der getestete verteilte globale Konformitätsabschluss verwendet höchstens vier vorbereitete Prozesse. Jeder Block erhält einen frischen Hauptprozess; sein Pool wird für die Konfigurationen wiederverwendet. Die einzige Anpassung am bisherigen Wächter ist ein wählbarer Programmeinstieg. Die eigentlichen Gitteralgorithmen bleiben unverändert; bestehende Versuchsarchive werden nicht verändert.

Vor Produktionsbeginn bestanden 57 Tests. Die neuen Tests prüfen alle Konfigurationen gegen die rationale Referenz, feste Besetzungen, exakte Paarung, Abstände, Komponentenhistogramme und Fehlerfälle. Ein kleiner technischer Gesamtbaulauf prüfte den Versuchscontroller; er gehört nicht zur wissenschaftlichen Serie. Jeder Produktionsbaum erhält leichte Invariantenprüfungen, zusätzlich muss die Summe der Komponentenhistogramme exakt dem globalen Histogramm besetzter Blätter entsprechen. Auf eine unabhängige Vollvalidierung jedes großen Baums wird wie vereinbart verzichtet.

Die Aufbauzeit enthält Normalisierung, Präfixbildung, Abschluss und Merge, aber nicht Verschiebung, Prüfsummen, Poolstart oder Kennzahlen. Die Kennzahlenzeit enthält hier zusätzliche Blattsuchen je Komponente. Laufzeiten sind Begleitgrößen, keine neue Optimierungsstudie. Der Speicherwächter begrenzt die Prozessfamilie auf 6 GiB RSS, bei mindestens 2 GiB verbleibendem Systemspeicher; pro Prozess gilt ein Adressraumlimit von 6 GiB. Das Zeitlimit je Block beträgt 1800 Sekunden.

4 Kennzahlen und Auswertung

Erfasst werden L/n , S , E , E/S , $E/(L-1)$, die vollständigen Tiefenhistogramme aller, besetzter und leerer Blätter, sowie Mittelwert, Quartile, 95-/99-Prozent-Quantile und maximale besetzte Tiefe jeder erzeugenden Komponente. Bei überlappenden Wolken ist dieses Label keine räumliche Gebietszuordnung. Leere Blätter besitzen kein erzeugendes Label und werden nicht künstlich einer Komponente zugeteilt.

Primärer gepaarter Vergleich für Konfiguration a und Block r :

$$Q_{a,r} = \frac{L_{a,r}}{L_{\text{single},r}}, \quad \Delta_{a,r} = 100(Q_{a,r} - 1) \%.$$

Die Prozentwerte im Bericht sind Mittelwerte dieser gepaarten Änderungen, nicht Quotienten zweier nachträglicher Gesamtmittel. Für E/S wird zusätzlich die gepaarte Differenz gespeichert. Ein Blockbootstrap mit 10000 Ziehungen vollständiger Blöcke, Analyse-Seed 2026091704, liefert punktweise 95-Prozent-Perzentilbereiche. Sie sind explorativ, nicht simultan über alle Konfigurationen abgesichert. Es gibt keine nachträgliche Auswahl günstiger Wiederholungen.

4.1 Heuristik für gut getrennte Komponenten

Für eine ausreichend isolierte Komponente legt die lokale Dichte nahe

$$\bar{d}_j - \bar{d}_{\text{single}} \approx \log_2 w_j.$$

Teilchengewichtet ergibt sich damit näherungsweise

$$\bar{d} - \bar{d}_{\text{single}} \approx \sum_j w_j \log_2 w_j = -H_2(w).$$

Gleiche Aufteilung auf zwei bzw. vier getrennte Komponenten würde also etwa eine bzw. zwei Ebenen geringere mittlere besetzte Tiefe bedeuten. Für eine Ein-Prozent-Komponente wären es etwa 6,64 Ebenen. Dies ist eine Interpretationshilfe, kein exakter Satz über Konformitätsabschluss, endliche Stichproben oder überlappende Cluster.

5 Ergebnisse

k	d/σ	Anteile [%]	L/n	Änderung [%]	E/S
1	0	100	3,298	0,000	1,285
2	2	50:50	3,306	0,232	1,289
2	2	90:10	3,293	-0,151	1,283
2	2	99:1	3,294	-0,143	1,283
2	4	50:50	3,307	0,264	1,289
2	4	90:10	3,305	0,202	1,288
2	4	99:1	3,301	0,067	1,286
2	8	50:50	3,310	0,345	1,289
2	8	90:10	3,310	0,363	1,290
2	8	99:1	3,307	0,249	1,288
2	16	50:50	3,304	0,164	1,287
2	16	90:10	3,308	0,305	1,289
2	16	99:1	3,306	0,216	1,288
4	4	25:25:25:25	3,301	0,082	1,286
4	4	70:10:10:10	3,308	0,280	1,290
4	4	97:1:1:1	3,305	0,203	1,288
4	8	25:25:25:25	3,309	0,329	1,290
4	8	70:10:10:10	3,311	0,380	1,293
4	8	97:1:1:1	3,313	0,444	1,294
4	16	25:25:25:25	3,312	0,423	1,293
4	16	70:10:10:10	3,310	0,363	1,291
4	16	97:1:1:1	3,308	0,295	1,291

Die größte absolute Änderung eines mittleren gepaarten Blattverhältnisses beträgt 0,444 %. Dies ist ein nachträglich ausgewähltes Maximum, kein unverzerrter Schätzer eines universellen Worst Case.

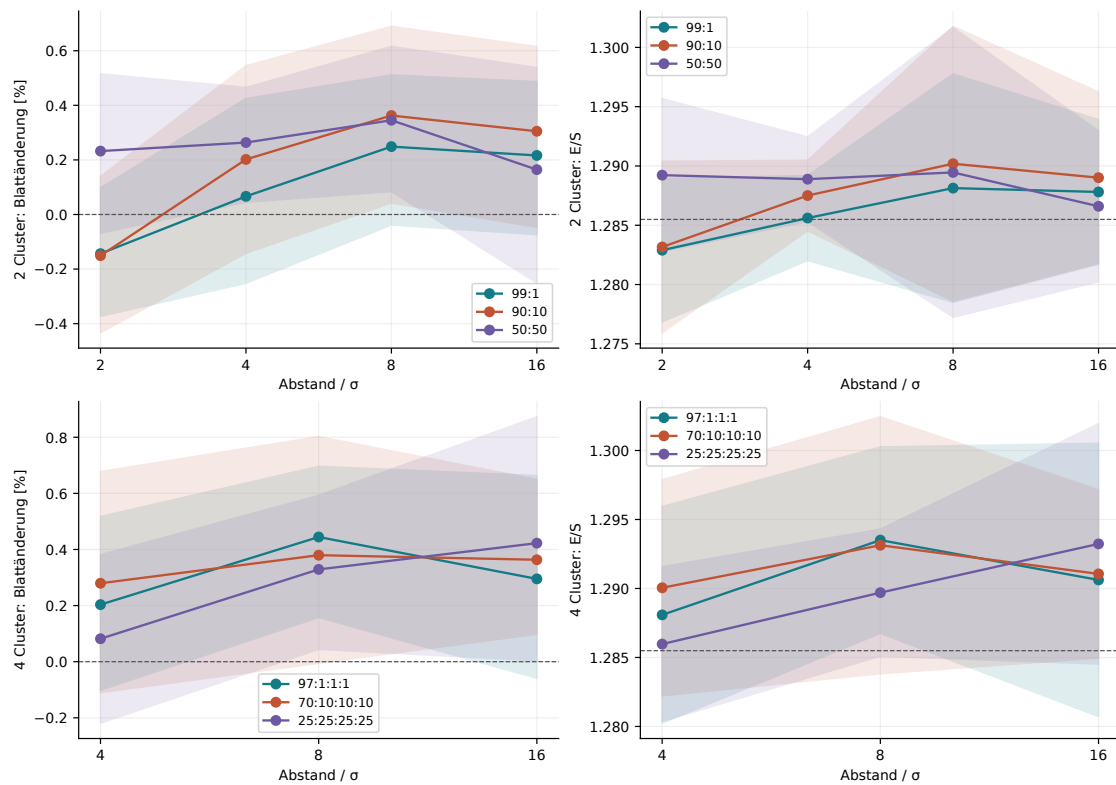


Abbildung 1: Links: mittlere gepaarte Blattänderung zur Ein-Cluster-Referenz; Bänder sind punktweise 95-Prozent-Blockbootstrapbereiche. Rechts: E/S mit einer Standardabweichung zwischen Blöcken; gestrichelte Linie ist der Referenzmittelwert.

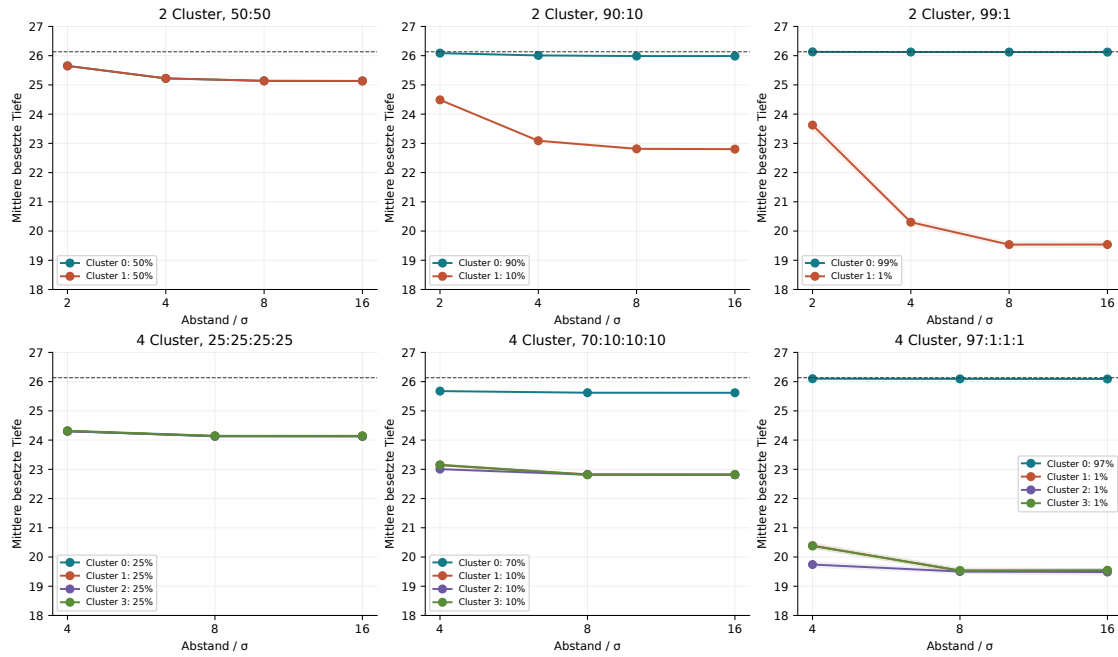


Abbildung 2: Mittlere besetzte Tiefe nach erzeugender Komponente; Bänder zeigen Standardabweichungen. Gestrichelte Linie: Ein-Cluster-Referenz. Clusterlabels sind bei überlappenden Verteilungen keine getrennten räumlichen Gebiete.

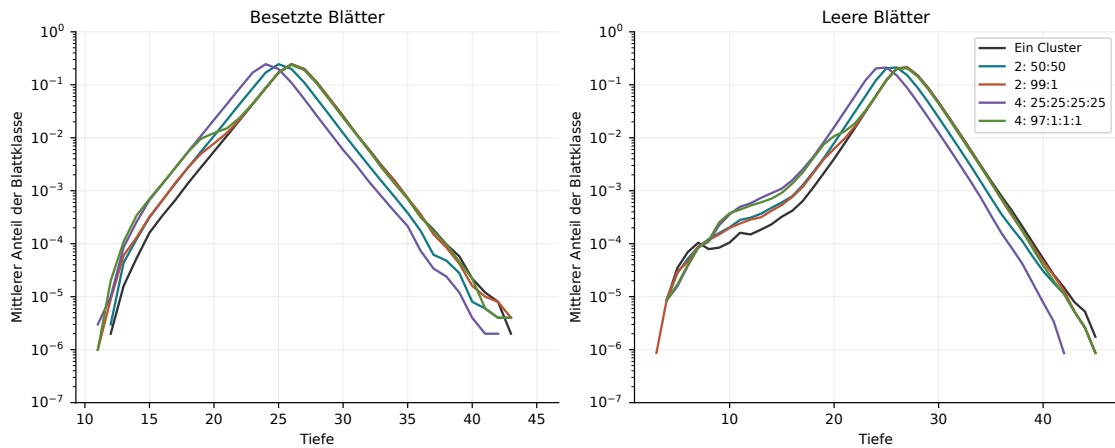


Abbildung 3: Tiefenprofile der Referenz und der vier ausgewählten weit getrennten Konfigurationen. Je Block wird innerhalb der Blattklasse normiert, danach über die zehn Blöcke gemittelt. Logarithmische Ordinate; Nullwerte werden nicht dargestellt.

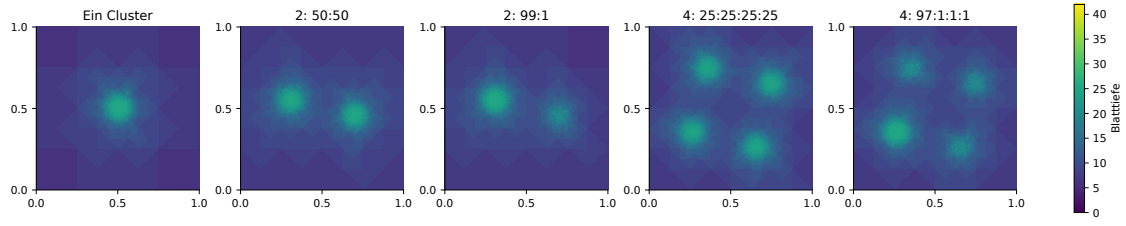


Abbildung 4: Vorab ausgewählte Bäume aus Block 1 bei 100000 Teilchen. Gleicher Ausschnitt des Einheitsquadrats und gemeinsame Tiefenskala. Alle Blätter werden eingefärbt, ohne Kantenlinien; das vollständige Wurzeldreieck reicht über den Ausschnitt hinaus.

6 Einordnung

Alle 220 geplanten Produktionsbäume wurden erfolgreich berechnet, ohne Grenzübertritt oder Abbruch. Die Serie benötigte etwa 22,1 Minuten einschließlich Kennzahlen und Ablaufverwaltung, ohne die nachfolgende Berichtserstellung. Alle archivierten Eingabeprüfsummen und Paarungsbeziehungen wurden vor der Auswertung geprüft.

6.1 Kleine Veränderungen der Blattzahl

Die Konfigurationsmittel von L/n reichen von 3,293470 bis 3,313096; die Ein-Cluster-Referenz liegt bei 3,298470. Der größte Betrag einer mittleren gepaarten Änderung beträgt 0,4444 %. Er tritt bei vier Komponenten, $d/\sigma = 8$ und Gewichten 97:1:1:1 auf. Der zugehörige punktweise Bootstrapbereich reicht ungefähr von 0,158 bis 0,696 %. Die mittleren Werte von E/S liegen zwischen etwa 1,283 und 1,294.

Die Blattzahlen verhalten sich damit über die untersuchten Geometrien und Gewichte tatsächlich sehr regelmäßig. Dies bedeutet nicht, dass alle Konfigurationen dieselbe erwartete Blattzahl besitzen. Einige punktweise Bootstrapbereiche schließen die Nulländerung aus; kleine systematische Mehrbedarfe sind mit den Daten vereinbar. Die Intervalle sind weder simultan korrigiert noch ein vorab spezifizierter Äquivalenztest. Aus dem größten beobachteten Mittelwertunterschied folgt keine universelle Grenze für andere Mischungen, Breiten oder Clusterzahlen.

6.2 Deutliche Änderungen der Auflösung

Die mittlere besetzte Tiefe der Referenz beträgt 26,133147. Bei $d/\sigma = 16$ sinkt sie für zwei gleich gewichtete Komponenten gegenüber der gepaarten Referenz um 0,997842, für vier um 1,997462 Ebenen. Die lokale Dichteheuristik von einer bzw. zwei Ebenen beschreibt diese Fälle sehr gut.

Bei zwei Komponenten mit Gewichten 99:1 und $d/\sigma = 16$ hat die dominante Komponente im Mittel Tiefe 26,121828, die kleine dagegen 19,538800. Das teilchengewichtete Gesamtmittel unterscheidet sich trotzdem nur um $-0,077149$ Ebenen von der Referenz. Das zeigt, wie leicht die Auflösung eines kleinen Nebenclusters im Gesamtmittel unsichtbar bleibt.

Mit zunehmendem Abstand nimmt die Tiefenwirkung der Überlappung ab. Insbesondere werden kleine Komponenten in der Nähe eines dominanten Clusters feiner aufgelöst als isolierte Komponenten derselben Größe, weil dort auch die anderen Teilchen zur lokalen Gesamtdichte beitragen. Bei vier Komponenten sind gegenüberliegende und benachbarte Nebencluster bei endlichem Abstand nicht geometrisch gleich gelegen. Ihre Herkunftslabels werden deshalb einzeln ausgewiesen.

6.3 Grenzen und nächste Fragestellung

Die Studie untersucht eine feste Gesamtteilchenzahl, gleiche isotrope Breiten und höchstens vier Komponenten mit mindestens einem Prozent Gewicht. Sie beantwortet nicht das Verhalten extrem kleiner Satelliten, stark verschiedener Breiten, anisotroper Wolken, randnaher Zentren oder sehr vieler Cluster. Lage und Orientierung ändern sich gemeinsam mit der Punktwolke zwischen Blöcken; ihre einzelnen Varianzbeiträge werden nicht getrennt. Bei ungleichen Gewichten verschiebt sich außerdem der Massenschwerpunkt relativ zur geometrischen Referenz.

Ein nächster kontrollierter Schritt wären unterschiedliche Komponentenbreiten bei festgehaltener Geometrie und Besetzung. Damit ließe sich untersuchen, ob sehr dichte kleine Komponenten die hier beobachtete Trennung zwischen fast konstanter Blattzahl und stark veränderter lokaler Auflösung beibehalten. Eine größere Teilchenzahl ist hierfür zunächst nicht erforderlich.

7 Archiv und Reproduktion

Die primären Einzelprotokolle mit Histogrammen stehen unter `data/trials/`; `data/trials.csv`, `data/clusters.csv` und `data/summary.csv` erleichtern die Weiterverarbeitung. `manifest.json` sichert die archivierten Eingaben; `results.json` enthält die berechneten Vergleiche und Intervalle. Quellen, Seedschema, konkrete Lagen und Eingabeprüfsummen erlauben die Replikation in der dokumentierten Umgebung.

Die Auswertung ohne neue Experimente wird im Berichtsverzeichnis mit folgenden Befehlen erneut erzeugt:

```
MPLCONFIGDIR=/tmp/cfk-matplotlib python3 analyze.py
latexmk -pdf -interaction=nonstopmode -halt-on-error Bericht.tex
```

Für eine neue Ausführung derselben Serie mit archivierten Quellen:

```
python3 data/source/benchmark_mixture.py --output /tmp/cfk-g3-replay
```

Das Ausgabeverzeichnis muss neu sein; bestehende Archive werden nicht überschrieben. Beispiellabbildungen zeigen vorab festgelegte Bäume aus Block 1, keine nach Sichtung ausgewählten Extremfälle.