Gaußverteilte Teilchen und die Struktur adaptiver CFK-Bäume

G1: Zentrierte Verteilungen mit unterschiedlicher Konzentration

Forschungsprotokoll des Projekts CFK

17. September 2026

1 Fragestellung und Umfang

Nach den Gleichverteilungs- und Skalierungsversuchen vom 16. September steht nun nicht die Codegeschwindigkeit, sondern die Morphologie der Bäume im Vordergrund. Untersucht werden die Zahl und Tiefe ihrer Blätter und die zusätzlich durch Konformität erforderlichen Teilungen. Alle Versuche verwenden dieselbe Wurzel

$$T_1 = \mathrm{conv}\{(0,0),(2,0),(0,2)\}, \qquad |T_1| = 2,$$

dieselbe Integercodierung und dieselbe Hypotenusenbisektion. Teilchen liegen ausschließlich im Einheitsquadrat $Q = [0,1]^2$. Das Netz bedeckt weiterhin das gesamte Wurzeldreieck, einschließlich des unbesetzten Bereichs außerhalb des Quadrats.

Diese erste Etappe fixiert das Zentrum bei $(0{,}5, 0{,}5)$ und variiert die Breite. Verschobene Zentren und gemeinsam variierende Zentren/Breiten sind ausdrücklich spätere Studien. Bestehende Versuchsarchive werden nicht verändert. Die bereitgestellte LaTeX-Vorlage wird wie im Hauptbericht auf eine wissenschaftliche Dokumentation ohne Aufgaben-/Lösungsteil angepasst.

2 Verteilung und vorab festgelegtes Design

2.1 Bedingte Normalverteilung

Die Ausgangsverteilung ist isotrop,

$$X \sim \mathcal{N}(\mu, \sigma^2 I_2), \quad \mu = (0{,}5, 0{,}5), \quad \sigma \in \{0{,}025;\ 0{,}05;\ 0{,}1;\ 0{,}2;\ 0{,}4\}.$$

Ein Koordinatenpaar außerhalb Q wird vollständig verworfen und neu gezogen. Es werden genau n akzeptierte Punkte erzeugt. Die Dichte der tatsächlichen Eingabe ist damit

$$f_Q(x) = \frac{\exp(-\|x-\mu\|^2/(2\sigma^2))}{2\pi\sigma^2 Z_\sigma}\,\mathbf{1}_Q(x), \qquad Z_\sigma = \left[2\Phi\left(\frac{1}{2\sigma}\right) - 1\right]^2.$$

Es erfolgt kein Abschneiden auf den Rand, keine periodische Fortsetzung und keine Entfernung doppelter Punkte. Duplikate wären wie zuvor ein Eingabefehler. σ bezeichnet die Breite der unbeschränkten Ausgangsverteilung, nicht die empirische Streuung der akzeptierten Punkte. Insbesondere verändert die Konditionierung die breite Verteilung deutlich.

Für eine Koordinate folgt durch Integration mit $a = 1/(2\sigma)$

$$\mathrm{Var}(X_j \mid X \in Q) = \sigma^2 \left(1 - \frac{2a\varphi(a)}{2\Phi(a) - 1}\right).$$

Die Symmetrie erhält den theoretischen Mittelwert 0,5. Dabei bezeichnen Φ und φ Verteilungsfunktion und Dichte der eindimensionalen Standardnormalverteilung. Diese Formeln dienen zur Interpretation; die tatsächlichen Mittelwerte, Streuungen und Verwerfungsquoten werden in jedem Versuch gemessen.

2.2 Versuchszahlen und Zufall

Für jede der sechs Größen $n = 10, 100, 1000, 10^4, 10^5, 10^6$ sind 25 Wiederholungen bei $\sigma = 0{,}1$ und je zehn Wiederholungen bei den übrigen vier Breiten vorgesehen. Das ergibt

$$6(25 + 4 \cdot 10) = 390 \quad \text{Bäume.}$$

Der Ausgangsfall wird innerhalb der Breitenstudie nicht doppelt gemessen. Es gibt keine nachträgliche Auswahl von Wiederholungen nach ihrem Ergebnis. Die unterschiedlichen Wiederholungszahlen sind bei der Interpretation von Streuungen und Unsicherheit zu beachten. Die Studie ist explorativ; zehn oder 25 Wiederholungen garantieren keine vorgegebene Genauigkeit für seltene Extremereignisse wie die maximale Tiefe.

Der Master-Seed ist 20260917. Der konkrete Seed ist die Big-Endian- Interpretation des SHA-256-Digests der ASCII-Zeichenfolge

```
cfk-gaussian-g1:20260917:n:sigma_label:repetition
```

mit den wörtlichen Breitenlabels `0.025`, `0.05`, `0.1`, `0.2`, `0.4`, dezimalen Größen und Wiederholungsindizes ab 1. `random.Random.gauss` erzeugt nacheinander beide Koordinaten. Damit sind die Seeds für verschiedene Größen, Breiten und Wiederholungen verschieden. Es handelt sich nicht um gepaarte oder geschachtelte Punktmengen über verschiedene Breiten oder Größen hinweg. Genaue Reproduktion bezieht sich auf die archivierte Python-/Systemumgebung.

Die Größen werden aufsteigend bearbeitet; die Reihenfolge der Breiten wird pro Wiederholung zyklisch verschoben. Sobald zehn Wiederholungen der vier Nebenbreiten abgeschlossen sind, folgen nur die restlichen Wiederholungen des Ausgangsfalls. Dies ist kein vollständig randomisiertes Laufzeitdesign. Da Baumkennzahlen bei identischen Koordinaten deterministisch sind, beeinflusst die Reihenfolge deren Werte nicht.

3 Algorithmus und Prüfumfang

Die Fassung `particles_distributed.py` verwendet die bereits getestete Integer-Präfixarithmetik und globale Konformitätsabschlüsse. Sind S_i die Teilungsanforderungen einzelner Präfixgruppen, gilt in dieser Hierarchie

$$\mathcal{C}\left(\bigcup_i S_i\right) = \bigcup_i \mathcal{C}(S_i).$$

Jede Gruppe liefert daher eine global konforme Knotenmenge, die zentral vereinigt wird. Die vollständige mathematische Herleitung steht im Hauptbericht vom Vortag. Grobe Anforderungen aus der Bucketbildung werden ebenfalls geschlossen. Die notwendigen Teilungszahlen der disjunkten Präfixbereiche können addiert werden.

Bis 2048 Teilchen wird seriell gerechnet, darüber mit höchstens vier vorbereiteten Prozessen. Jeder Versuch erhält einen frischen Hauptprozess. Die schlanke Fassung verzichtet auf Jobzeitstrahlen, geschätzte Serialisierungsgrößen und die vollständige Nachprüfung jedes großen Baums. Zwischenergebnisse werden nach der Vereinigung freigegeben. Die zugrunde liegenden Geometrie-, Präfix- und Abschlussregeln bleiben unverändert. Die neue Fassung wurde vor Beginn gegen die rationale Referenz getestet; insgesamt bestanden 49 Tests. Eine separate Vorabprobe mit 10.000 Teilchen je Breite gehört nicht zu den 390 Produktionsmessungen.

Gemessen wird auf dem Intel Core Ultra 7 255U mit 12 physischen Kernen, 14 logischen CPUs und ungefähr 16 GB RAM unter Linux und Python 3.13.9. Die detaillierte Prozessorabfrage liegt in `hardware.txt`; Python-, Plattform- und Überwachungsbibliotheksversion sind außerdem in den Studienmetadaten archiviert.

Jeder Lauf prüft die vollständige binäre Knotenzahl, die exakte Flächensumme und einfache Histogrammbeziehungen. Nur Wiederholung 1 mit $n \leq 1000$ erhält zusätzlich eine vollständige strukturelle Konformitätsprüfung und einen Test der Eindeutigkeit der Teilchen-Blatt-Zuordnung. Die Statistik der besetzten Blätter erfordert eine tatsächliche Integer-Blattsuche, aber keine unabhängige geometrische Vollvalidierung. Die Eigenschaft höchstens eines Teilchens je Blatt wird für die großen Messungen aus dem getesteten Algorithmus vorausgesetzt.

Die Aufbauzeit enthält Eingabenormalisierung, Packung, Präfixarbeit, Konformität und Merge. Ziehung, Poolstart, Kennzahlen, Kontrollprüfungen und Dateiausgabe sind getrennt erfasst. Aufbauzeiten dieser Fassung sind nicht ohne Weiteres mit den früheren instrumentierten Messungen vergleichbar und sind hier nur Begleitgrößen.

Der Speicherwächter begrenzt die RSS-Summe der Versuchsprozesse auf höchstens 6 GiB und reserviert mindestens 2 GiB systemweit verfügbaren Speicher. Zusätzlich gilt pro Prozess ein entsprechendes Adressraumlimit. Die RSS-Abtastung erfolgt alle 0,2 Sekunden; kurzzeitige Spitzen können unterschätzt und gemeinsame Seiten mehrfach gezählt werden. Nach 1800 Sekunden oder einem Fehler wird der Versuch abgebrochen und die Serie unterbrochen, ohne stille Wiederholung.

4 Baumkennzahlen

Mit L Blättern, N Knoten, S notwendigen Teilchentrennungen und E zusätzlichen Konformitätsteilungen werden erfasst:

$$N = 2L - 1, \qquad E = L - 1 - S, \qquad L/n, \qquad E/S, \qquad E/(L-1).$$

Es werden keine unnötigen Teilungen außerhalb des minimalen Konformitätsabschlusses eingeführt. „Zusätzlich“ bedeutet daher zusätzlich zur reinen Teilchentrennung, nicht algorithmisch überflüssig.

Drei Histogramme unterscheiden alle, teilchentragende und leere Blätter. Für die besetzten Blätter zählt jedes Teilchen genau einmal; bei den allgemeinen Blättern zählt dagegen auch jede leere Zelle. Mittelwerte dieser beiden Verteilungen beantworten unterschiedliche Fragen. Leere Blätter und Konformitätsaufwand sind nicht dasselbe: Schon die reine Trennung von Teilchen kann leere Geschwister erzeugen. Deshalb darf $L - n$ nicht als Anzahl zusätzlicher Konformitätsteilungen interpretiert werden; dafür ist ausdrücklich E vorgesehen. Quantile verwenden den kleinsten ganzzahligen Tiefenwert, bei dem mindestens der geforderte Anteil erreicht wird (Nearest-Rank-Konvention). Neben vollständigen Histogrammen werden Quartile, 95-/99-Prozent-Quantile und Maximum gespeichert.

Da die Fläche eines Blatts auf Tiefe d gleich 2^{1-d} ist, legt eine lokale Dichteheuristik für wenig abgeschnittene Gaußverteilungen nahe:

$$d_{\mathrm{typ}} \approx \log_2 n - 2\log_2 \sigma + C.$$

Eine Halbierung der Breite könnte damit etwa zwei zusätzliche Ebenen erfordern, ohne die Blattzahl proportional zu vergrößern. Dies ist eine Interpretationshypothese, keine exakte Formel für den zufälligen Baum, seine maximale Tiefe oder seinen Konformitätsabschluss. Randkonditionierung, die Einbettung in die feste Wurzel und die diskrete Hierarchie können davon abweichen.

5 Ergebnisse und Darstellung

Alle 390 vorgesehenen Bäume wurden erzeugt. 15 vorab festgelegte kleine Kontrollfälle wurden vollständig nachgeprüft; alle Läufe bestanden die leichten Invariantenprüfungen. Die folgenden Streubereiche und Fehlerbänder sind Stichprobenstandardabweichungen zwischen Wiederholungen, keine Konfidenzintervalle.

5.1 Blattzahl und Konformitätsaufwand

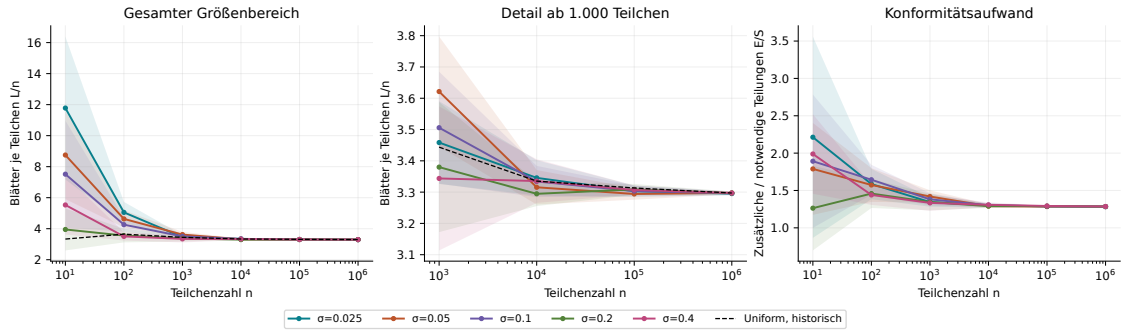

Abbildung 1: Links: gesamter Bereich. Mitte: Detail der großen Punktmengen. Rechts: zusätzlicher Konformitätsaufwand. Gleiche Breite hat in allen Abbildungen dieselbe Linienfarbe.

Zusammenfassung für $n = 1000000$:

σ	r	L/n	E/S	$\bar{d}_{\text{alle}}$	$\bar{d}_{\text{besetzt}}$	Bereich $d_{\max}$
0,025	10	$3{,}295 \pm 0{,}005$	1,284	29,899	29,459	45–55
0,05	10	$3{,}298 \pm 0{,}005$	1,285	27,901	27,458	43–50
0,1	25	$3{,}296 \pm 0{,}006$	1,284	25,901	25,458	41–50
0,2	10	$3{,}297 \pm 0{,}005$	1,285	24,065	23,622	40–46
0,4	10	$3{,}297 \pm 0{,}004$	1,286	23,412	22,977	40–45

Bei $n = 10$ reichen die Breitenmittel von 3,95 bis 11,78 Blättern je Teilchen; bei $n = 1000000$ von 3,295 bis 3,298. Die Konzentration wirkt sich damit auf kleine Bäume anders aus als auf große. Eine nahe beieinanderliegende Punktmenge muss bis in einen kleinen Teil des festen Wurzelgebiets erschlossen und konform eingebettet werden. Dieser geometrische Aufwand kann pro Teilchen bei kleinen Stichproben erheblich sein. Die gemessenen großen Bäume allein beweisen keine universelle asymptotische Konstante.

5.2 Tiefen statt nur Zellzahlen

Bei der Halbierung von $\sigma = 0.05$ auf 0.025 beträgt die Differenz der mittleren besetzten Tiefe bei $n = 1000000$ 2,001 Ebenen.

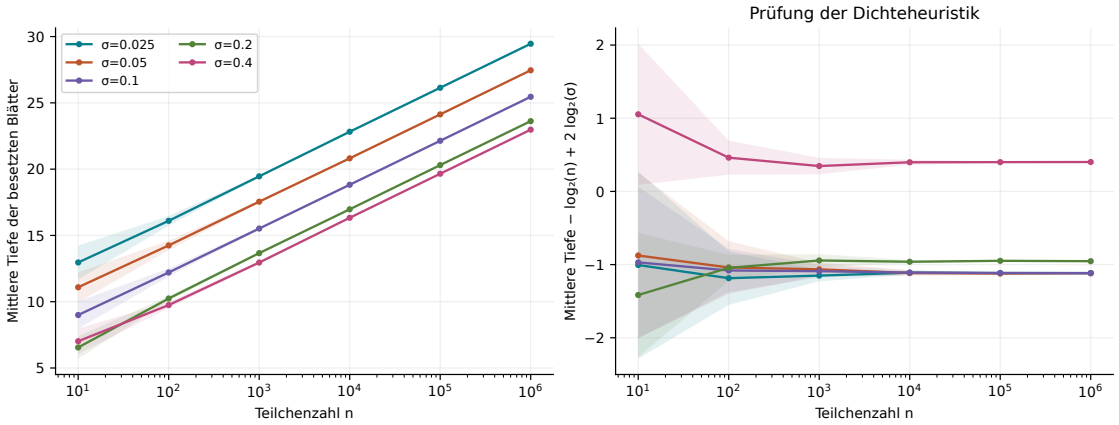

Abbildung 2: Besetzte Blätter: absolute Mittelwerte und die um Größen- und Breiteneffekt korrigierte Tiefe. Ein Zusammenfallen der schmalen Verteilungen unterstützt die Dichteheuristik, ist aber kein Beweis.

Bei der Halbierung von $\sigma = 0.1$ auf 0.05 beträgt die Differenz der mittleren besetzten Tiefe bei $n = 1000000$ 2,000 Ebenen.

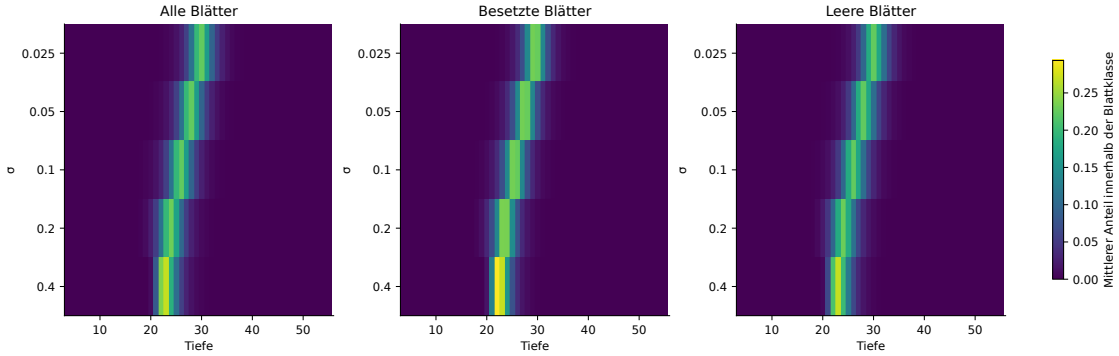

Abbildung 3: Tiefenverteilungen bei einer Million Teilchen. Pro Wiederholung wird jedes Histogramm innerhalb seiner Blattklasse auf Summe 1 normiert; anschließend werden diese Anteile gemittelt. Alle drei Tafeln verwenden dieselbe Farbskala.

Die Unterscheidung zwischen allen und besetzten Blättern ist wesentlich: Eine flächenmäßig große, leere Umgebung kann viele gröbere Zellen enthalten, ohne die typische Auflösung an den Teilchen zu beschreiben. Die Maximalwerte werden separat gespeichert; sie reagieren stark auf einzelne nahe Punktpaare und sind nicht mit einer typischen Tiefe gleichzusetzen. Leere Blätter sind allerdings nicht grundsätzlich gröber: Konformität und leere Geschwister können auch viele feine unbesetzte Zellen erzeugen. Deshalb darf aus einem großen leeren Flächenanteil nicht auf eine geringe mittlere Tiefe aller leeren Blätter geschlossen werden.

5.3 Räumliche Beispiele und geeignete Ansichten

Für die weitere gemeinsame Auswertung empfehle ich die Kombination aus den Strukturkurven und den drei Tiefen-Heatmaps. Die Kurven beantworten, wie teuer ein Baum ist; die Histogramme zeigen, wo sich seine Auflösung hin verschiebt. Die räumlichen Beispiele erklären die Geometrie, sind aber für quantitative Aussagen weniger geeignet als die wiederholten Messungen.

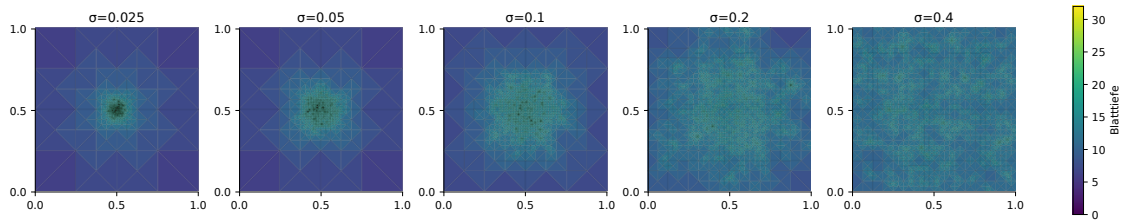

Abbildung 4: Vorab festgelegte erste Wiederholung mit 1000 Teilchen je Breite. Gezeigt wird immer dasselbe Einheitsquadrat, nicht die gesamte Wurzel. Farbe bezeichnet die Tiefe, schwarze Linien die Zellkanten. Es handelt sich um einzelne Beispiele, nicht um gemittelte Gitter.

5.4 Tatsächliche Streuung und Randkonditionierung

σ	s_x gemessen	s_x theoretisch	Verworfen [%]	Theorie [%]
0,025	0,02499	0,02500	0,0000	0,0000
0,05	0,05000	0,05000	0,0000	0,0000
0,1	0,09999	0,10000	0,0001	0,0001
0,2	0,19094	0,19092	2,4604	2,4684
0,4	0,25949	0,25955	37,7672	37,7952

Die Unterschiede der breiten Verteilungen zur einfachen Skalierungsheuristik müssen auch vor dem Hintergrund dieser Konditionierung betrachtet werden. Die empirischen Mittelwerte und beide Koordinatenstreuungen jedes Laufs bleiben in den Rohdaten erhalten.

6 Einordnung und nächste Etappe

Die breite konditionierte Gaußverteilung ist nicht identisch mit einer Gleichverteilung. Historische Uniformdaten liefern einen sinnvollen Orientierungswert für die Baumkennzahlen, sind aber kein neuer zeitgleicher Kontrollarm. Bei $n \leq 10^5$ enthalten sie jeweils 25 Wiederholungen; bei 10^6 nur drei. Es werden keine alten und neuen Laufzeiten vermischt. Für teilchentragende Tiefen fehlen historische Uniformhistogramme; dafür wird ausdrücklich keine Vergleichskurve erfunden.

Die erste Etappe isoliert den Einfluss der Breite bei einem festen Zentrum. Das Zentrum $(0{,}5, 0{,}5)$ ist zugleich speziell zur dyadischen Hierarchie gelegen. Eine spätere Zentrenstudie soll klären, welche Effekte auf Konzentration, räumliche Lage zur Hierarchie und Randnähe zurückgehen. Die hier gefundenen Zusammenhänge sind nicht automatisch auf anisotrope Gaußverteilungen, Mischverteilungen oder andere Domänen übertragbar. Auch die Breitenwahl ist zu beachten: Benachbarte Werte unterscheiden sich exakt um Faktor zwei und passen damit zur diskreten Skalierung der Bisektionshierarchie. Ein Zusammenfallen transformierter Kurven ist deshalb noch kein Nachweis für beliebige Zwischenbreiten. Nichtdyadische Zwischenwerte und gegenüber der Hierarchie verschobene Zentren sind sinnvolle spätere Kontrollen, werden aber nicht nachträglich in dieses vorab festgelegte Design aufgenommen.

7 Archiv und Reproduktion

Alle einzelnen JSON-Protokolle, Metadaten, Fehler-/Ressourcenprotokolle, Quellkopien und die fünf vorab festgelegten Beispielgitter liegen unter `data/`. `uniform_reference.json` enthält die verwen-

deten historischen Vergleichsdaten samt Herkunft und ursprünglichen Prüfsummen. `manifest.json` sichert die archivierten Eingaben. Primär sind die Einzelprotokolle, nicht die gerundeten Tabellen des Berichts.

`build_report.py` leitet sämtliche Abbildungen und Tabellen neu aus dem Archiv ab. Es startet keine Messreihe. Die Hauptdatei dieses Berichts ist `Bericht.tex`, der Testlauf liegt in `test_results.txt`. Im Berichtsverzeichnis:

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

Eine vollständige neue Serie mit den archivierten Quellen wird in ein neues Verzeichnis geschrieben:

```
python3 data/source/benchmark_gaussian.py \
  --output /tmp/cfk-gaussian-g1-replay
```

Die Option `-resume` setzt einen unterbrochenen Lauf mit derselben Quellkopie und Pythonversion fort; fertige Einzelprotokolle bleiben erhalten. Bereits dokumentierte Versuchsarchive sollen nicht überschrieben oder als neue unabhängige Messungen ausgegeben werden.