# Asymptotic theory for sequencial random partition of integers

Takuya Koriyama

University of Chicago

September 4, 2026

- Asymptotic mixed normality of maximum likelihood estimator for Ewens–Pitman partition
  T.Koriyama, T.Matsuda, F.Komaki
  Advances in Applied Probability, 58(1) 214-234, 2026
- Asymptotic Inference for Exchangeable Gibbs Partitions
  T.Koriyama
  Stochastic Processes and their Applications, 202, 105074, 2026.

# Motivation: Modelling dynamics of clustering

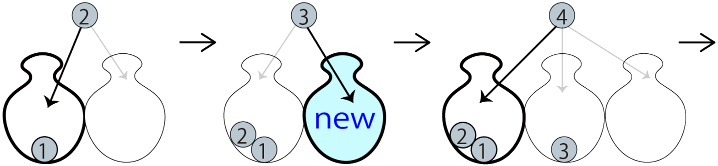
- Suppose we observe some units (e.g., customers, animals, etc) sequentially, which form clusters according to an unknown rule.
- Goal-1: Characterize the dynamics of the macroscopic behavior as the items increase, e.g., Zipf's law in NLP.
- Goal-2: Prediction of the allocations of the new unit based on the clusters formed by previous units.

# Sequential partitions of integers

- For each $n \in \mathbb{N}$, consider partitions of $[n] \equiv \{1, 2, \dots n\}$.
- We use $\Pi_n$ to denote a partition of $[n]$. For instance, $\Pi_3$ can be

$$\{\{1,2,3\}\} \quad \text{or} \quad \begin{matrix} \{\{1,2\},\{3\}\} \\ \{\{2,3\},\{1\}\} \\ \{\{3,1\},\{2\}\} \end{matrix} \quad \text{or} \quad \{\{1\},\{2\},\{3\}\}$$

- Consider a sequential partition of integers $(\Pi_n)_{n=1}^{\infty}$.




$$\Pi_1 = \{\{1\}\} \to \Pi_2 = \{\{1,2\}\} \to \Pi_3 = \{\{1,2\},\{3\}\} \to \Pi_4 = \{\{1,2,4\},\{3\}\}$$

- Ewens–Pitman partition is a specific rule to assign balls to cells randomly.

# Definition of Ewens–Pitman partition

**Algorithm 1:** Ewens–Pitman Partition

Fix $\alpha \in (0,1)$ and initialize $\Pi_1 \leftarrow \{\{1\}\}$;

**for** $n = 1, 2, \ldots$ **do**

Writing $\Pi_n = \{U_1, \ldots, U_{\mathsf{k}_n}\}$, assign the next $(n+1)$-th balls randomly as

$$(n+1) \in \begin{cases} U_i & \text{with prob. } \dfrac{1}{n}(|U_i| - \alpha), \quad \text{for } i \in \{1, 2, \ldots \mathsf{k}_n\} \\ \text{a new set} & \text{with prob. } \dfrac{\mathsf{k}_n}{n}\alpha \end{cases}$$

- $\mathsf{k}_n$ is the number of nonempty cells.
- $(|U_i| - \alpha)$ induces “The richer gets richer” mechanism
- $\alpha \in (0,1)$ serves as a discount parameter

# Likelihood formula of the Ewens–Pitman partition

**Theorem (e.g., Pitman (2003))**

Given a partition $\Pi_n$ of $[n]$, let

$$\mathsf{k}_{n,j} \equiv \text{Number of cells of size } j, \quad \text{for } j \in [n]$$

and let $\mathsf{k}_n = \sum_{j=1}^n \mathsf{k}_{n,j}$ be the number of nonempty cells. Then, we have

$$\mathbb{P}(\Pi_n; \alpha) = \frac{(\mathsf{k}_n - 1)!}{(n-1)!} \cdot \alpha^{\mathsf{k}_n - 1} \cdot \prod_{j=2}^{n} \left( \prod_{i=1}^{j-1} (i - \alpha) \right)^{\mathsf{k}_{n,j}}$$

- $(\mathsf{k}_{n,j})_{j \geq 1}$ are sufficient statistics
- The likelihood is invariant under the permutations of the labels of balls.

# Macroscopic behavior

**Theorem (Pitman (2006))**

*Let* $\mathsf{k}_{n,j}$ *be the number of cells of size* $j$ *and* $\mathsf{k}_n = \sum_{j\geq 1} \mathsf{k}_{n,j}$.

- *There exists a positive random variable* $\mathsf{M}_\alpha$ *such that*

$$\frac{\mathsf{k}_n}{n^\alpha} \xrightarrow{\text{a.s.}} \mathsf{M}_\alpha \quad \text{as } n \to +\infty.$$

- *For each fixed* $j \in \mathbb{N}$, *as* $n \to +\infty$,

$$\frac{\mathsf{k}_{n,j}}{\mathsf{k}_n} \xrightarrow{\text{a.s.}} \mathfrak{p}_\alpha(j) \equiv \frac{\alpha \prod_{i=1}^{j-1}(i-\alpha)}{j!},$$

*where* $\mathfrak{p}_\alpha(j)$ *is a probability mass function over* $j \in \mathbb{N}$.

- The discrete distribution with p.m.f $\mathfrak{p}_\alpha(j)$ is called the Sibuya distribution
- This is heavy-tailed:

$$\mathfrak{p}_\alpha(j) \sim \frac{\alpha}{\Gamma(1-\alpha)} \cdot j^{-(\alpha+1)} \text{ as } j \to +\infty,$$

# Macroscopic behavior

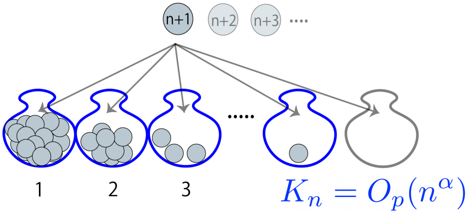
**Theorem (Pitman (2006))**

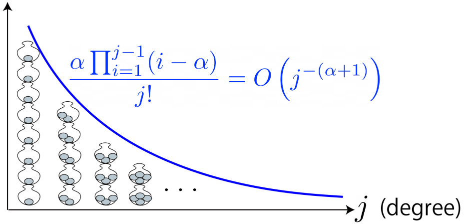
- *There exists a positive random variable* $\mathsf{M}_\alpha$ *such that*

$$\frac{\mathsf{k}_n}{n^\alpha} \xrightarrow{\text{a.s.}} \mathsf{M}_\alpha \quad \textit{as } n \to +\infty.$$

- *For each fixed* $j \in \mathbb{N}$, *as* $n \to +\infty$,

$$\frac{\mathsf{k}_{n,j}}{\mathsf{k}_n} \xrightarrow{\text{a.s.}} \mathfrak{p}_\alpha(j) \equiv \frac{\alpha \prod_{i=1}^{j-1}(i-\alpha)}{j!}.$$



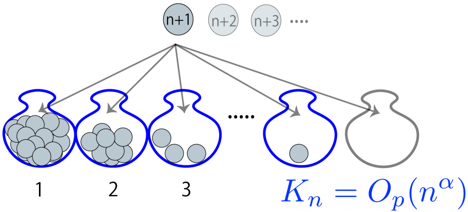



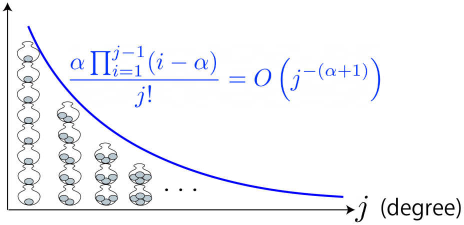

$\frac{\mathsf{k}_n}{n^\alpha} \to \mathsf{M}_\alpha$ (left) and $\frac{\mathsf{k}_{n,j}}{\mathsf{k}_n} \to \frac{\alpha \prod_{i=1}^{j-1}(i-\alpha)}{j!}$ (right)

# Application

| | $k_n$ | $k_{n,j}$ |
|---|---|---|
| Ecology | Species | Species $j$ times observed |
| Nonparametric Bayes | Clusters | Clusters of size $j$ |
| Network Analysis | Vertices | Vertices with $j$ edges |

**Estimation of $\alpha$ from observed partitions $\Pi_n$ is of particular interest.**



# Prior works on the estimation of $\alpha$

- By $n^{-\alpha}\mathsf{k}_n \xrightarrow{\text{a.s.}} \mathsf{M}_\alpha$, one can show that the estimator

$$\hat{\alpha}_n^{\mathsf{naive}} \equiv \frac{\log \mathsf{k}_n}{\log n}$$

  is consistent, but convergence rate is slow ($(\log n)^{-1}$ rate).
- The MLE $\hat{\alpha}_n$

$$\hat{\alpha}_n \equiv \arg\max_{\alpha\in(0,1)} \mathbb{P}(\Pi_n;\alpha) = \arg\max_{\alpha\in(0,1)} \alpha^{\mathsf{k}_n-1}\cdot\prod_{j=2}^{n}\left(\prod_{i=1}^{j-1}(i-\alpha)\right)^{\mathsf{k}_{n,j}}$$

  was known to achieve a faster rate (Balocchi et al. (2022))
- However, the asymptotic distribution of the MLE was unknown.

# Table of content

# Asymptotic distribution of the MLE

**Theorem (Koriyama et al. (2026))**

*Let* $\mathfrak{i}(\alpha)$ *be the Fisher Information of the Sibuya distribution:*

$$\mathfrak{i}(\alpha) \equiv \sum_{j=1}^{\infty} \mathfrak{p}_\alpha(j) \left( \frac{\partial}{\partial \alpha} \log \mathfrak{p}_\alpha(j) \right)^2, \quad \textit{with} \quad \mathfrak{p}_\alpha(j) = \frac{\alpha \prod_{i=1}^{j-1} (i - \alpha)}{j!}$$

*Let* $\mathsf{k}_n$ *be the number of nonempty cells. Then,*

$$\sqrt{\mathsf{k}_n \mathfrak{i}(\alpha)} \cdot (\hat{\alpha}_n - \alpha) \to \mathcal{N}(0, 1)$$

- Since $\mathsf{k}_n \sim n^\alpha$ (a.s.), we have $\hat{\alpha}_n - \alpha = O_p(n^{-\alpha/2})$

# Scaling by $\mathsf{k}_n$ is the key

Recall $\sqrt{\mathsf{k}_n \mathfrak{i}(\alpha)} \cdot (\hat{\alpha}_n - \alpha) \to \mathcal{N}(0,1)$

**Theorem (Koriyama et al. (2026))**

*Let $I_n(\alpha)$ be the Fisher Information of the Ewens–Pitman partition:*

$$I_n(\alpha) = \mathbb{E}\Big[\big(\partial_\alpha \log \mathbb{P}(\Pi_n;\alpha)\big)^2\Big]$$

*Then, the MLE error scaled by $\sqrt{I_n(\alpha)}$ converges to a variance mixture of normals:*

$$\sqrt{I_n(\alpha)}(\hat{\alpha}_n - \alpha) \to^d \sqrt{\frac{\mathbb{E}[\mathsf{M}_\alpha]}{\mathsf{M}_\alpha}} \cdot N \quad \textit{where} \quad \begin{cases} \mathsf{M}_\alpha =^d (\lim_{n\to+\infty} n^{-\alpha}\mathsf{k}_n) \\ N =^d N(0,1) \end{cases}$$

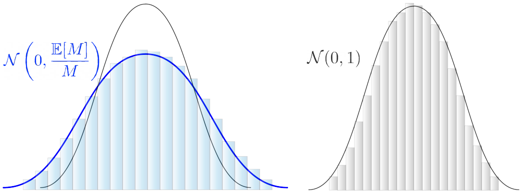



Figure: Limit laws of the MLE error scaled by $\sqrt{I_n(\alpha)}$ (left) vs $\sqrt{\mathsf{k}_n \mathfrak{i}(\alpha)}$ (right)

# Several questions remain

- Q1. What if the model is miss-specified?
- Q2. Can we predict the allocation of the $(n+1)$-th item given the partitions of $[n]$?

# Table of content

# Gibbs partitions – Extension of Ewens–Pitman partitions

- Recall that Ewens–Pitman's allocation was

$$(n+1) \in \begin{cases} U_i & \text{with prob. } \dfrac{1}{n}(|U_i| - \alpha) \quad \text{for } i \in \{1, 2, \dots \mathsf{k}_n\} \\ \text{new cell} & \text{with prob. } \dfrac{\mathsf{k}_n}{n}\alpha \end{cases}$$

- Gibbs partitions generalize the rule as:

$$(n+1) \in \begin{cases} U_i & \text{with prob. } \dfrac{v_{n+1,\mathsf{k}_n}}{v_{n,\mathsf{k}_n}} \cdot (|U_i| - \alpha), \quad \text{for } i \in \{1, 2, \dots \mathsf{k}_n\} \\ \text{new cell} & \text{with prob. } \dfrac{v_{n+1,\mathsf{k}_n+1}}{v_{n,\mathsf{k}_n}} \end{cases},$$

where $v_{n,k}$ are another parameters paired with $\alpha \in (0, 1)$.

# Gibbs partitions – Extension of Ewens–Pitman partitions

$$(n+1) \in \begin{cases} U_i & \text{with prob. } \dfrac{v_{n+1,\mathsf{k}_n}}{v_{n,\mathsf{k}_n}} \cdot (|U_i| - \alpha), \quad \text{for } i \in \{1, 2, \dots \mathsf{k}_n\} \\ \text{new cell} & \text{with prob. } \dfrac{v_{n+1,\mathsf{k}_n+1}}{v_{n,\mathsf{k}_n}} \end{cases}$$

- For the probability to sum up to $1$, $v_{n,k}$ must satisfy a backward recursion:

  For all $1 \leq k \leq n, \quad v_{n,k} = (n - \alpha k) \cdot v_{n+1,k} + v_{n+1,k+1}, \quad$ with $\quad v_{1,1} = 1.$

- Example: two parameters $(\alpha, \theta)$ family

$$v_{n,k} \equiv \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{\prod_{i=0}^{n-1}(\theta + i)}.$$

(Ewens–Pitman partition corresponds to $\theta = 0$)

# Representation ability of Gibbs partition

**Theorem (Pitman (2006))**

*Let $\Pi_n$ be a random partition of $[n]$ such that $(\mathsf{k}_{n,j})_{j=1}^{n}$ are sufficient statistics. Then, for any $\alpha \in (0,1)$, the followings are equivalent:*

- *$\exists$ triangle seq. $\{v_{n,k}\}$ such that $\Pi_n$ follows the Gibbs partition $(\{v_{n,k}\}, \alpha)$.*
- *$\exists$ positive random variable $\mathsf{S}$ such that*

$$\frac{\mathsf{k}_n}{n^\alpha} \xrightarrow{\text{a.s.}} \mathsf{S} \quad \text{as } n \to +\infty.$$

- *For each fixed $j \in \mathbb{N}$, as $n \to +\infty$,*

$$\frac{\mathsf{k}_{n,j}}{\mathsf{k}_n} \xrightarrow{\text{a.s.}} \mathfrak{p}_\alpha(j) \equiv \frac{\alpha \prod_{i=1}^{j-1}(i-\alpha)}{j!}.$$

- Indeed, $\{v_{n,k}\}$ and $\mathsf{S}$ are in one-to-one relation

# Working assumption on $v_{n,k}$

Recall the two parameters $(\alpha, \theta)$ family

$$v_{n,k} \equiv \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{\prod_{i=0}^{n-1}(\theta + i)}$$

satisfies the backward recursion.

$$v_{n,k} = (n - \alpha k) \cdot v_{n+1,k} + v_{n+1,k+1} \tag{1}$$

**Assumption**

*There exists a mixture distribution $\mu$ on $(-\alpha, +\infty)$ with a bounded support such that*

$$v_{n,k} = \int \mathrm{d}\mu(\theta) \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{\prod_{i=0}^{n-1}(\theta + i)}$$

- Taking $\mu = \delta_0$ recovers Ewens–Pitman partition.
- Many classes can be covered by our assumptions (see our paper in details)

# The MLE $\hat{\alpha}_n$ still works under the misspecified setting

- Let us use the same MLE $\hat{\alpha}_n$ to estimate $\alpha$:

$$\hat{\alpha}_n = \underset{\alpha \in (0,1)}{\arg\max} \; \alpha^{\mathsf{k}_n - 1} \prod_{j=2}^{n} \left\{ \prod_{i=1}^{j-1} (-\alpha + i) \right\}^{\mathsf{k}_{n,j}} .$$

That is, we estimate $\alpha$ as if $\Pi_n$ is generated by the Ewens–Pitman parition.

**Theorem (Koriyama (2026))**

*Suppose $\Pi_n$ follows the Gibbs partition. Then,*

$$\sqrt{\mathsf{k}_n \mathsf{i}(\alpha)} \cdot (\hat{\alpha}_n - \alpha) \to \mathcal{N}(0,1).$$

- The MLE works in the misspecified setting too.

# Prediction of the next allocation

- Suppose we observe a partition $\Pi_n = \{U_1, \dots, U_{\mathsf{k}_n}\}$.
- We aim to predict the allocation of the next item.
- More precisely, the estimand is the simplex $\mathbf{p}_n := (\mathsf{p}_{n,0}, \mathsf{p}_{n,1}, \dots, \mathsf{p}_{n,\mathsf{k}_n}) \in \Delta^{\mathsf{k}_n+1}$, where

$$\mathsf{p}_{n,i} = \begin{cases} \mathbb{P}\big((n+1) \in \text{new cell} \mid \Pi_n\big) & i = 0 \\ \mathbb{P}\big((n+1) \in U_i \mid \Pi_n\big) & 1 \leq i \leq \mathsf{k}_n \end{cases}$$

# Prediction of the next allocation

$$\mathsf{p}_{n,i} = \begin{cases} \mathbb{P}\big((n+1) \in \text{new cell} \mid \Pi_n\big) & i = 0 \\ \mathbb{P}\big((n+1) \in U_i \mid \Pi_n\big) & 1 \le i \le \mathsf{k}_n \end{cases}$$

- (Recall) If $\Pi_n$ were to follow the Ewens–Pitman partition,

$$\mathsf{p}_{n,i} = \begin{cases} \frac{\mathsf{k}_n}{n}\alpha & i = 0 \\ \frac{1}{n}(|U_i| - \alpha) & 1 \le i \le \mathsf{k}_n \end{cases}$$

- We define the estimator $\hat{\mathbf{p}}_n = (\hat{\mathsf{p}}_{n,0}, \hat{\mathsf{p}}_{n,1}, \ldots, \hat{\mathsf{p}}_{n,\mathsf{k}_n})$ by plugging the MLE $\hat{\alpha}_n$:

$$\hat{\mathsf{p}}_{n,i} = \begin{cases} \frac{\mathsf{k}_n}{n}\hat{\alpha}_n & i = 0 \\ \frac{1}{n}(|U_i| - \hat{\alpha}_n) & 1 \le i \le \mathsf{k}_n \end{cases}$$

# Limit distribution of the TV distance

**Theorem (Koriyama (2026))**

*Suppose $\Pi_n$ follows the Gibbs partition. Then,*

$$\frac{n}{\sqrt{\mathsf{k}_n}}\sqrt{\mathrm{i}(\alpha)} \cdot TV(\hat{\mathbf{p}}_n, \mathbf{p}_n) \to |\mathcal{N}(0,1)|.$$

- By $\mathsf{k}_n \sim n^\alpha$ (a.s.), we have $TV(\hat{\mathbf{p}}_n, \mathbf{p}_n) = O_p(n^{-1+\alpha/2})$.
- We obtain

$$\lim_{n\to+\infty} \mathbb{P}\left(\forall I \subset \{0,1,\ldots,\mathsf{k}_n\}, \quad \mathbf{p}_n(I) \in \left[\hat{\mathbf{p}}_n(I) \pm \frac{\sqrt{\mathsf{k}_n}}{n}\frac{\tau_{0.975}}{\sqrt{\mathrm{i}(\hat{\alpha}_n)}}\right]\right) = 0.95.$$

- We can derive limit distributions for general $f$-Divergence.

# Summary

- We derive the asymptotic distribution of the MLE $\hat{\alpha}_n$ for the Ewens–Pitman partition.
  - $\hat{\alpha}_n$ is asymptotically mixed normal.
  - By a proper *random* normalization, the randomness is cancelled out.

- We extend the result to Gibbs-partition.
  - The MLE retains a similar asymptotic distribution.
  - We proposed an estimator to predict the next allocation and derive the limit distribution.

# Reference I


Cecilia Balocchi, Stefano Favaro, and Zacharie Naulet. Bayesian nonparametric inference for" species-sampling" problems. *arXiv preprint arXiv:2203.06076*, 2022.

Takuya Koriyama. Asymptotic inference for exchangeable gibbs partitions. *Stochastic Processes and their Applications*, page 105074, 2026.

Takuya Koriyama, Takeru Matsuda, and Fumiyasu Komaki. Asymptotic mixed normality of maximum-likelihood estimator for ewens–pitman partition. *Advances in Applied Probability*, 58(1):214–234, 2026.

Jim Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.

Jim Pitman. *Combinatorial Stochastic Processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer, 2006.

# Confidence interval with fixed width can be vacuous

Recall: $\lim_{n\to+\infty} \mathbb{P}\left(\forall I \subset \{0,1,\ldots,\mathsf{k}_n\},\ \mathbf{p}_n(I) \in \left[\hat{\mathbf{p}}_n(I) \pm \frac{\sqrt{\mathsf{k}_n}}{n}\frac{\tau_{0.975}}{\sqrt{\mathfrak{i}(\hat{\alpha}_n)}}\right]\right) = 0.95.$

- The width shrinks at the rate $\frac{\sqrt{\mathsf{k}_n}}{n}$.
- However, the width is fixed across $I$, and hence can become vacuous when $\mathbf{p}_n(I) \ll \frac{\sqrt{\mathsf{k}_n}}{n}$
- We propose a local confidence interval of the form

$$\left[\hat{\mathbf{p}}_n(I) \pm \frac{|I|}{n\sqrt{\mathsf{k}_n}}\frac{\tau_{0.975}}{\sqrt{\mathfrak{i}(\hat{\alpha}_n)}}\right]$$

- We show that this confidence interval has approximate $95\%$ coverage for any $I$ such that $\mathbf{p}_n(I) \ll \frac{\sqrt{\mathsf{k}_n}}{n}$ (see our paper in details).

# Connection to Nonparametric Bayesian Inference

- Let us fix a non atomic measure $F$(say $F = N(0,1)$).
- For any $\alpha \in (0,1)$, Poisson Dirichlet prior is a discrete random measure

$$\mathsf{PD}(\alpha) := \sum_{i=1}^{\infty} p_i \delta_{y_i}, \quad y_i \overset{\text{iid}}{\sim} F,$$

where the weight $(p_i)_{i=1}^{\infty}$ are constructed in the following "stick breaking" process:

**Algorithm 2:** Stick breaking process

**for** $i = 1, 2, \ldots$ **do**
  Sample $v_i \sim \mathsf{Beta}(1-\alpha, i\alpha)$
  Take $p_i = v_i \cdot \sum_{j=1}^{i-1}(1 - v_j)$

- Let $P \sim \mathsf{PD}(\alpha)$ and $(X_i)_{i=1}^{n} \overset{\text{iid}}{\sim} P$. Since some of $(X_i)_{i=1}^{n}$ have ties, they induce a partition of $[n]$.

**Theorem (Pitman (2006))**

*A partition $\Pi_n$ induced by the conditional iid sample from PD$(\alpha)$ has the same law as the Ewens–Pitman partition $(\alpha)$.*

## The law of $\mathsf{M}_\alpha = \lim_{n \to +\infty} n^{-\alpha} \mathsf{k}_n$

- For each $\alpha \in (0,1)$, let $S_\alpha$ be the positive random variable characterized by

$$\lambda \geq 0, \; E[S_\alpha^\lambda] = e^{-\lambda^\alpha}.$$

- Mittag-Leffler distribution $(\alpha)$ is the law of $M_\alpha$ defined as

$$M_\alpha := (S_\alpha)^{-\alpha}$$

# Simulation

Simulation setting: $\alpha = 0.8$, $\theta = 0$, $100000$ Monte Carlo simulations.

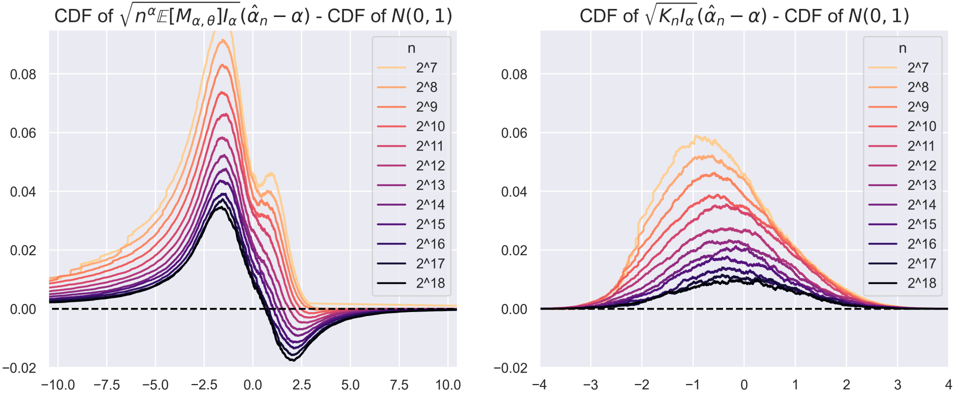



Figure: The visualization of the asymptotic mixed normality. The left figure plots the difference of CDF of $\sqrt{\mathrm{i}_{\alpha\alpha}^{(n)}}(\hat{\alpha_n} - \alpha)$ to the CDF of $N(0,1)$, while the right figure plots the difference of the CDF of $\sqrt{K_n \mathrm{i}_\alpha}(\hat{\alpha}_n - \alpha)$ to the CDF of $N(0,1)$.