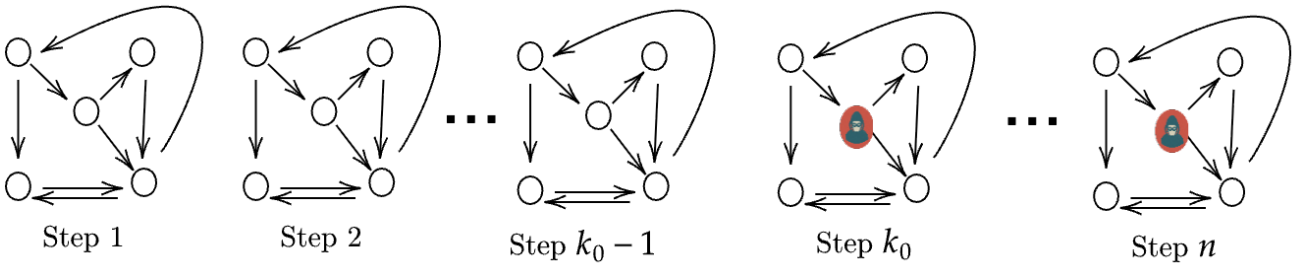
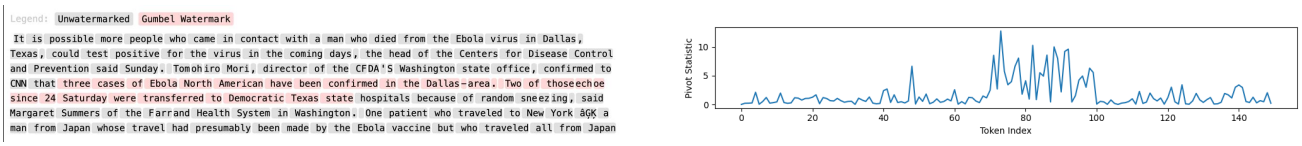
# Research Statement

**Soham Bonnerjee**
Department of Statistics, University of Chicago
sohambonnerjee@uchicago.edu

My core research interests lie at the intersection of time-series and machine learning, with an emphasis on tackling complex non-stationarity arising in various modern applications. I develop valid theoretical tools to perform statistical inference for dependent datasets, which have been increasingly commonplace as outputs of iterative optimization algorithms, autoregressive next token generation by large language models, as well as more classical temporal and spatial datasets from disciplines such as epidemiology, geography, climate science and archaeology.




(a)




(b)

Figure 1: (a): In a distributed learning system, client(s) may turn malicious at some step; identification of this step as well as the client, are change-point problems [6]. (b) In a text possibly contaminated by LLM-generated watermarked outputs, detecting the watermarked segments is an epidemic change-point problem.[4]

A unifying thread in much of my work is to analyze dependent observations from myriad settings (such as SGD iterates, or pivot statistics from LLM-generated tokens) through a time-series lens. For both classical time-series as well as learning theory research, my theoretical interests often go beyond central limit theory to obtain *sharp Gaussian approximations* with finite-sample guarantees, facilitating validity of bootstrap, confidence sets, and hypothesis tests in non-i.i.d. settings. So far, a majority of my research can be categorized into three intertwined directions.

- **Gaussian approximation and its applications for time series.** I develop sharp, *time-uniform* Gaussian approximation results (also called *Komlos-Major-Tusnady* approximation for multivariate stationary and non-stationary processes [5], augmented with explicit construction of said approximations, along with practically usable valid bootstrap strategies. These approximations has numerous applications such as in change-point literature [3, 12], simultaneous inference [3, 16], and conditional independence testing[14].
- **Inference for spatially dependent data.** I develop Gaussian-approximation based valid inference strategy for spatially dependent datasets. Most of the literature in this area employs Bayesian methodology to capture dependence. Instead, we employ a direction-agnostic notion of dependence to perform valid inference for *spatial random effect* models. In an ongoing work, I have also come up with a notion of *spatial change-point*, and have introduced fast, accurate algorithms to localize such change-points.

- **Inference for dependent datasets in modern machine learning.** I develop distributional theory such as *stable* convergence, *Berry-Esseen* bounds or time-uniform couplings for the iterates of various stochastic approximation algorithms (SA) such as *Stochastic Gradient Descent* (SGD), *Q-learning* and *Federated Learning*. Concurrently, on the theory of large language models (LLM), I have also worked on leveraging change-point tools to perform provably consistent watermark detection in mixed-source (containing both human and LLM generated) texts. The dependence prevalent in these datasets (such as iterates of SA algorithms, or words of a text) cannot be easily classified into any class of "weak" or "strong" dependence vis-à-vis classical time series literature, and usually requires a case-by-case treatment [6, 4, 2].

In parallel to my core interests as described above, I have also been working on (i) *privacy-aware* training of Transformers using novel variants of *Differentially-private SGD* (DP-SGD) [7]; (ii) valid inference for temporally dependent *object* data (e.g. data taking values in a *Hadamard* space).

In the following, I describe my main completed research projects in detail, modularized by their respective themes.

## Inference for time series: theory and application

**Sharp Gaussian approximations for non-stationary time series with explicit construction.**

In this project, joint with Sayar Karmakar and Wei Biao Wu [5] and published in *Annals of Statistics*, we tackle the problem of *strong invariance principle* in non-stationary time series. Previous results indicated that there exists a Gaussian process that can form a coupling with the partial sums of the non-stationary processes while maintaining an optimal rate. However, no concrete structure about such processes were known, preventing their use in statistical inference. In this article, we present an explicit construction of a Gaussian coupling for univariate non-stationary time-series - including the *first ever* construction of a Gaussian coupling whose covariance structure exactly matches that of the non-stationary process- while maintaining the optimal rate. Additionally, we propose a consistent estimation strategy of the said covariance structure, thereby facilitating Gaussian bootstrap-based inference. This paper has applications in change-point analysis, constructing simultaneous confidence band and wavelet analysis.

**Testing for synchronization of change-points in multiple time-series.**

As a non-trivial use-case of the strong invariance principle for practically relevant problems, we tackle the problem of testing for synchronization of multiple change-points in this joint work with Sayar Karmakar, Maggie Cheng and Wei Biao Wu [3]. In many multivariate time-series with change-points, cross-sectional dependence might lead to shared or clustered change-points in different covariates; however, usual literature on change-point analysis seems to assume a shared change-point for the most part. In this work, we describe a valid, Gaussian-bootstrap based algorithm to test such hypotheses of shared change-points. This work has received a *Major Revision* request from *Biometrika*. It has also received a *Hannan Graduate Student Travel Award* from the *Institute of Mathematical Statistics*(IMS).

## Inference for modern machine learning algorithms

**Sharp asymptotic inference for Decentralized Federated Learning (DFL)**

In decentralized federated learning systems, multiple clients collaborate periodically through a connection graph $\mathbf{C}$ to perform an optimization problem through *local SGD* algorithm. Concretely, suppose

$\theta^\star_K = \arg\min_\theta \sum_{k=1}^K w_k F_k(\theta) \in \mathbb{R}^d$; then, one follows the iterative algorithm

$$\boldsymbol{\Theta}_t = (\boldsymbol{\Theta}_{t-1} - \eta_t \mathbf{G}_t)\, C_t, \quad C_t = \begin{cases} \mathbf{C}, & t \in E_\tau, \\ I_K, & \text{otherwise.} \end{cases},$$

where $\boldsymbol{\Theta}_t = (\theta_t^1, \ldots, \theta_t^K) \in \mathbb{R}^{d\times K}$ denotes the local parameter updates of each client at the $t$-th step, $\mathbf{G}_t$ denotes the corresponding local gradient updates and $\eta_t$ are the step-sizes. The final estimate is $Y_n := K^{-1}\boldsymbol{\Theta}_n\mathbf{1}$. In a joint article [6] with Sayar Karmakar and Wei Biao Wu, we address two key theoretical gaps. Firstly, we derive the *first-ever* Berry-Esseen theorem for $Y_n$, which also sheds light into optimal choice of the learning rate $\eta_t$. Secondly, mirroring the KMT-approximations, we present two, first-of-its-kind *time-uniform* and rate-optimal Gaussian approximations for the local SGD iterates $\boldsymbol{\Theta}_t$ and $Y_t$. We also discuss how these results motivate valid Gaussian bootstrap-based algorithm to identify the onset of adverserial attacks such as Man-In-the-Middle. This has been accepted as a *Spotlight* poster (Top 3%) in *NeurIPS 2025*.

### Segmentation of watermarked texts

With the advent of Large Language Models, detection of machine-generated texts have become important. A common method uses *Watermarking*, which makes the use of LLM detectable *only* when keys corresponding to each token are present, and un-detectable otherwise. Recent theoretical work in this area has mostly focused on the testing problem of unwatermarked vs. watermarked texts. In a joint research [4] with Sayar Karmakar and Subhrajyoty Roy, we tackle the more difficult problem of identifying such watermarked segments from a mixed-source texts by introducing a novel perspective of *epidemic change-point*. Adapting key ideas from time-series literature to the novel setting of watermarked texts, we propose `WISER`: a provably consistent and computationally efficient algorithm to identify multiple watermarked segments from a text input. Apart from rigorous theoretical guarantees, our algorithm is shown to out-perform all competitive algorithms-most of which lack theoretical guarantees or are computationally expensive- on benchmark datasets for various watermarking schemes. This work is submitted to *ICLR 2026*.

## Inference for spatially dependent data

### Inference for spatial random effect model

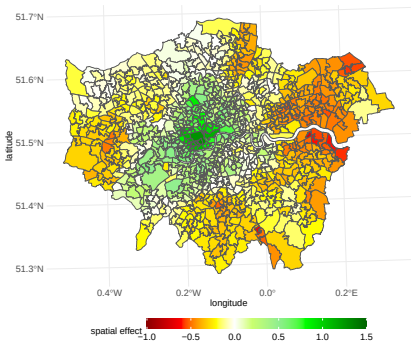



Figure 2: Spatial effect on London housing price dataset

In spatial data analysis, the spatial dependency is often left unaddressed, or tackled by putting a Gaussian prior with some covariance structure eg. Matérn or squared-exponential kernels. Such assumptions are often un-testable, necessitating a more general treatment. In a joint work with Soudeep Deb and Wei Biao Wu, we consider inference on spatial random effect model $Y_{ij} = X_{ij}^\top\beta + U_i + \varepsilon_{ij}$, where $(Y_{ij}, X_{ij})$ are the observed response-covariate pairs and $\varepsilon_{ij}$ are the i.i.d. random errors. Here, the dependency structure of the spatial effects $(U_i)_{i\in\mathbb{Z}^d}$ is simply characterized by $U_i = g(e_{i-s} : s \in \mathbb{Z}^d)$, where $e_i, i \in \mathbb{Z}^d$ are i.i.d. This characterization is quite general, and arises naturally out of writing out the joint distribution of $(U_k)_{k\in\mathbb{Z}^d}$ in terms of compositions of conditional quantile functions of i.i.d. uniform random variables. Then, under mild assumptions on the covariates $X_{ij}$, we establish a central limit theory for the least square estimate $\hat\beta$ of $\beta$, and then proceed to provide a consistent estimate of the corresponding asymptotic variance. These inference results motivate valid inferential procedures to test for presence

of spatial effects as well presence of spatial correlations, which we implement to investigate *London house price dataset*. Our results indicate a considerable spatial effect, as can also be seen in Figure 3. The preprint will be online soon, and we plan to submit it to *Biometrika*.

### Localization of spatial change-points

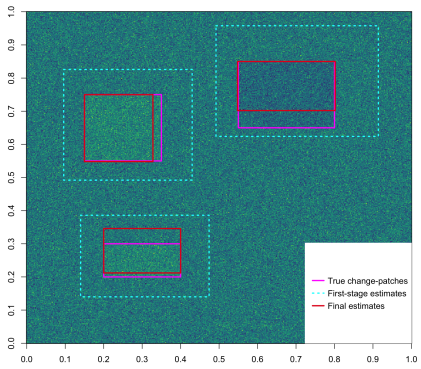



Figure 3: Spatial change-patches localization via two-stage algorithm

Despite huge literature on change-points in time-series, research on change-points on spatially dependent datasets is almost non-existent. In an *ongoing* work with Sayar Karmakar and George Michailidis, we deviate from the scan statistics literature and their Gaussian/i.i.d. assumptions and generalize the notion of epidemic change-points to define spatial *change-patches*. Subsequently, we propose a theoretically valid, two-stage algorithm to localize all rectangular change-patches in a spatially dependent datasets. The first stage produces the efficiency by obtaining a coarse estimate of the patches, before a further fine-tuning produces final estimate. The emphasis on speed is important, since traditional algorithm are prohibitively slow on large spatial datasets. Apart from rigorous theoretical results, our simulation studies have established the efficacy of our method.

## Future research directions

### Change-point estimation on Random Object

Time-series data taking values in metric spaces, which we refer to as *random objects*, are increasingly common in real-world applications, such as graph Laplacians, covariance matrices, probability distributions, and compositional vectors with examples in various domains like brain imaging, social networks, income histograms, microbiome data, and genetics [1] `[some more references]`. However, most of the statistical literature have investigated change-point problems in object data either by assuming i.i.d. observations, or assuming an embedding onto Hilbert space or some tangent space. We aim to investigate the change-point localization problem as well as multiple change-point detection in a general dependent framework, by potentially generalizing *functional dependence measure*[15] and establishing guarantees on object-versions of *Wild Binary Segmentation* [8] or *Seeded Binary Segmentation* [9].

### Statistical foundations on Watermarking in LLMs

Building on [4], we aim to investigate several pertinent problems in watermarking. For example, as opposed to the usual offline tests [10, 11], it is relevant to propose a valid watermark detection-scheme that parses a text sequentially, stopping whenever it has determined to have encountered watermarked segments. On the other hand, in realistic scenarios it might not be known if a specific watermarking scheme is the only candidate to have been potentially used. In these scenarios, it is useful to employ the concept of *e-values* [13] to combine different watermark-detection tests on the same text input.

Some other potential research directions include change-point analysis for high-dimensional logistic regression, capturing cliques or clusters in a spatio-temporal setting, and inference on various popular iterative algorithms.

## References


[1] J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, and R. L. Buckner. Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–562, 2010.

[2] S. Bonnerjee, Y. Han, and W. B. Wu. Stable convergence of stochastic gradient descent for non-convex objectives. *Preprint*, 2025.

[3] S. Bonnerjee, S. Karmakar, M. Cheng, and W. B. Wu. Testing synchronization of change-points for multiple time series. *Major Revision from Biometrika*, 2025.

[4] S. Bonnerjee, S. Karmakar, and S. Roy. Wiser: Segmenting watermarked region - an epidemic change-point perspective. *arXiv preprint arXiv:2509.21160*, 2025.

[5] S. Bonnerjee, S. Karmakar, and W. B. Wu. Gaussian approximation for nonstationary time series with optimal rate and explicit construction. *The Annals of Statistics*, 52(5):2293–2317, 2024.

[6] S. Bonnerjee, S. Karmakar, and W. B. Wu. Sharp gaussian approximations for decentralized federated learning. *arXiv preprint arXiv:2505.08125; NeurIPS 2025, Spotlight*, 2025.

[7] S. Bonnerjee, Z. Wei, A. Asch, S. Nandy, P. Ghosal, et al. How private is your attention? bridging privacy with in-context learning. *arXiv preprint arXiv:2504.16000*, 2025.

[8] P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[9] S. Kovács, P. Bühlmann, H. Li, and A. Munk. Seeded binary segmentation: a general methodology for fast and optimal changepoint detection. *Biometrika*, 110(1):249–256, 2023.

[10] X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025.

[11] X. Li, G. Wen, W. He, J. Wu, Q. Long, and W. J. Su. Optimal estimation of watermark proportions in hybrid ai-human texts. *arXiv preprint arXiv:2506.22343*, 2025.

[12] F. Mies. Strong gaussian approximations with random multipliers. *arXiv preprint arXiv:2412.14346*, 2024.

[13] A. Ramdas and R. Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

[14] M. Wieck-Sosa, M. F. Haddad, and A. Ramdas. Conditional independence testing with a single realization of a multivariate nonstationary nonlinear time series. *arXiv preprint arXiv:2504.21647*, 2025.

[15] W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005.

[16] W. B. Wu and Z. Zhao. Inference of trends in time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(3):391–410, 2007.