RESEARCH STATEMENT

Soham Bonnerjee*

Overview and Research Vision. My research develops statistical theory and methodology for learning and inference under dependence, nonstationarity, and algorithmic noise. A unifying thread is to obtain *finite- or sharp asymptotics* that (i) expose operational trade-offs (e.g., step-size schedules in SGD), (ii) deliver robust, scalable testing/localization procedures for complex dependence (spatial/temporal, networked, and object-valued data), and (iii) translate to practical algorithms with verifiable uncertainty. This agenda spans: (A) stochastic approximation and optimization (online/streaming inference for SGD and variants), (B) change-point detection and synchronization tests in dependent high-dimensional and non-Euclidean data, and (C) spatial statistics with latent random effects and unknown mean structure.

1. Completed Work (selected)

1.1 A unified, non-asymptotic theory for SGD with *general* learning-rate schedules.

Consider online SGD

$$\theta_t \;=\; \theta_{t-1} - \eta_t \,\nabla f(\theta_{t-1}, \xi_t), \qquad t \geq 1,$$

with potentially *finite-horizon* and/or *cyclical* step-sizes. I develop a single umbrella bound for the pth moment error that isolates the initialization-forgetting and variance-accumulation terms:

$$\|\theta_n - \theta^\star\|_p^2 \;\leq\; \exp\Big(-c_0 \sum_{k=1}^{n} \eta_k\Big)\,\|\theta_0 - \theta^\star\|^2 \;+\; C \sum_{j=1}^{n} \eta_j^2 \,\exp\Big(-c_0 \sum_{k=j+1}^{n} \eta_k\Big),$$

which directly yields sharp consequences for (i) linearly decaying to zero (Linear-D2Z) schedules—showing consistency *and* exponential forgetting of initialization with a $\tilde{O}(n^{-1/2})$ term—and (ii) *cyclical* schedules, where the iterates converge to a *cyclostationary* limit law rather than a stationary one. This explains the persistent periodic error fluctuations empirically observed with cosine/periodic schedules and clarifies their exploration $\leftrightarrow$ convergence tradeoffs.

1.2 Stable end-term CLTs for SGD under nonconvexity and local inference with momentum.

For step-sizes $\eta_t \asymp t^{-\alpha}$, $\alpha \in (1/2, 1)$, I establish *stable* (conditional) CLTs for end-term SGD (and momentum-SGD) around a *local* minimizer a:

$$\Sigma(a)^{-1/2}\,\eta_n^{-1/2}\,(\theta_n - a)\;\Big|\;\{\theta_n \to a\;\} \quad\Rightarrow\quad \mathcal{N}(0, I_d),$$

where $\Sigma(a)$ solves the continuous Lyapunov equation $A\Sigma + \Sigma A^\top = S$ with $A = \nabla^2 F(a)$ and S the noise covariance. The results (i) rigorously justify *local* uncertainty quantification without Polyak–Ruppert averaging, (ii) adapt to momentum with the correct scaling, and (iii) underpin a data-driven GMM-based post-selection procedure to distinguish and infer around multiple minima in nonconvex landscapes.

*Name guessed from co-authorship in the uploaded papers; please replace if incorrect.

1.3 High-dimensional synchronization testing under dependence via valid bootstrap. I develop a max-type synchronization test for multivariate/functional time series under general dependence and nonstationarity, together with a *uniformly* valid high-dimensional bootstrap that controls size while powerfully detecting misalignment. Theory supplies nonasymptotic Gaussian and anti-concentration controls for the studentized maxima that remain valid as both the time horizon and dimension diverge.

1.4 Spatial epidemic change-point localization with near-optimal accuracy. I propose a two-stage "ISEP" localizer for epidemic-type mean shifts in spatio-temporal data: a coarse scan using aggregated patch statistics followed by a refinement stage that provably sharpens the localization error. The method tolerates spatial dependence and heteroskedasticity and remains effective when the signal is sparse and spatially fragmented.

1.5 Estimation and testing for spatial random effects with unknown mean structure. I study a pragmatic spatial random-effects model that *decouples* mean misspecification from spatial correlation learning. The contributions include (i) identification of spurious correlation inflation when the mean is unknown, (ii) a consistent, distance-aware estimator of spatial correlation that is robust to mean misspecification, and (iii) a level-α test of spatial independence with finite-sample calibration improvements.

1.6 Gaussian/local-polynomial approximation for nonstationary series with explicit bias correction. For nonstationary processes with smooth trend $\mu(\cdot)$ and dependence, I obtain uniform expansion for a local-polynomial estimator,

$$\widehat{\mu}_h(t) - \mu(t) - h^2\beta\,\mu''(t) - Q_h(t) = o_p\big((nh)^{-1/2} + h^2\big),$$

where $Q_h(t)$ is an explicit Gaussian/Brownian functional capturing long-run variance. This delivers implementable CIs with data-driven bandwidths under time-varying dynamics.

2. Ongoing and Future Work

2.1 Robust preference learning via Nash sharpening (robust NLHF). I am developing a robust, *game-theoretic* aggregation for human preference data that stabilizes against label noise, adversarial flips, and misspecification. The formulation yields minimax-type guarantees and calibrated uncertainty for preference scores, with practical algorithms that integrate seamlessly with RLHF/LLM fine-tuning.

2.2 Watermark localization and change-point detection under dependence. We design detectors that (i) implant weak "watermarks" and (ii) localize them by matched filtering that remains valid under temporal and spatial dependence, enabling *post-deployment* integrity checks and tamper-detection in streaming systems.

2.3 Multiple change-point localization for dependent *object-valued* data. I am extending localization theory to random objects on manifolds/metric spaces (e.g. SPD matrices, distributions), developing geometry-aware CUSUM/scan statistics with bootstrap calibration that respect curvature and Fréchet means.

2.4 Beyond convexity for general step-size schedules: CLTs and inference. Building on the unified non-asymptotic SGD framework, I am targeting (i) end-term CLTs for cyclical schedules (*cyclostationary* limits) and (ii) inference under Linear-D2Z in nonconvex regimes, including momentum and adaptive optimizers.

3. Broader Impact and Software

The methodological focus on *explicit rates, valid bootstrap, and implementable uncertainty* enables safe deployment in scientific and engineering workflows: e.g., monitoring and rapid localization in sensor networks and neuroimaging (spatial change-points), calibrated synchronization in multimodal experiments (time-varying dependence), and trustworthy online learning (SGD schedules and robust preferences). Reproducible code accompanies the learning-rate studies and will be extended to spatial/temporal toolkits.

4. Selected Technical Highlights (at a glance)

- **SGD with general schedules:** moment bounds isolating forgetting vs. exploration; first rigorous analysis of Linear-D2Z; cyclostationary limit laws for periodic schedules.
- **Nonconvex SGD inference:** stable CLTs at local minima (with momentum) enabling post-selection inference without averaging.
- **High-d sync tests:** uniformly valid bootstrap under dependence; nonasymptotic anti-concentration for maxima.
- **Spatial inference:** random-effects estimation/testing robust to unknown mean; epidemic change-point localization with near-optimal accuracy.

Note. Items §1.5 (spatial random effects) and §1.4 (spatial epidemic change-points) are now treated as *completed*, per the current drafts, while §2.1–§2.3 remain in progress and will be integrated into the final statement as manuscripts mature.