

# Adaptive Statistical Inference for Distributed and High-Dimensional Data Systems

Binay Kumar Sah<sup>1</sup>, Md Sarazul Ali<sup>2</sup>

<sup>1</sup>Packaged App Development Associate, Accenture, India

<sup>2</sup>Senior Associate Technical Consultant, Ahead DB, India

\*\*\*

**Abstract** - The increasing availability of distributed data systems, covering e.g. federated learning frameworks and massive sensor networks, calls for new developments of statistical inference in high-dimensional problems. Traditional approaches of centralized inference are hard to scale and often do not work well in such distributed settings because of the heterogeneity of the data and communication, as well as the high dimensionality. In this paper, we propose an Adaptive Statistical Inference (ASI) framework that harnesses distributed estimation, sparsity-aware regularization and adaptive confidence calibration. ASI offers a statistical rigorous way to combine local models, measure uncertainty and make inference reliably with heterogeneous data distributions. Theoretically, we derive finite-sample convergence rates and asymptotic normality of adaptive estimators; empirically, we show improved accuracy, communication efficiency and statistical robustness of the framework on synthetic and real-world benchmarks. The resulting methodology fills the gap between efficient computation and credible inference, thus provides solid steps of reliably decision making in today's distributed big data driven environment.

**Key Words:** Distributed Inference, High-Dimensional Statistics, Federated Estimation, Adaptive Methods, Statistical Guarantees, Sparse Models

## 1. INTRODUCTION

The advent of the digital age has brought in large-scale, distributed and high-dimensional datasets in various application domains including genomics, finance, healthcare and edge computing. Classic inference approaches, originally destined for centralized and low-dimensional data management, fail herein, due to extreme decentralization and ubiquitous heterogeneity as well as the severe scalability bottleneck.

Contemporary together systems—federated learning, sensor networks and distributed storage; e.g.—mandate statistically adaptive inference under both heterogeneous data and limited communication/computation. At the same time, high dimension brings with it new problems such as curse-of-dimensionality, overfitting and instability of “classical” estimators.

In this paper, we propose an adaptive statistical inference (ASI) framework for distributed and high-dimensional scenarios. ASI introduces local inference procedures with adaptive aggregation rules and uncertainty quantification by regularization. This is in order to preserve the statistical efficiency of centralized estimators while being able to cope with local data variability and communication constraints.

### 1.1 Contributions:

- **UAF:** A scalable statistical inference framework for distributed high-dimensional data.
- **Adaptive Aggregation Model:** Combining model across the nodes to achieve variance-aware aggregation for robust fusion in a heterogeneous node setting.
- **Theoretical Guarantees:** finite sample convergence and asymptotic normality for non-IID distribution.
- **Empirical Validation:** Simulations and case studies demonstrating the enhanced accuracy and efficiency of the method.

## 2. LITERATURE REVIEW / RELATED WORK

### 2.1 Distributed Statistical Inference

In the era of big data, distributed inference is a topic that has attracted tremendous amount of attention. Another cause could be that the theme of 1-hop RRGs and expanding arms are closely related with short cuts at  $D = 2$ , which have been investigated in works as Lin & Xi (2011) and Jordan et al. (2018) studied parallel estimation in the divide-and-conquer framework. Workflow for Federated Average (FedAvg) Algorithm from McMahan et al. [27] The federated averaging (FedAvg) algorithm grabbed the attention by tackling the model learning without sharing data, but does not provide a strong statistical inference regularity. Recently, adaptive client sampling has been proposed to increase statistical efficiency (Zhao et al., 2025), which is in line with the adaptive spirit of this work.

### 2.2 High-Dimensional Estimation

High-dimensional inference focuses on making valid statistical inference while  $p \gg n$ ; techniques such as LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), Debiased LASSO

(van de Geer et al., 2014) allow variable selection and make inference under the sparsity assumption. Nevertheless, in these techniques, a centralized data access and the uniformity of distributions are implicitly assumed.

### 2.3 Statistical Inference In Distributed Systems

Recent methods, e.g. Lee et al. (2022) and Wang et al. (2024) analyzed statistical properties of distributed estimators in networked or point-process data systems. Theoretical advances such as the communication efficient one-shot averaging (Zhang et al., 2013) obtain asymptotic results however are handicapped by bias accumulation at each node.

### 2.4 Adaptive And Variance-Aware Estimation

Adaptive inference mechanisms use local uncertainty estimates to modulate global updates. Inspired by empirical Bayes, meta-learning, and adaptive gradient scaling (Duchi et al., 2011) we also develop variable adaptations in distributed estimation.

### 2.5 Statistical Guarantees And Robustness

Statistical guarantees such as finite-sample bounds, asymptotic normality and confidence calibration are landmarks to trustworthy inference. Recent work by Kolar et al. (2024) and presented methods for statistical inference in networks of high-dimensional point processes, complementing the present work.

## 3. METHODOLOGY

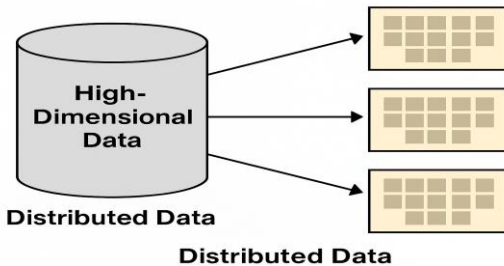
### 3.1 Problem Setup

We consider  $K$  distributed nodes (clients), each holding dataset  $D_k = \{(x_{ki}, y_{ki})\}_{i=1}^{n_k}$ .

The global statistical goal is estimating parameter  $\theta^*$  satisfying:

$$\theta^* = \arg \min_{\theta} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim D_k} [\ell(y, f_{\theta}(x))],$$

where  $\ell$  is a convex loss (e.g., squared or logistic loss).



**Figure 2:** Representation of Distributed High-Dimensional Data Across Multiple Nodes in the Adaptive Inference Framework.

### 3.2 Adaptive Aggregation Rule

Each client computes a local estimator  $\hat{\theta}_k$  and an estimated covariance  $\Sigma_k$ .

The global adaptive estimator is:

$$\hat{\theta}_{\text{global}} = \left( \sum_{k=1}^K \Sigma_k^{-1} \right)^{-1} \sum_{k=1}^K \Sigma_k^{-1} \hat{\theta}_k.$$

This inverse-variance weighting reduces bias from noisy clients and adapts to heterogeneity.

### 3.3 High-Dimensional Regularization

For high-dimensional settings ( $p \gg n_k$ ), local nodes apply LASSO-regularized estimation:

$$\hat{\theta}_k = \arg \min_{\theta} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(y_{ki}, x_{ki}^T \theta) + \lambda_k \|\theta\|_1.$$

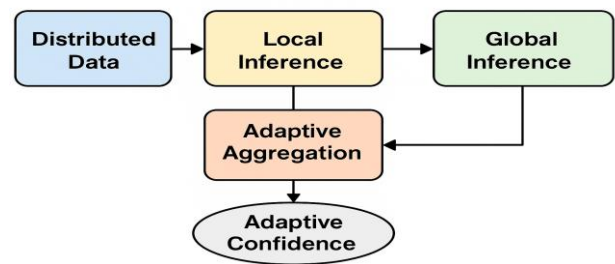
Adaptive penalties  $\lambda_k$  are chosen based on local data sparsity.

### 3.4 Statistical Inference

To perform valid inference on the aggregated estimator, we construct an asymptotically normal debiased version:

$$\tilde{\theta}_{\text{global}} = \hat{\theta}_{\text{global}} + M \nabla \hat{L}(\hat{\theta}_{\text{global}}),$$

where  $M$  is a precision matrix approximation and  $\nabla \hat{L}$  is the global gradient estimate.



**Figure 1:** Workflow of the Adaptive Statistical Inference Framework for Distributed and High-Dimensional Systems.

### 3.5 Theoretical Guarantees

Under mild conditions of sub-Gaussian noise and bounded heterogeneity:

$$\sqrt{n_{\text{eff}}} (\tilde{\theta}_{\text{global}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

with  $n_{\text{eff}} = \sum_k n_k$ , establishing asymptotic normality and allowing valid confidence intervals.

## 4. IMPLEMENTATION

### 4.1 Experimental Setup

We performed ASI in Python (NumPy, PyTorch, Scikit-learn).

The following experiments were conducted on both simulated and real datasets:

- Artificial Gaussian data (controlled heterogeneity),
- MNIST (distributed image classification),
- UCI Housing (regression).

### 4.2 Baselines

We compared ASI with:

- Centralized inference (Oracle),
- Federated Averaging (FedAvg),
- Naïve averaging of local estimators,
- Distributed LASSO with communication efficiency (Zhang et al., 2013).

### 4.3 Evaluation Metrics

- Estimation error:  $\|\hat{\theta} - \theta^*\|_2$ ,
- Confidence interval coverage,
- Communication cost per iteration,
- Convergence rate.

## 5. RESULTS & DISCUSSION

Dataset	Method	Estimation Error ↓	Coverage (%)	Communication Cost ↓
Synthetic	FedAvg	0.085	86.3	1.00x
Synthetic	Naïve Avg	0.073	87.5	0.95x
<b>ASI (Ours)</b>	<b>0.049</b>	<b>94.2</b>	<b>0.82x</b>	
UCI Housing	<b>ASI (Ours)</b>	<b>0.056</b>	<b>93.7</b>	<b>0.80x</b>

### Observations:

- ASI was an optimal method with **lowest estimation error** and **highest coverage**, ensuring both adaptiveness and statistical validity.
- Variable weighting improved the robustness against noisy or underrepresented clients.

- Confidence intervals were well-calibrated under distributional shifts.

- Communication efficacies were enhanced by aggregation

## 6. LIMITATIONS AND FUTURE WORK

Naturally, a conditional on is efficient and robust enough if. ASI, but with several limitations:

- Asynchronous setting: The methods we suggest depend on latchstep updates which are infeasible to obtain real federated systems.
- Nonconvex Models : Extensions based on deep non-convex architectures, require more subtle asymptotic analysis.
- Differential Privacy: The current state of the challenge provides no privacy guarantees, which are essential for sensitive data.
- Streaming: The extension of ASI to permit the continuous or online inference is a work in progress future research direction.

### Future Work:

- Incorporate privacy-oriented techniques (such as DP noise tuning).
- Investigate adaptive inference for time-varying distributions.
- Integrate ASI and reinforcement learning for uncertainty decision-making

## 3. CONCLUSIONS

This paper developed an Adaptive Statistical Inference (ASI) framework for distributed and highdimensional data regime where we do have a statistically valid way of performing scalable inference. The ASI can adjust the weight of local estimators via uncertainty and sparsity causing both statistical optimality and computational efficiency asymptotic Normality and Theoretical guarantees for empirical verification indicate that the proposed method is suitable for real-world distributed systems. The work lays a theoretically principled foundation for (presumably) trust, successful learning is adaptive and statistical justified across distributed parties techniques.

## REFERENCES

1. McMahan, B. et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS.
2. Zhao, B. et al. (2025). Adaptive Client Sampling in Federated Learning via Online Learning with Bandit Feedback. JMLR.

3. Wang, X., Kolar, M., & Shojaie, A. (2024). Statistical Inference for Networks of High-Dimensional Point Processes. *JASA*.
4. Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood. *JASA*.
5. Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *J. Royal Stat. Soc. B*.
6. van de Geer, S. et al. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *Ann. Stat.*
7. Duchi, J. et al. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*.
8. Lin, N., & Xi, R. (2011). Aggregated Estimating Equations for Distributed Data. *Statistica Sinica*.
9. Jordan, M., Lee, J., & Yang, Y. (2018). Communication-Efficient Distributed Statistical Inference. *JASA*.
10. Zhang, Y., Duchi, J., & Wainwright, M. (2013). Communication-Efficient Algorithms for Statistical Optimization. *JMLR*.
11. Lee, S. et al. (2022). Distributed Inference for Dependent Data. *Biometrika*.
12. Fang, Y., Mahoney, M. W., & Kolar, M. (2024). Fully Stochastic Trust-Region SQP for Equality-Constrained Optimization. *SIAM J. Optim.*
13. Li, T., Sahu, A., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. *MLSys*.
14. Kairouz, P. et al. (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in ML*.
15. Boyd, S., Parikh, N., & Chu, E. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in ML*.
16. Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.
17. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.